Machine Learning for Classification of ADHD

Atif Riaz

Department of Computer Science

City, University of London

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

Doctor of Philosophy

City, University of London January 2020

To my parents whose love, affection, encouragement and prays of day and night make me
able to achieve such success and honour,
To my family and other loved ones,
To my lovely little heroes, Musa and Ayan,
To anyone who has shown me friendship, care and kindness.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 65,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures. I grant powers of discretion to the City, University of London librarian to allow the dissertation to be copied in whole or in part without further reference to myself (the author). This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgment.

Atif Riaz
January 2020

Acknowledgements

In the name of God, the most beneficent, the most merciful. Praise be to Him, who has enabled me to pursue this program and allowed me to complete my dissertation in due time with good health. I would like to acknowledge the role of my parents and family members who supported me in achieving this degree.

My profound and sincere gratitude is to my supervisors Eduardo Alonso and Greg Slabaugh, for providing me the opportunity to work on a topic of my passion and guiding me throughout my journey as a researcher. I have been incredibly blessed to have supervisors who were always approachable, available, friendly and willing to help. I would like to replicate their values in my future endeavours.

I would also thank my examiners Xujiong Ye and Christopher Child, for their time and interest in my work. Their feedback was very helpful and influenced the quality of this dissertation. I would also like to extend my gratitude to Constantino Carlos Reyes-Aldasoro for his advice and suggestions on my thesis. Special thanks to senior tutor for research Esther Mondragón for chairing my viva.

This research would not have been possible without the three-year scholarship from Erasmus Mundus LEADERS program. I would like to thank the staff associated with the program, Ana Hajduk and Rehman Aziz. My experience at the City would not have been smooth and comfortable without the academic support staff. My heartiest gratitude goes to Savita Afonso, Nathalie Chatelain, Paula Green and David Mallo-Ferrer for their help on numerous occasions.

Last but not least, I gratefully acknowledge the valuable feedback on my thesis from Muhammad Asad, SM Masudur Rahman Al Arif, Rilwan Remilekun, Radha Kpparti, Shaila Hussain and Sarah Scott.

Abstract

Attention Deficit Hyperactive Disorder (ADHD) is well-known common causation of childhood behavioural disorders. It is estimated that around 5-10% of children globally are affected with this disorder. ADHD is attributed to problematic behaviours that include inattention and impulsivity. Children find it extremely difficult to focus, be attentive and to organise themselves. It contributes to a lifetime of impairment, poor quality of life and long-term burden on affected families. Since there is no single cause found in the prevalence or absence of ADHD. The usual method of diagnosis is merely dependent on behavioural analysis which are all subjective. Clinicians usually take months to diagnose the condition. To date, there are no biological markers that exist for ADHD. To measure neurobiological data objectively, an assessment of the brain behaviour relationship is essential to transform the method of diagnosis. Automatic diagnosis is a profound way for an effective cure.

In this dissertation, we aim to solve the problem of automatic diagnosis of ADHD using machine learning methods based on functional MRI (fMRI) data. The proposed methods begin with classical machine learning and move to deep learning as a way to improve the classification performance. Interpretability of results is an important aspect, so functional connectivity is a central theme in the work and the proposed methods utilise functional connectivity in increasingly more complex ways.

In the first method, we have evaluated a clustering based novel method to calculate functional connectivity. After calculating functional connectivity, we employ Elastic Net feature selection to select the discriminant features and integrate non-imaging data. Finally, a Support Vector Machine (SVM) classifier is trained to classify ADHD.

The second method presents a deep learning based novel method, called FCNet, that calculates functional connectivity from fMRI time-series signals. The FCNet consists of two networks, i) a convolutional neural network in a Siamese architecture that extracts abstract features from a pair of time-series signals and, ii) a similarity measure network that computes the strength of similarity between the extracted features which serves as functional connectivity. Similar to the previous method, an Elastic Net and SVM is applied to classify ADHD.

In the third method, we have proposed an end-to-end trainable model to classify ADHD from preprocessed fMRI time-series data. The model takes fMRI time-series signals as input and outputs the predicted labels, and is trained end-to-end using back-propagation. The proposed model is comprised of three networks, namely i) a feature extractor, ii) a functional connectivity network, and iii) a classification network.

Our findings highlight that functional connectivity serves as an important biomarker towards classification of ADHD and the frontal lobe is altered the most in the case of ADHD. The frontal lobe is known to be associated with cognitive functions like attention, memory, planning and mood. Our findings of the frontal lobe anomalies in ADHD support findings of the earlier studies. Our results reveal that an end-to-end trainable deep network incorporating functional connectivity yields higher detection rates.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

This dissertation explores a set of computer vision problems related to the classification of neurological disorders and proposes different frameworks to classify a subject as healthy control or ADHD based on functional MRI data. In this first chapter, we begin by motivating the need for studying neurological disorders and the application of machine learning in this domain. Next, we review the key original contributions. We end this chapter with a list of articles published during the course of the research and an outline of the remaining chapters.

1.1 Motivation

Brain disorders have emerged as one of the greatest threats to human health. It is estimated that up to one-third of the population suffers from any mental disorder [9] each year. Mental illness is considered to have more impact on human health globally than any other group of chronic diseases [10]. Mental, neurological and substance use disorders constitute 13% of the global burden of disease exceeding both cancer and cardiovascular diseases [11]. In the UK, brain disorders affect approximately 45 million people, accounting for a cost of £120 billion per annum [12]. For many disorders, early and reliable diagnosis is considered critical for mitigating disease effects. Despite the advances in imaging technologies and data analysis techniques, proper clinical diagnosis is not well established and in most cases diagnosis of a neurological disorder is achieved based on physical observations. For example, Attention

Deficit Hyperactivity Disorder (ADHD) is one of the most common neuro-developmental and mental disorders affecting 5-10% of school-going children [5] contributing to lifetime impairment [13], poor quality of life [14] and long-term burden on affected families [13, 14]. Like many other neurological disorders, the underlying mechanism of ADHD is still unknown [5]. To date, no neural biomarker exists that can be used to diagnose ADHD [15] and diagnosis is dependent on observations conducted by medical practitioners or parents and it may take months to make diagnosis based on such observations. The impact of brain disorders on human well-being is very alarming and unfortunately, most of them remain undetected at their early stages. Early diagnosis of any brain disorder is very important as proper medical care can mitigate or possibly eliminate the effects of the disorder. We intend to improve our understanding of brain related disorders with the help of the state-of-the-art advances in machine learning algorithms. Our motivation is to explore novel machine learning algorithms to help medical experts for diagnosis of a brain disorder, ADHD in particular.

1.2 Aims and Objectives

The overall aim of this dissertation is to contribute and evaluate methods for the diagnosis of a brain disorder from functional MRI brain scan data. Our aim is to propose machine learning frameworks that are able to learn the differences between healthy and brain disorder groups from the data. Based on the learned differences, the proposed methods are able to predict whether a new scan belongs to healthy or disorder subjects. The survey and limitations of the existing work motivate us to explore the following aims and objectives:

1. Can functional connectivity (a concept detailed in the next chapter) of brain regions be presented as an important biomarker for diagnosis of a brain disorder? Can it improve the performance of the proposed machine learning or deep learning-based method? This question is explored in Chapter 7.

2. Can we propose novel machine learning methods for evaluation of functional connectivity that can yield better performance as compared to the state-of-the-art? This is addressed in Chapters 5 and 6.

3. Do the non-imaging features (such as age, gender) carry important information for prediction of a brain disorder? This issue is highlighted in Chapters 5 and 6.

4. How can we build a deep neural network using functional connectivity to classify a brain disorder? This question is addressed in Chapter 7.

5. Can a convolutional neural network be used to map time-series functional MRI signals to features that can perform better? This question is addressed in Chapter 6 and 7.

In the next section, we present the contributions of the dissertation that target the aims and research questions mentioned above.

1.3 Original Contributions

In this dissertation, we have proposed machine learning methods for classification of a brain disorder based on functional MRI. The following are the main contributions of the work with respect to the aims and research questions presented in Section 1.2.

1. We have evaluated the importance of the non-imaging features for classification of a brain disorder. For this, the non-imaging features (such as age, gender) were integrated with imaging features in a machine learning model.

2. We have proposed affinity propagation clustering based novel method for estimation of functional connectivity.

3. A novel convolutional neural network architecture has been presented to calculate functional connectivity of brain regions. The convolutional neural network is being used to extract features from time-series signals of functional MRI data. These features are employed to calculate functional connectivity.

4. We have presented an innovative end-to-end trainable deep network for classification of a brain disorder. The convolutional neural network model incorporates functional connectivity. We have also discussed the importance of functional connectivity in this deep network. Our work is the first to propose an end-to-end network, incorporating functional connectivity for classification of a brain disorder.

1.4 List of Publications

A number of articles have been published and under-review in several workshops, conferences and journals during the course of the research work. Parts of this dissertation are based on some of these:

1.4.1 Journals

- Atif Riaz, Muhammad Asad, Eduardo Alonso, and Greg Slabaugh, “Fusion of fMRI and Non-Imaging Data for ADHD Classification”, Computerized Medical Imaging and Graphics Volume 65, April 2018, Pages 115-128. Chapter 5 is related to this publication.

- Atif Riaz, Muhammad Asad, Eduardo Alonso, Greg Slabaugh, “DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI", Journal of Neuroscience Methods. The publication is related to the chapter 7.

1.4.2 Conferences

- Atif Riaz, Muhammad Asad, S M Masudur Rahman Al Arif, Eduardo Alonso, Danai Dima, Philip Corr and Greg Slabaugh, “Deep FMRI: An end-to-end deep network for classification of fMRI data”, IEEE International Symposium on Biomedical Imaging (ISBI), 2018. The publication is related to the chapter 7.

- Atif Riaz, Muhammad Asad, S M Masudur Rahman Al Arif, Eduardo Alonso, Danai Dima, Philip Corr and Greg Slabaugh, “FCNet: A Convolutional Neural Network for

Calculating Functional Connectivity from functional MRI", 1st International Workshop on Connectomics in NeuroImaging (CNI), MICCAI 2017, Proceedings (Vol. 10511, p. 70). Springer. Chapter 6 is related to this publication.

- Atif Riaz, Eduardo Alonso, Greg Slabaugh, "Phenotypic Integrated Framework for Classification of ADHD using fMRI", International Conference on Image Analysis and Recognition (ICIAR) 2016, Pages 217-225, Springer. The publication is related to the chapter 5.

1.5 Dissertation Outline

This dissertation is structured as follows: Chapter 2 provides a comprehensive clinical background, a brief overview of some of the neuroimaging modalities explored by brain studies and a brief description of the key concept of functional connectivity. Chapter 3 provides a brief overview of some fundamentals of machine learning and deep learning. Chapter 4 provides a comprehensive literature review. Chapter 5 highlights the importance of integration of non-imaging features with imaging features for the classification of ADHD. In this chapter, a clustering based novel method is presented to calculate functional connectivity. Chapter 6 presents a convolutional neural network based novel method to calculate functional connectivity. Chapter 7 provides the details of the innovative end-to-end trainable network for classification of ADHD. This leads to the conclusion of the dissertation in Chapter 8 where we discuss the limitations of the current framework, possible improvements and direction towards future research on the topic.

Chapter 2

Clinical Background

In this chapter, we start by describing the brain structure and some of the commonly used imaging modalities to study brain function, followed by the description of functional MRI. In the second section, we describe functional connectivity, which is the key concept in functional MRI studies.

2.1 Brain Structure

The human brain is considered the most important part of the body controlling all actions, emotions, feelings and responding to all body events. The brain is composed of glial cells, specialized cells called neurons and blood vessels. Glial cells provide structural and metabolic support to neuron cells. Neuron cells are the central processing units of the brain and there are about 86 billion neurons [16] in a brain. The neurons process and transmit information through electrical and chemical signals.

A neuron cell (Figure 2.1) is comprised of synapse, axon, cell body and dendrites. The neuron receives information from other cells through a number of dendrites. The information is passed and processed in the axon and finally transmitted to other cells through the synapses. Neurons can have over 1000 dendrite branches, allowing connections with thousands of other neurons.

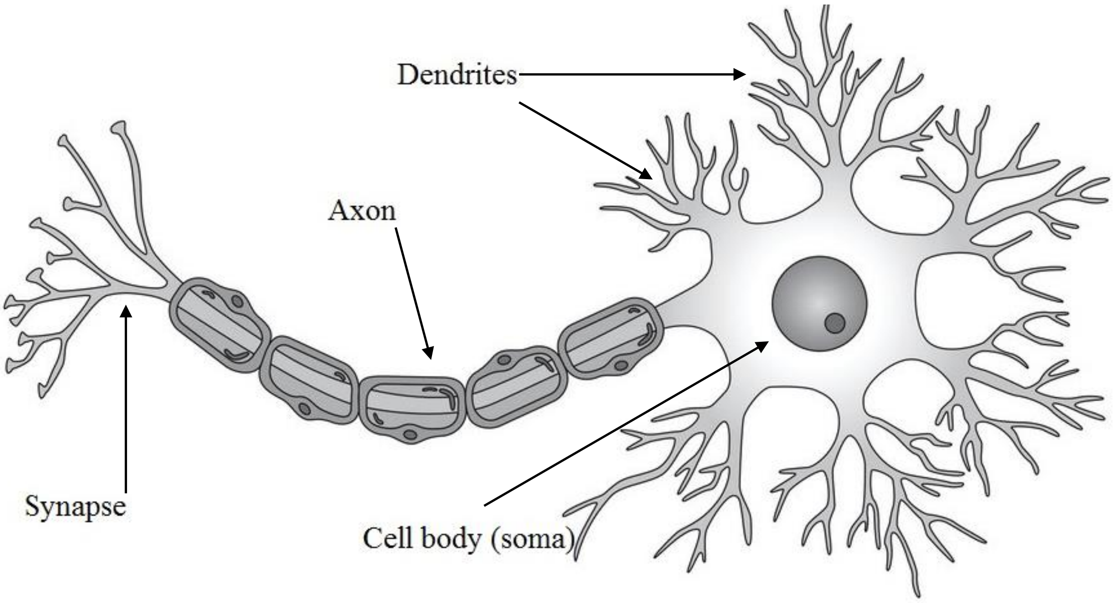


Fig. 2.1 Schematic view of a neuron.

The human brain matter can be divided into two parts, i) white matter, that contains the nerve fibres and ii) the grey matter, that contains neural cells. The brain surface is called cerebral cortex and is highly folded. This special folded structure allows a large surface area to fit into a fixed available brain volume. A cortical fold is termed as sulcus and the area between two sulci is termed as gyrus. The left and right hemispheres of the brain are similar in structural shape and most cortical areas show similarity in both hemispheres. However, both hemispheres are not essentially the same with respect to their functionality. How the different anatomical brain regions and parts are related to a particular cognitive function is one of the oldest debates in the field of neuroscience and is being explored widely by the research community. Different imaging modalities are being developed and used to study the brain in both healthy and disorder conditions. Studies have been using these imaging modalities to i) study the correlation of anatomical brain regions to some particular cognitive task and ii) study the effect of a specific brain disorder in brain regions and their functionalities. A few of the imaging modalities used to study brain functionalities are listed below.

2.2 Functional Brain Imaging Modalities

With advances in imaging technologies, different neuroimaging modalities have emerged like positron emission tomography, electroencephalograms, magnetic resonance imaging and functional MRI. This chapter provides a brief background about these neuroimaging modalities with an emphasis on functional MRI (fMRI).

2.2.1 Electroencephalogram (EEG)

Electroencephalogram (EEG) is one of the most common imaging modalities used for studying brain functional activity. EEG records the electrical activity along the scalp. The EEG measure represents the synchronous activity of a number of neurons that have a similar spatial orientation. If the neurons do not have a similar orientation, they do not create a detectable signal. In an EEG scan, small sensors are attached to the scalp which record the electrical signals that brain cells send to each other. The temporal resolution of EEG is high (on the order of milliseconds), however, its spatial resolution is low (typically around 1cm). One main disadvantage of EEG is that these sensors can only measure signals on the surface of the head and can not capture signals from all brain.

2.2.2 Positron Emission Tomography (PET)

Positron Emission Tomography (PET) is an imaging technique that uses radioactive material as a source to map the functional activity of the brain. The subject is first injected with a radioactive tracer isotope and a scanner is used to record radioactive emission of the tracer. The scanning technology is based on the assumption that areas of comparatively high radioactive emission are associated with brain functional activity. PET can detect glucose intake rates, thereby providing indirect measurement of brain functional activity. PET data has a high spatial resolution (typically around 4mm), however it has a poor temporal resolution (around 30-40 seconds). Apart from poor temporal resolution, the invasive nature of PET is considered another major shortcoming of the method.

2.2.3 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) scan is well suited for soft tissues, especially the brain. It uses a strong magnetic field to image the anatomy of the brain. An MRI scanner is comprised of a cylindrical tube with a very powerful electric magnet which creates high strength of magnetic field such as 3.0 Tesla. Normally, without any external magnetic field, the atomic nuclei of body cells are randomly oriented. Due to the effect of this powerful external magnetic field, these nuclei get aligned with the external field. When the external magnetic field is removed, nuclei are re-aligned to their original states, causing the release of radio frequency signals. The tiny signals from multiple nuclei coherently add to form combined signals which are measured by the detector and form an image. MRI can reveal the anatomical details of the brain, but it lacks the ability to measure functional activity of the brain.

2.2.4 functional MRI (fMRI)

Functional MRI (fMRI) is a neuroimaging modality that provides an opportunity to study the functional activity of the whole brain. fMRI has evolved as a popular neuroimaging modality with main advantages being: it is non-invasive, it avoids harmful radiation to the subject being scanned and it can record signals from all brain regions.

Different imaging modalities have different trade-offs in terms of spatial and temporal resolution. For example, EEG provides a higher temporal resolution, however, its spatial resolution is very low. While fMRI provides high spatial resolution, its temporal resolution is low (around one second). Due to its better spatial resolution, fMRI is considered a preferred approach to study certain neurological disorders.

During the last two decades, fMRI has developed as one of the most common and prominent methods used for functional brain imaging [17, 18] and it is increasingly employed to study the functional activity of brain regions and networks. It is accepted that fMRI may help in developing an objective diagnostic tool for brain disorders, particularly, in identifying biomarkers that can distinguish between healthy subjects and subjects with brain disorders.

2.3 Functional MRI and BOLD signal

fMRI is a non-invasive technique that primarily measures the oxygen contrast in the blood flow. This is known as the blood oxygen level-dependent (BOLD) contrast and is used to explore brain functional activity. There is a special protein present in blood cells called hemoglobin. The primary purpose of hemoglobin is to transport oxygen to different body cells. Hemoglobin can bind with oxygen to carry it to cells and it can also detach the oxygen wherever it is required. Therefore, hemoglobin can have two forms i) oxyhemoglobin, where it is attached to oxygen and causes the red colour of the blood, and ii) deoxyhemoglobin, where oxyhemoglobin releases its oxygen. The basic principles of BOLD contrast are i) oxygenated and deoxygenated hemoglobin have different magnetic properties (oxygenated haemoglobin is diamagnetic, while deoxygenated haemoglobin is paramagnetic [19, 20]) ii) blood oxygenation level of a particular region varies according to the strength of the regional neural activity. These properties can be used to indirectly assess brain functional activity [21].

The brain does not store glucose which is considered as a primary source of energy for body cells. Due to this fact, the brain requires a continuous supply of glucose for the functionality of different brain regions. When certain neurons become active, they require more energy which is mainly produced from glucose. More blood flows to transport more glucose and thereby brings in more oxygen in the form of oxygenated hemoglobin molecules in the blood cells. This causes an increase of the oxygenated blood inflow. Figure 2.2 shows an illustration of the process. As more oxygenated blood flows in, it pushes away a portion of deoxygenated blood in the venous capillaries, causing a higher concentration of oxygenated hemoglobin as compared to deoxygenated hemoglobin. The contrast in the oxygenated and deoxygenated hemoglobin is picked up by the BOLD signals due to the difference in their magnetic properties. Therefore, changes in the BOLD signal can be used to identify areas of increased or decreased neuronal functional activity [22, 23].

fMRI is an extended form of MRI where MRI volumes are acquired in multiple time points and these three-dimensional volumes are stacked together to generate fMRI data that

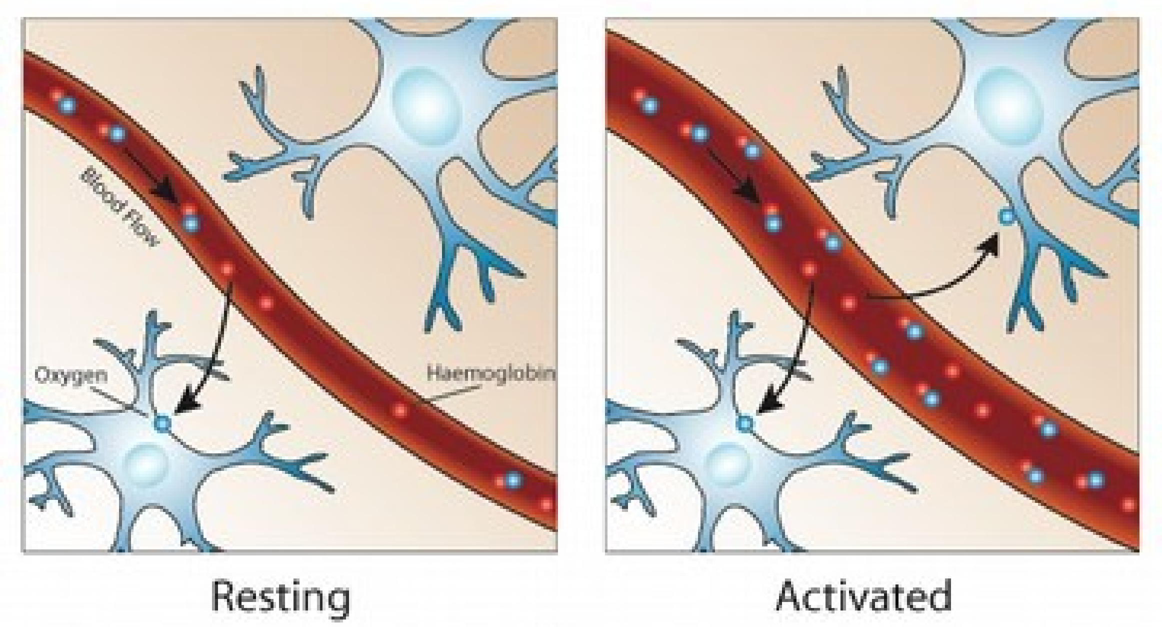


Fig. 2.2 Illustration of blood flow for brain cells. The left image depicts resting (normal) state and the right image shows an activated state. Thickness of the vessel represents the blood volume. A thick volume in the right image indicates high blood volume. The activated region (right image) is supplied with more oxygenated hemoglobin as compared with the resting state (image from www.sbirc.ed.ac.uk).

contains the three-dimensional brain volume for each time point as illustrated in Figure 2.3. The temporal resolution of an fMRI scanner is presented by the repetition time (TR).

fMRI studies can be grouped into two categories: task-based and resting-state. In a standard task-based fMRI, the subject is presented with a specific task of interest. The task depends on the nature of the experiment being conducted and may include responding to a stimulus, solve some arithmetic operation, moving the limb etc. The BOLD signal during the experimental task is compared to the BOLD signal during the control condition (rest) [17]. In the early era of fMRI, studies mainly focused on task-based fMRI [17]. In resting-state, the fMRI scan is recorded while the subject is not performing any specific task. The approach is task-independent and is known as resting-state fMRI (rs-fMRI).

Data generated by an fMRI scanner is referred to as raw data and is typically corrupted with noise. For the last two decades, a number of studies have put efforts to characterize and mitigate the effects of noise in fMRI signals [24, 25]. The improvements in methods to

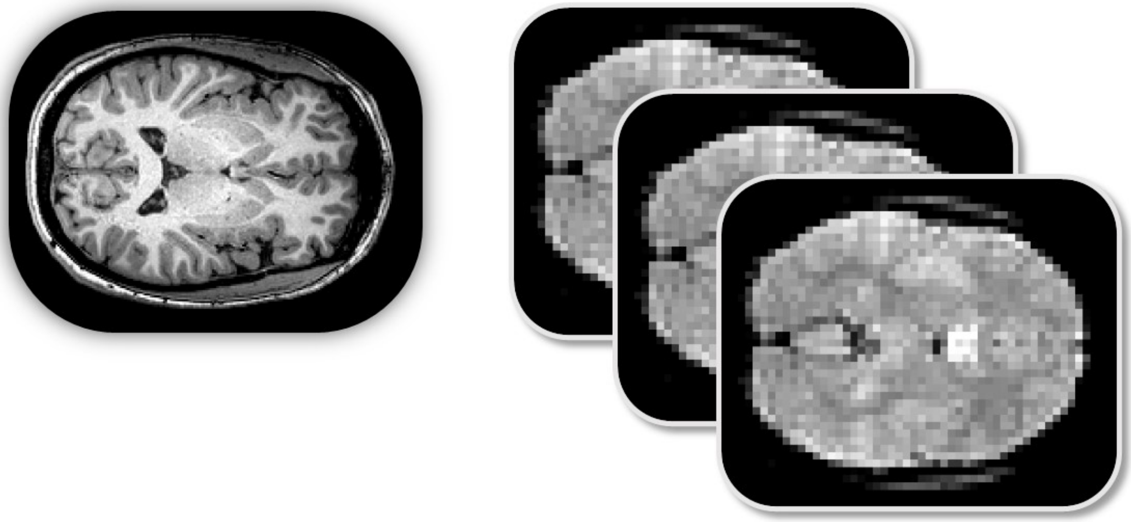


Fig. 2.3 MRI and fMRI scan. The left image is an MRI scan and the right is representing an fMRI scan. In this figure, there are three time points of fMRI data and each fMRI time point comprises a MRI scan. The scans are stacked together to constitute fMRI data.

distinguish signals from noise have advanced brain activity detection. The noise in an fMRI scan can be grouped into two broad categories, i) background noise and ii) physiological processes.

Background noise is characterized by the contributions of the sources that are independent of the signal of interest [26]. The main source of noise lying in this category is the thermal noise induced by the MRI apparatus. With an increase in the temperature, heat may attenuate electron movement which may distort the current in the fMRI detector. Radiofrequency (RF) noise is another contributor to background noise [27]. The presence of RF noise is considered undesirable in fMRI analysis and an important engineering effort is being invested in minimizing this noise. Thermal noise is unavoidable, however, its effects can be reduced through filtering. To achieve this, low pass filtering is applied to eliminate noise from the frequencies that lie outside of the signal band of interest.

Physiological processes include subject motion, cardiac pulsation and respiratory activity [26]. The main source of noise in this category is the subject motion, particularly head movement. During the scan, it is very difficult to maintain the head in a fixed position for the whole course of fMRI acquisition. Even with a carefully administered setup, head motion is unavoidable. In some cases, noise may be introduced due to physical responses to the experiments (for example, providing feedback through some buttons). Even with prior training of the subjects, certain inherent factors like random thoughts, scanner noise, different physical sensations, etc. may introduce noise in captured data. Such noise can affect the analysis of fMRI data. For example, head motion during the scan can change the signal to region mapping. In order to mitigate the effects of noise, preprocessing of the raw data is essential prior to analysis. The main steps involved in preprocessing are listed below.

Slice Timing Correction

Most fMRI scanners use a two-dimensional pulse sequence that scans one three-dimensional volume by acquiring multiple two-dimensional images at a time and combining them to form a three-dimensional volume. Each new volume is acquired after every acquisition time (TR), typically 2-3 seconds. The individual two-dimensional slices are acquired during this time sequentially or mostly in an interleaved manner (all odd numbered slices are scanned first, followed by the even numbered slices). For sequential scanning, the last slice is acquired almost one TR after the first slice and in the case of interleaved scanning, adjacent slices are acquired almost TR/2 time apart. This causes inconsistent acquisition time among different slices within a volume. Slice timing error may introduce severe distortion in the analysis of fMRI data. Temporal interpolation is a widely used approach to address this source of error. The slice timing correction technique estimates the signal amplitude of a slice by interpolating between the same slice and neighbour TRs to estimate the signal that would have been acquired at the same time.

Realignment

Head motion can introduce strong artefacts in fMRI data and is a prominent concern in most fMRI studies. Even with a carefully administered experimental setup, subjects may show displacements of head up to several millimetres which may have detrimental effects on the results. Head motion errors can be suppressed by during-scan or post-scan techniques. During scan acquisition, head immobilization techniques may be used. These include fixation pads, masks, inflatable airbags, use of a mock scanner, which can minimize head motion during the scan. Even with these during-scan techniques, head motion is inevitable. Therefore, post-scan techniques are mandatory to avoid errors due to head motion. The objective of post-scan realignment techniques is to minimize the artefacts due to head motion and to determine rigid body transformations that best map functional images to a common space. A rigid body transformation is applied to each three-dimensional volume and can be parametrized by three translations and three rotational parameters. Depending on the algorithm implemented, a particular scan (first, last or middle) is taken as the reference scan and all other images are aligned to this reference scan. Realignment involves the optimization of the six parameters that minimize the mean squared difference between the reference scan and all other scans.

Coregistration

The acquired stack of three-dimensional functional images scans and anatomical scans generally do not match due to the difference in MR contrasts and acquisitions. This may cause errors in mapping the activity from the functional data to the anatomical images. The computation techniques that map images of different modalities (structural and functional) of each subject are called coregistration or functional-structural coregistration. Coregistration allows functional data to be superimposed on anatomical images for clear visualization. Also, spatial normalisation (next step) is more precise when warps are calculated from high spatial resolution structural images as compared to the functional images. These techniques typically map anatomical data to the spatial resolution of functional data (typically mean functional image) as a first step and, similarly to realignment, perform a rigid body transformation parametrized by translation and rotation parameters, where a cost function is minimized [28].

Spatial Normalisation

In most fMRI studies, it is required to aggregate, compare and analyse brain functional activity across multiple subjects. The shape and size of the brains are not consistent across multiple subjects. If this issue is not addressed, a voxel belonging to a particular subject may correspond to a different voxel across subjects. To address this problem, a standard approach is to normalize each brain scan to a common template space. This technique enables the data to be compared across different subjects. The common template may be estimated from a specific population [29] or a standard defined template. Most commonly used standard templates include Talairach [30] and Montreal Neurological Institute (MNI) template space.

Spatial Smoothing

fMRI data is inherently spatially correlated, this is due to the fact that adjacent brain voxels tend to show functional similarity [31]. Therefore, spatial smoothing can suppress noise sources uncorrelated among adjacent pixels. A typical implementation of spatial smoothing includes convolving the data with a Gaussian kernel that matches the spatial correlation of fMRI data. The main advantages of spatial smoothing are i) it increases the functional signal to noise ratio of data by suppressing the noise [31, 32], and ii) it reduces anatomical or functional variations among different subjects.

SPM [33], FSL [34] and DPARSFA[35] are popular tools that are usually employed for preprocessing of fMRI data. DPARSFA is based on SPM and uses its functionality to build a preprocessing pipeline in a user-friendly manner.

In the next section, we will introduce an important concept called functional connectivity which is widely used in fMRI data analysis and in this dissertation.

2.4 Functional Connectivity

Functional connectivity (FC) can be defined as the temporal coherence between spatially remote neurophysiological events [36]. In the context of functional imaging studies, FC describes the relationship of functional activity patterns of anatomically separated brain regions, thereby reflecting functional communication between brain regions [37].

Brain regions that show coherent functional activity patterns are assumed to be functionally connected. Figures 2.4 and 2.5 represent the functional activity of brain regions. In Figure 2.4, two brain regions show coherent functional activity, so these regions can be assumed to be functionally connected, while in Figure 2.5 the functional activity of the brain regions is not coherent, so these regions are not functionally connected.

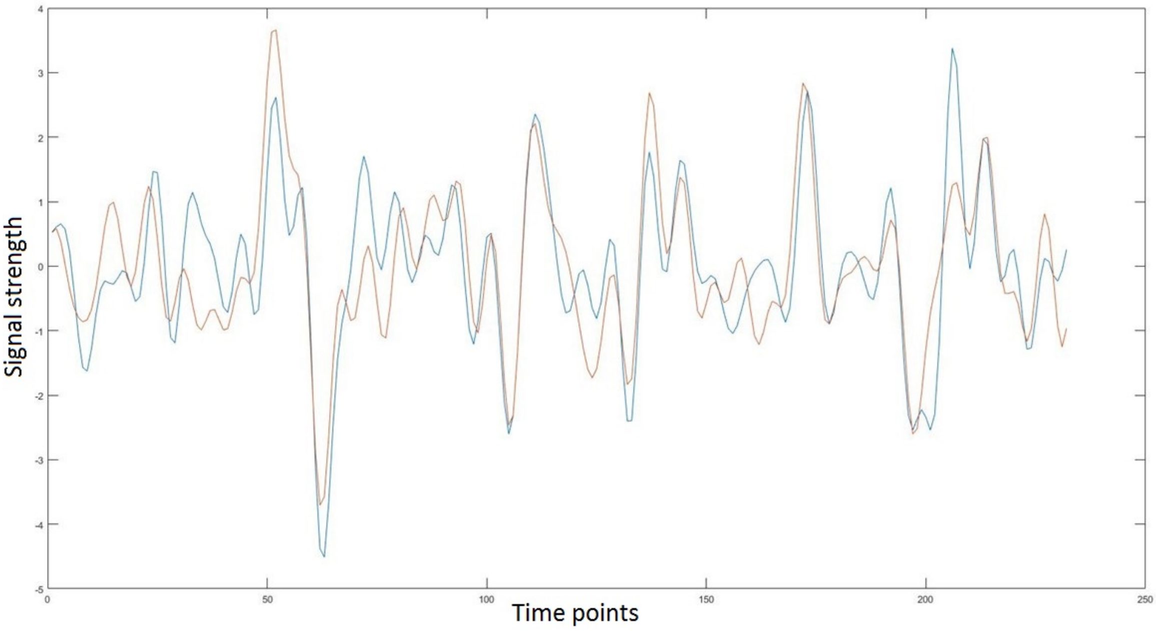


Fig. 2.4 Functional activity of two brain regions. Similar functional activity represents that the two regions are functionally connected.

The human brain can be envisioned as a complex, hierarchical network where each brain region can be represented as a node of the network and an edge represents the connectivity between brain regions. These networks continuously coordinate with each other forming a hierarchy of efficient networks. Each brain network may be responsible for a particular activity and the connections between networks may represent the coordinated activities of different systems of the body. A typical brain network is illustrated in Figure 2.6, where sub-networks of the brain and interaction between sub-networks are visualized.

The first resting state fMRI study was conducted by Biswal et al. [38], in which the authors found coherent functional activity between the left and right hemispheric regions belonging to the primary motor cortex of the brain during rest. Their findings suggested possible functional connectivity between the regions of the motor cortex during the rest. After this pioneering work, many studies explored functional connectivity and found that different brain regions show functional connectivity during the resting state of the brain.

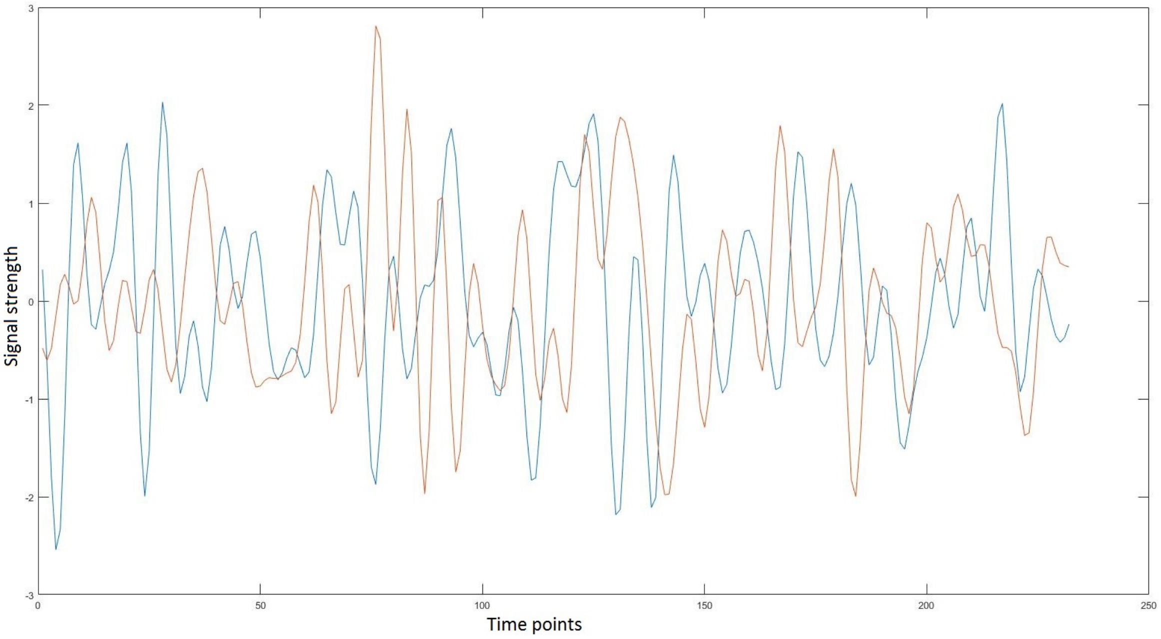


Fig. 2.5 Functional activity of two brain regions. The functional activity is different for two regions so the regions are not functionally connected.

The connectivity of brain regions is considered very important for monitoring and regulating bodily functions. In the case of a brain disorder, the normal functionality of the brain is altered resulting in different connectivity patterns or the creation of new abnormal connectivity patterns. This alteration in connectivity patterns is illustrated in Figure 2.7. The figure depicts a comparison between a healthy brain and a disorder-affected brain. The right image in Figure 2.7 illustrates a disorder-affected brain where the connectivity between regions A, B and C is altered (represented by red colour), and new connections also appear, so that regions belonging to these abnormally created links are “miswired".

Functional connectivity of all brain regions can be represented as a matrix, where each column and row score represents brain regions and values score the connectivity strength between any two regions. Such a matrix is represented in Figure 2.8, for a brain parcellated into 90 regions.

A number of studies have explored functional connectivity of brain regions in different neurological disorders such as schizophrenia, epilepsy [1, 39], ADHD [5, 7, 40] and have

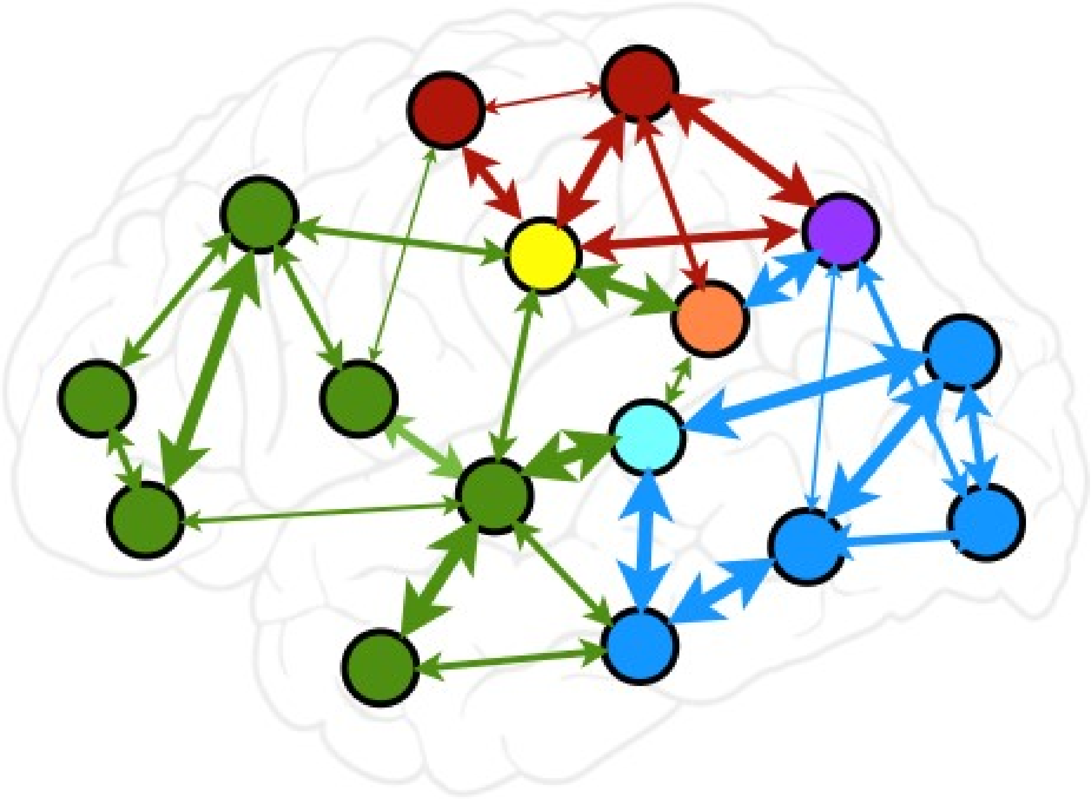


Fig. 2.6 Illustration of the brain network. Multiple sub networks are shown with different colours. There are three main networks presented with green, red and blue colours and these networks are also interconnected with each other.

shown that certain disorders affect functional connectivity. Such functional connectivity alterations are visualized in Figure 2.9.

In this dissertation, we have explored functional connectivity to study the Attention Deficit Hyperactivity Disorder (ADHD). We were interested in exploring novel methods to evaluate resting state functional connectivity of brain regions in healthy and ADHD.

In this chapter we have provided some important clinical concepts and background. In the next chapter we will describe the technical background related to this dissertation.

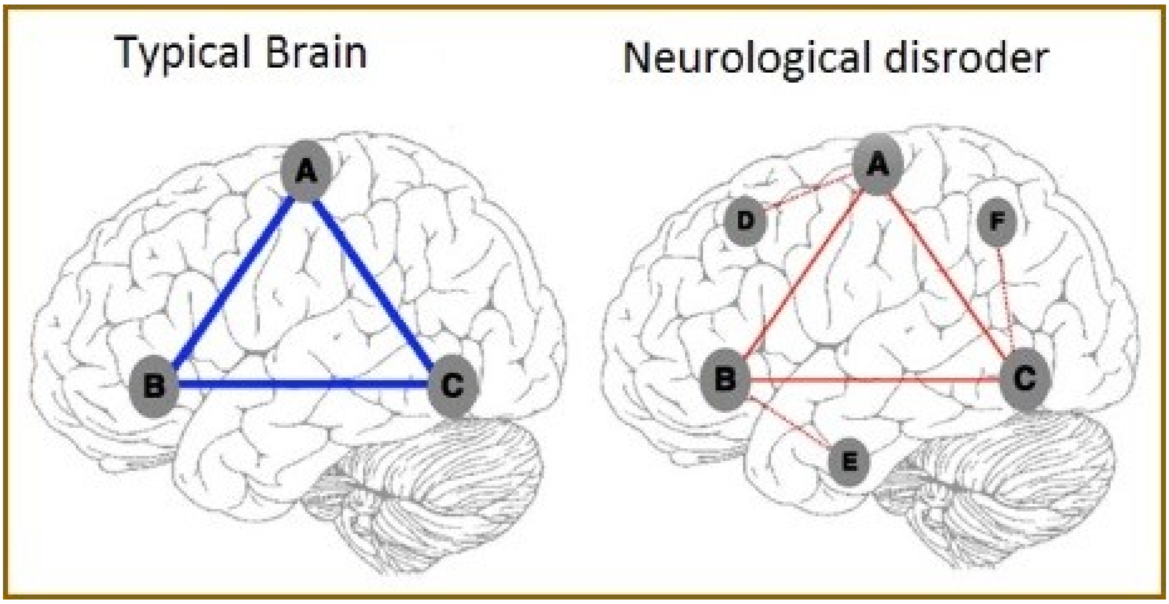


Fig. 2.7 The left image illustrates a healthy brain where regions labelled A, B and C are normally connected. The right image shows a brain with a disorder where the connectivity between the regions has been altered (connectivity may be increased or decreased) which is shown in a red colour. There are also some new abnormal connections to regions D, E and F.

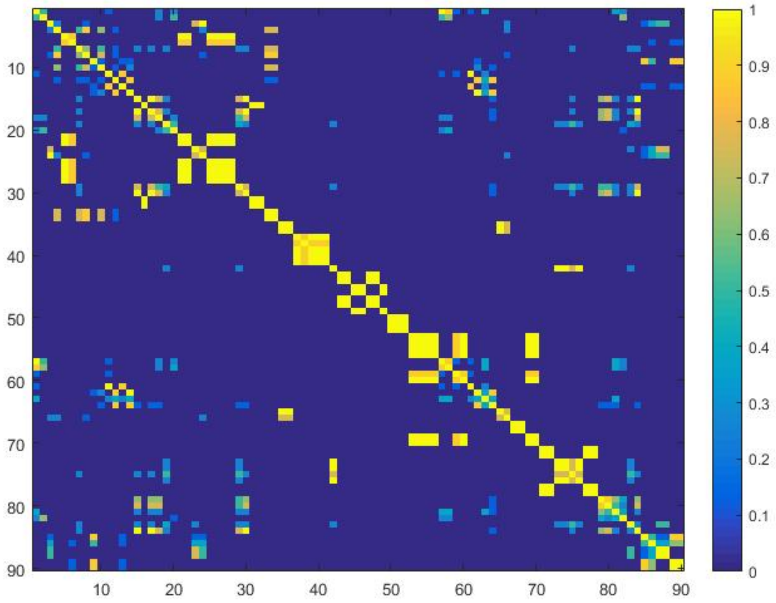


Fig. 2.8 Matrix representing functional connectivity of brain regions. Rows and columns represent the brain regions, in this case there are 90 brain regions so the matrix has dimensions of 90 × 90. The values represent strength of the functional connectivity, where a value close to one represents functionally connected regions.

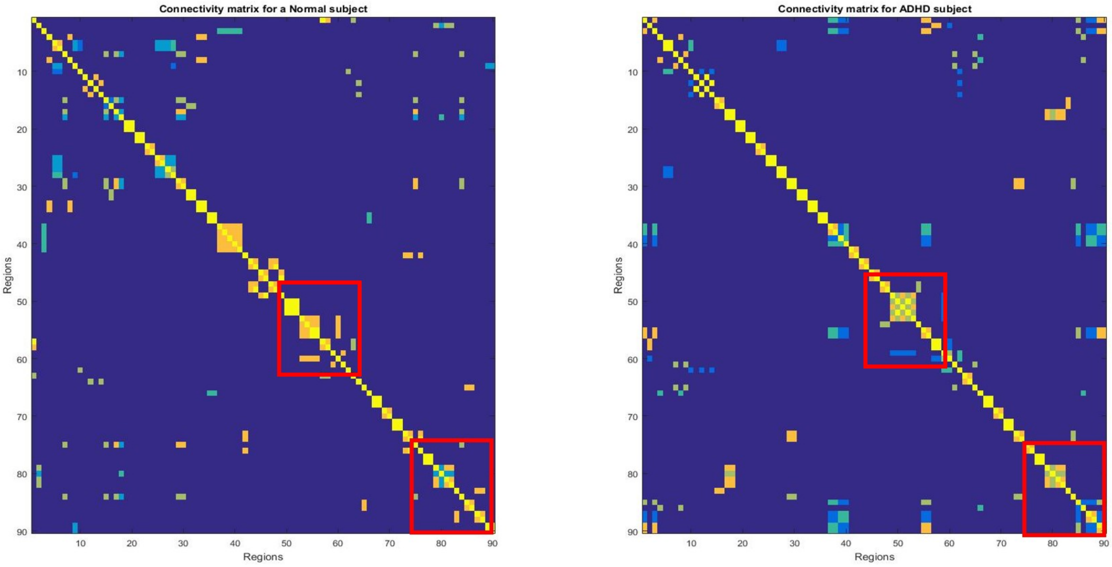


Fig. 2.9 Matrix representing functional connectivity alterations in ADHD. The left matrix is the functional connectivity matrix for a healthy subject and the right matrix is the functional connectivity matrix for an ADHD subject. Red boxes highlight some of the alterations.

Chapter 3

Background

This chapter is divided into three sections. We start by describing basic machine learning techniques. The second section highlights some important concepts of deep learning. In the final section of this chapter, we present a literature review of the state-of-the-art research in the field of brain disorder classification.

3.1 Machine Learning

Machine learning has evolved as a powerful tool in the domain of computer science. Machine learning refers to a set of algorithms that enable a machine to learn from the available data without being explicitly programmed.

Machine learning can be divided in to two broad categories: i) supervised learning and ii) unsupervised learning. These are described below.

3.1.1 Supervised Learning

Supervised learning can be presented as learning a mapping between input data and its labels such that the algorithm can predict labels when presented with unseen data [41]. The inference is based on the assumption that the label is not a random value, rather that there exists a relationship between the input data and the label. During the training phase, training data $\{x,y\}$ is presented to the machine learning algorithm, where x is the data and y is the

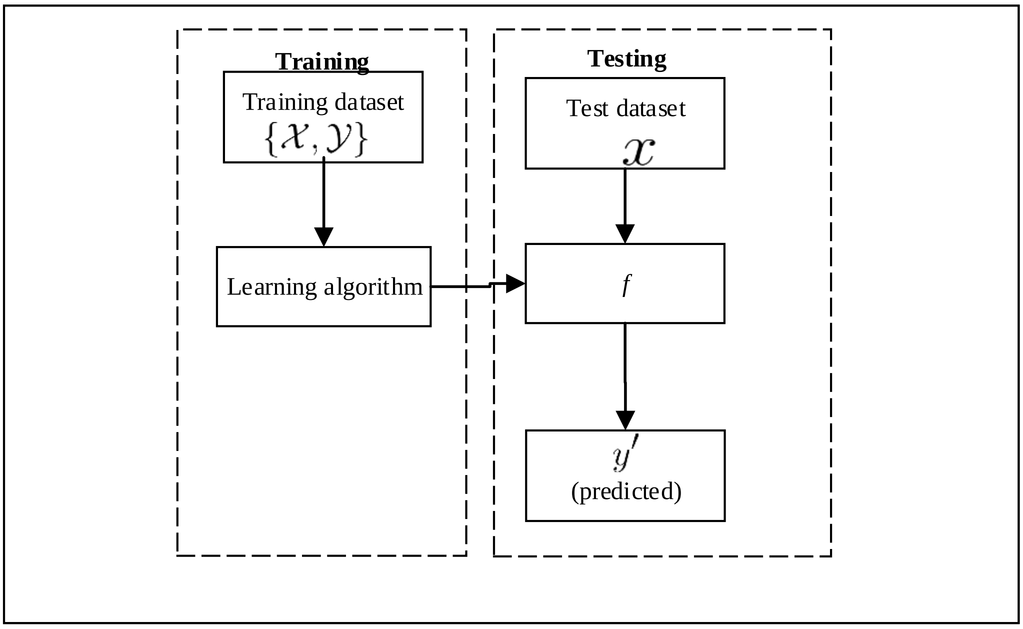


Fig. 3.1 Supervised machine learning.

class label (for instance, healthy or ADHD). The algorithm learns the mapping of x onto y as a function $f : x \to y$ from the training data. During the testing phase, unseen testing data x is presented to the learned machine learning algorithm. The algorithm predicts the label y. The process is presented in Figure 3.1.

Depending on the type of the predicted target label, supervised learning can be sub-divided into two categories, regression and classification.

Regression

In the case where the output label y is a continuous value, supervised learning is categorised as regression. For example, if the target is to predict the ADHD rating score from fMRI data. The ADHD rating score represents the severity of ADHD and is a continuous value. This is characterised as a regression problem.

Classification

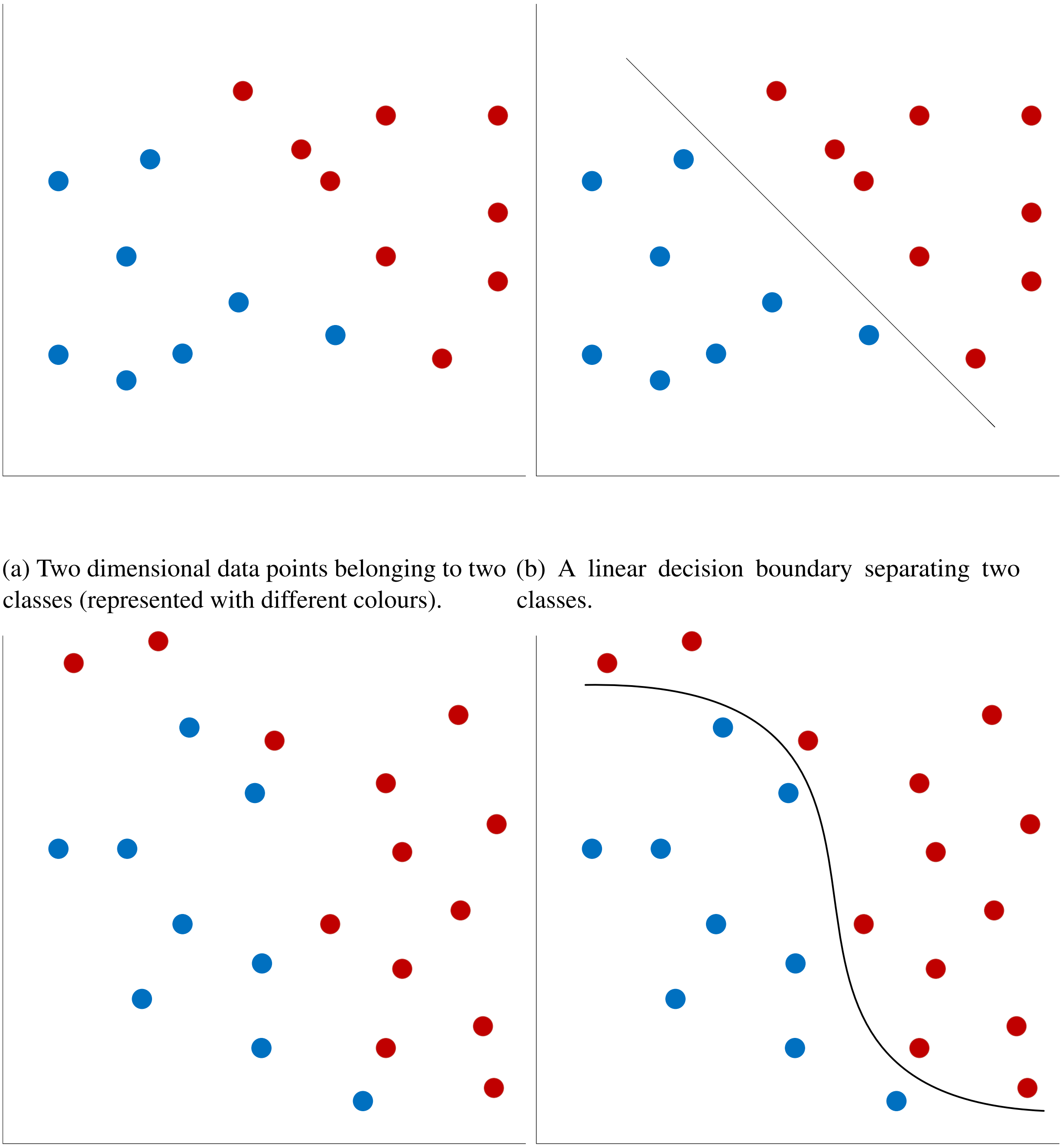
If the target label to predict is categorical data, supervised learning is classed as a classification problem. For example, predicting whether fMRI scan data belongs to a healthy individual or an individual with brain disorder, is categorised as a classification problem. The output is categorical (healthy or ADHD). Some notable classification algorithms include neural networks, random forest and support vector machine.

The goal of the classification algorithm is to learn the decision boundary that can discriminate between the different classes. The classifier learns this decision boundary during the training phase. The learnt decision boundary is used to make predictions for unseen data. Classification approaches can be linear or non-linear. In the linear classification problem, the classifier learns the mapping function as a linear combination of the input features. A non-linear algorithm learns a complex non-linear mapping of input features. Figure 3.2 illustrates the concept of linear and non-linear classification problems using abstract examples of two-dimensional data points. Figure 3.2 (a) shows a linear classification problem where a classifier is able to learn the linear combination of input features which is presented as a linear decision boundary in Figure 3.2 (b). Figure 3.2 (c) shows a non-linear classificaiton problem between two classes where the classifier is able to learn the non-linear mapping which is presented as a non-linear decision boundary in Figure 3.2 (d).

In this dissertation, we have addressed a supervised machine learning problem where we have input data (pre-processed fMRI data) and corresponding labels (healthy or ADHD). The machine learning model learns the mapping between the data and labels during the training phase. During the testing phase, previously unseen fMRI data is presented to the learned classifier, and the classifier predicts the label (healthy or ADHD) of the data.

3.1.2 Unsupervised Learning

Unsupervised machine learning deals with data for which output labels are not available. In these problems, machine learning algorithms learn the underlying characteristics of the



(a) Two dimensional data points belonging to two classes (represented with different colours).

(b) A linear decision boundary separating two classes.

(c) Two dimensional data points belonging to two classes (represented with different colours).

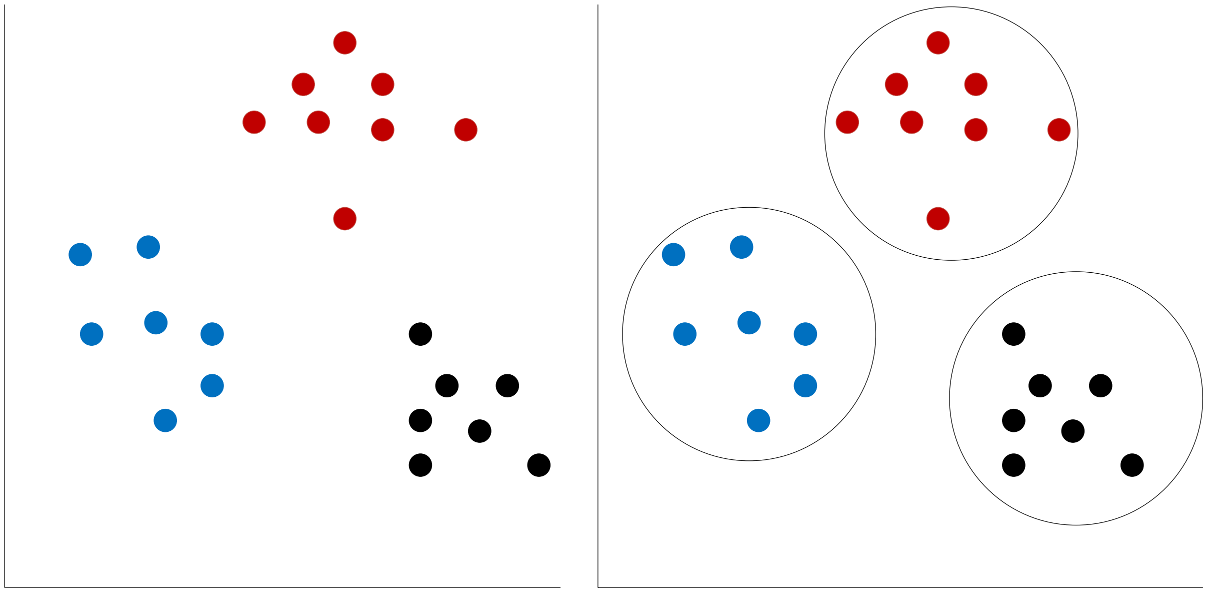
(d) Non-linear decision boundary separating two classes.

Fig. 3.2 Decision boundaries for linear and non-linear separable classes.

data by themselves. Clustering is one of the popular examples of unsupervised learning algorithms.

Clustering

Clustering is one of the classical problems in computer science. It has been extensively explored and has been applied to a wide range of application domains. Clustering can be defined as: given a set of objects with some defined characteristics (features or attributes), the goal is to group them in a meaningful way. The final emerged groups are called clusters. The ultimate goal of any clustering algorithm is to group the objects into clusters such that their inter-cluster differences are maximized while the intra-cluster differences are minimized. Figure 3.3 illustrates the clustering of abstract two-dimensional data.



(a) Two-dimensional data. (b) Data grouped into three different clusters.

Fig. 3.3 Illustration of clustering of two-dimensional data.

Clustering of fMRI data

In the case of fMRI data, a clustering algorithm will group brain regions into different clusters such that the regions lying in one cluster will have similar functional activity, which should be different from regions outside of this cluster. Figure 3.4 and Figure 3.5 present clustering results of fMRI data. Figure 3.4 represents two regions that are grouped into one cluster, similarly Figure 3.5 represents four regions grouped into one cluster.

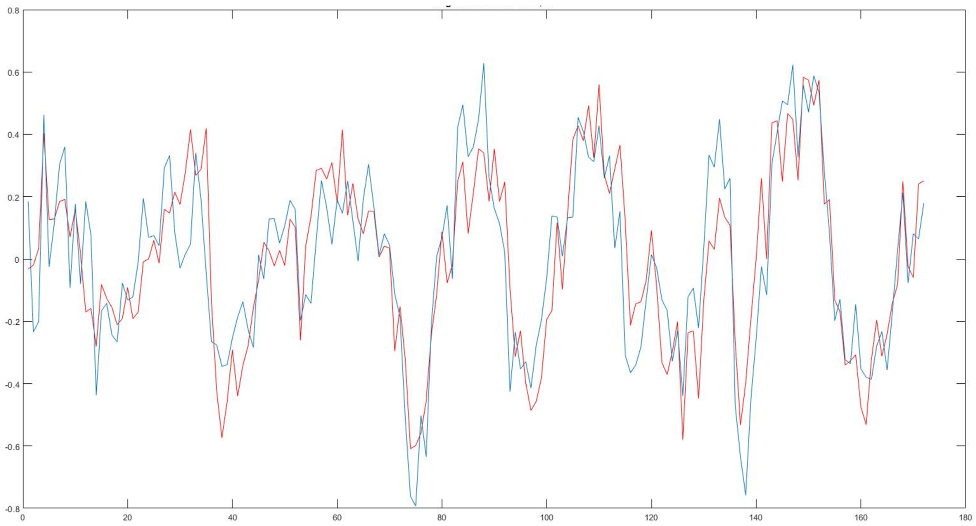


Fig. 3.4 A cluster of fMRI data with two regions.

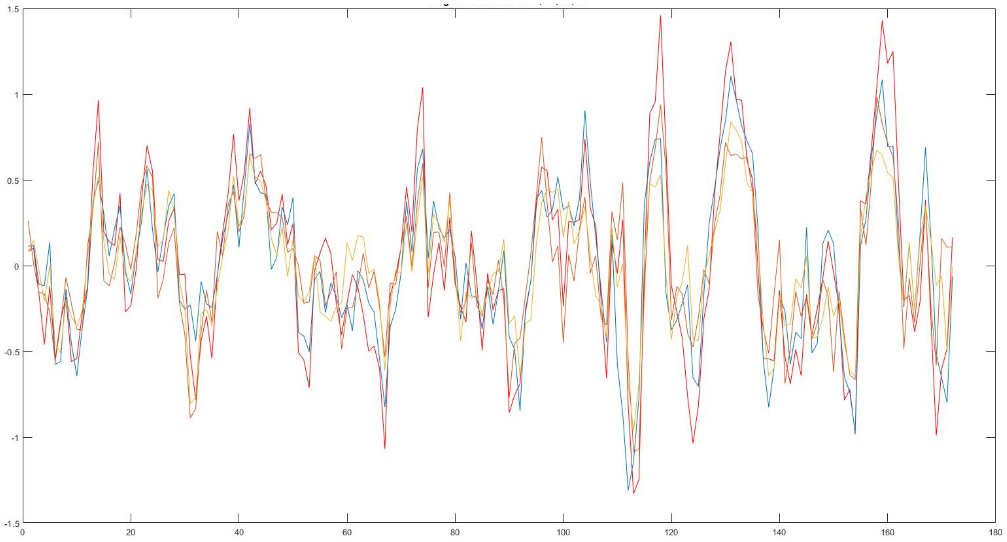


Fig. 3.5 A cluster of fMRI data with four brain regions.

3.1.3 Dimensionality Reduction

Handling high-dimensional data is considered a serious challenge in many machine learning applications [42] due to the so-called curse of dimensionality problem [43]. In most machine learning problems, large numbers of features can cause models to overfit [44, 45], thus degrading the performance of the machine learning model [46]. To overcome problems associated with high-dimensional data, various dimensonality reduction techniques have been explored. The motivation of dimensionality reduction techniques lies in the fact that out of all available features, only a subset of features is important and plays a key discriminant role towards the prediction problem. Dimensionality reduction techniques can be divided into two broad categories, i) feature extraction and ii) feature selection.

Feature Extraction

Feature extraction techniques project the original feature space into a new feature space with a lower dimension. The newly constructed features are usually a combination of the original features. Popular methods of this type include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA).

Feature Selection

Feature selection methods aim to select a subset of features which minimize redundancy and maximize the relevance to the target class. There is a wide range of algorithms which select a subset of features. Some examples of feature selection approaches include information gain, Fisher score, LASSO, and elastic nets.

Both feature extraction and feature selection methods improve the performance of a classifier in machine learning by reducing the number of available features for training. Feature extraction maps the original features into a new feature space. Detailed anlaysis of the mapped feature set is therefore considered challenging as it involves mapping from the original feature space to the new projected feature space [46]. In contrast, feature selection selects a subset of features from the original feature set without any transformation. Owing to this fact, feature selection is considered a better choice for dimensionality reduction in terms of interpretability [46].

A general classification model employing feature selection is presented in Figure 3.6. In the training phase, the subset of features is selected through the feature selection algorithm.

The selected features are presented to the classifier for training. In the testing phase, the same selected feature set is presented to the classifier for prediction. For example, consider the case of classification of a brain disorder from fMRI data. The first step in the training phase is generation of features, which in this case is calculation of functional connectivity. Next step is feature selection where discriminant features are selected from calculated functional connectivity. Finally, the selected features are presented to a classifier for training. In the testing phase, selected features from the test dataset are presented to a trained classifier for the final prediction. Feature selection can be divided into three main categories [44] which are presented below.

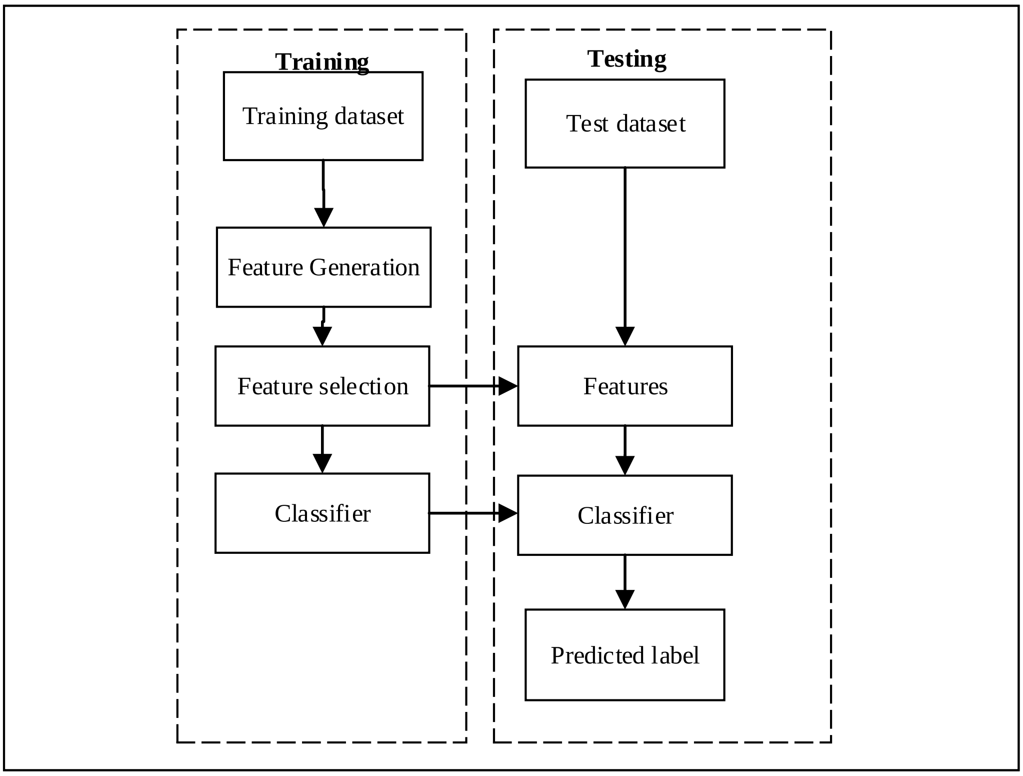


Fig. 3.6 A general process of classification with a feature selection algorithm.

Filter-based Methods

Filter-based methods evaluate the importance of features by relying on the characteristics of data, without utilizing any classification algorithm [42]. A typical filter based method is comprised of two steps. In the first step, these methods use simple statistical measures (e.g. mean, variance, correlation coefficients) to rank features according to their relevance in identifying class-level differences. Next, a subset of all features is selected based on a threshold value or the number of features. Filter-based methods are usually efficient to implement, however, they are independent of any classifier and ignore the effect of selected features on the performance of the classifier.

Wrapper-based Methods

Wrapper-based methods use an objective function from a classification machine learning model to rank the features with respect to their relevance to a particular classification problem. A wrapper-based method typically performs the following steps: i) it searches a subset of features, ii) evaluates the selected subset of features through the classifier, iii) repeats these steps until a required criterion is achieved or a maximum number of iterations is performed. The features with the best performance are selected. Based on the selection of a subset of features, these methods can be further sub-divided into two categories [47], i) forward selection and ii) backward elimination. In forward selection, the search of features begins with an empty feature set and features are added in iterative steps until an optimum set of features is found. In backward elimination, the search starts with all features and features are removed in iterative steps until an optimum subset of features is found. Wrapper-based methods typically yield better accuracy compared to the filter-based methods [47, 48] and allow feature dependency to be taken into account [49]. However, these methods tend to be very computationally extensive [46].

Embedded-based Methods

Embedded-based methods select discriminant features as a part of the machine learning process. These methods enforce certain penalties on a machine learning model and output a subset of relevant features. Such methods allow interaction with a classifier model but are less computationally expensive than wrapper-based methods [49]. Embedded-based methods incorporate statistical criteria, as the filter model does, to select feature subsets and choose the optimal subset of features with the highest classification accuracy. In other words, it achieves model fitting and feature selection simultaneously. Elastic Net [50] is a popular example of feature selection in this category and has been employed in this dissertation. The most important property of Elastic Net is that it encourages grouped selection of features, which makes it suitable for domains where input features might be correlated and all grouped features are required to be selected. Elastic Net is an embedded-based feature selection algorithm that takes advantages of both LASSO and ridge regressions. It combines penalties of both regression algorithms in one single solution. Similar to LASSO regression, it employs a L_1 penalty to enable variable selection and continuous shrinkage, and similar to ridge regression, a L_2 penalty is employed to encourage grouped selection of features.

3.2 Deep Learning

Recently, Deep Learning has emerged as a state-of-the-art tool in the domain of Artificial Intelligence (AI) and has outperformed other machine learning methods in a number of domains including computer vision [51, 52], natural language processing [53], semantic parsing [54] , transfer learning [55, 56] and many more. Deep Learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. Deep Learning typically uses an algorithm called back-propagation to learn how a machine should change its internal parameters that are used to compute the representations in each layer from the representation in the previous layer.

Deep Learning covers a range of artificial neural network-based algorithms. We will start by explaining a single perceptron which is essentially the basic unit of a neural network. Later on, we will introduce multi-layer perceptrons (MLP), convolutional neural networks (CNN) and fully connected networks.

3.2.1 Perceptron

A perceptron or artificial neuron is a basic processing unit of a deep learning network. In the work of Warren McCulloch and Walter Pitts [57], the authors proposed the idea of the perceptron for the first time. The authors presented an analogy between a biological neuron and simple logic gate with binary outputs. The perceptron can be viewed as a basic unit of a neural network in a biological brain. Multiple signals arrive at the dendrites of a neuron. All these inputs are accumulated in the cell body of the neuron. If the accumulated information exceeds a certain threshold, an output signal is produced which is transmitted to the next neuron through the axon.

Based on this basic working idea of the neuron, Frank Rosenblatt proposed the learning rule for the perceptron in the machine learning domain [58]. Each input is assigned a weight, and the inputs and their corresponding weights are multiplied together in order to determine whether a neuron fires or not, thus solving a binary classification problem. For an input vector $\boldsymbol{x}$, the network output is calculated as:

$$\hat{y} = f(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b) \tag{3.1}$$

where b is the bias, w is the weight and f is the activation function of the neuron. Both b and $\boldsymbol{w}$ are learned through training. The schematic of a neuron is presented in Figure 3.7.

A number of activation functions have been explored, where the most common is a sigmoid function which squashes any value in a range of zero to one. The sigmoid function is calculated as:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.2}$$

The sigmoid function is visualized in Figure 3.8.

If we know the ground truth (actual label) for the input vector $\boldsymbol{x}$, we can calculate the loss term L as:

$$L = \frac{1}{2}(y-\hat{y})^2, \tag{3.3}$$

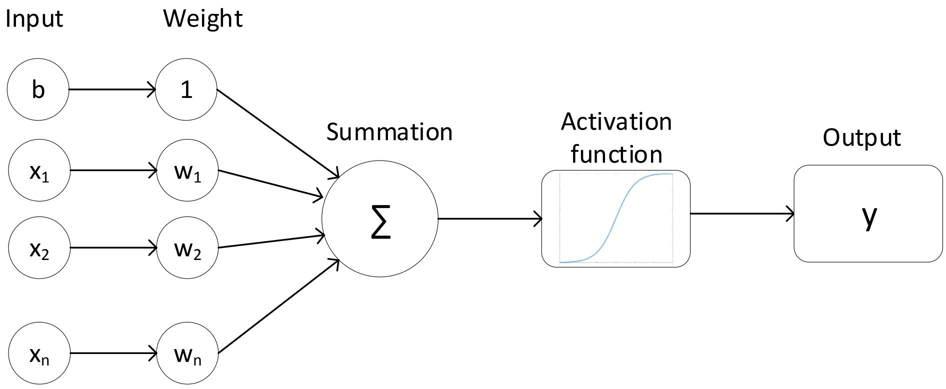


Fig. 3.7 Schematic of a neuron.

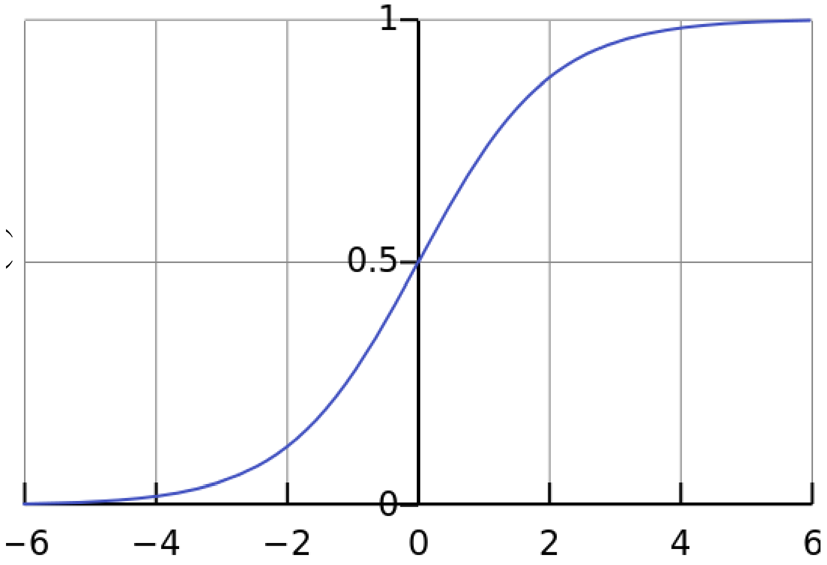


Fig. 3.8 Sigmoid function. The function squashes any value in a range of zero to one (y-axis).

where y is the actual label of the data and $\hat{y}$ is the prediction calculated from Equation 3.1. If we have a dataset with N training data, the loss function for the data can be defined as:

$$L = \frac{1}{2N} \sum_{i=1}^{N} L_i, \tag{3.4}$$

where L_i is loss of i^{th} sample which is :

$$L_i = (y_i - \hat{y}_i)^2. \tag{3.5}$$

Alternatively, we can write Equation 3.5 as:

$$L_i(\boldsymbol{w}, \boldsymbol{x}_i) = \left(y_i - \frac{1}{1 + e^{-(\boldsymbol{w}^T \boldsymbol{x}_i + b)}} \right)^2. \tag{3.6}$$

We can compute the derivatives as:

$$\frac{\partial L_i}{\partial w_j} = 2(y_i - \hat{y}_i)\hat{y}_i(1 - \hat{y}_i)x_{ij} \tag{3.7}$$

$$\frac{\partial L_i}{\partial b} = 2(y_i - \hat{y}_i)\hat{y}_i(1 - \hat{y}_i). \tag{3.8}$$

These derivatives can be used to update the weights using gradient descent:

$$w_j^t = w_j^{t-1} - \eta \frac{\partial L}{\partial w_j^{t-1}}, \tag{3.9}$$

$$b^t = b^{t-1} - \eta \frac{\partial L}{\partial b^{t-1}}, \tag{3.10}$$

where η is the learning rate of the gradient descent and t represents the iteration number. The training data is presented to the perceptron multiple times (each iteration is called an epoch), to allow the learning process to converge to the solution.

The perceptron detailed above can be used for a binary classification problem where we have only two classes to predict. For multiclass problems, multiple neurons equal to the number of classes can be used. A softmax layer is used for the final output. The softmax

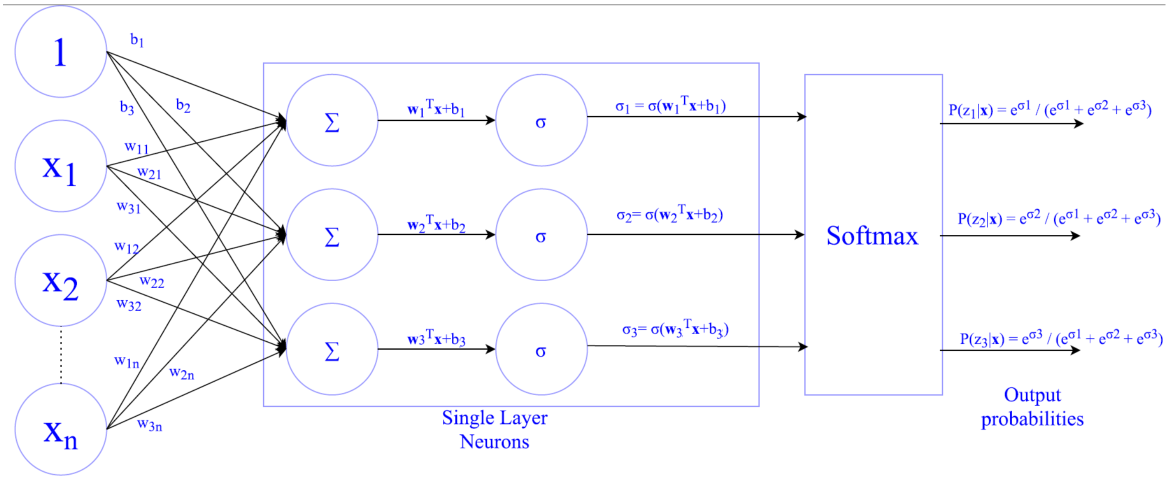

1
b_1
b_3
b_2
x_1
w_{11}
w_{21}
w_{31}
x_2
w_{12}
w_{22}
w_{32}
x_n
w_{1n}
w_{2n}
w_{3n}
Σ
Σ
Σ
$\mathbf{w}_1{}^T\mathbf{x}+b_1$
$\mathbf{w}_2{}^T\mathbf{x}+b_2$
$w_3{}^T\mathbf{x}+b_3$
σ
σ
σ
$\sigma_1 = \sigma(\mathbf{w}_1{}^T\mathbf{x}+b_1)$
$\sigma_2= \sigma(\mathbf{w}_2{}^T\mathbf{x}+b_2)$
$\sigma_3= \sigma(\mathbf{w}_3{}^T\mathbf{x}+b_3)$
Single Layer Neurons
Softmax
$P(z_1|\mathbf{x}) = e^{\sigma 1} / (e^{\sigma 1} + e^{\sigma 2} + e^{\sigma 3})$
$P(z_2|\mathbf{x}) = e^{\sigma 2} / (e^{\sigma 1} + e^{\sigma 2} + e^{\sigma 3})$
$P(z_3|\mathbf{x}) = e^{\sigma 3} / (e^{\sigma 1} + e^{\sigma 2} + e^{\sigma 3})$
Output probabilities

Fig. 3.9 Multiclass classification using perceptrons. All the inputs are fed to multiple neurons and output of every neuron is fed to a softmax layer. The softmax layer normalizes the output values to form a probability distribution. The output with a highest probability is considered as the final prediction.

layer normalizes the output values to form a probability distribution for all output classes. An example of a multiclass classification problem using perceptrons is presented in Figure 3.9.

3.2.2 Multi-layer Perceptron

The network discussed in Figure 3.9 can be used to solve the multiclass classification problem. The network is able to solve a linearly separable problem (a situation where classes can be separated by a line). However, the network is not suitable to for non-linear problems such as logical exclusive-OR (XOR) functions. The XOR function is a non-linear function where the classes cannot be separated by a line. To model such complex problems, the size of the network can be increased by cascading more layers between the first (input) layer and the last (output) layer. These intermediate layers are often termed as ‘hidden layers’ in the literature. With the increase in the number of layers, more parameters are included in the network enabling to model more complex problems. The multi-layer perceptron receives its input data in the vector form through the input layer. The data is multiplied by the assigned weights and passed through an activation function. The data is passed to every neuron in the next layer (the hidden layer). The data is propagated in a similar way to the next hidden layer and then finally to the last layer (the output layer). Note that there should be a unique weight between all possible pairs of the neurons of a layer (l) and the neurons of the next layer ($l+1$). In this manner, the data is propagated from the input layer to the output layer, which gives the final prediction of the network. The deviation of the predicted output from the ground truth is typically called error. With the differentiable layers, gradient descent can be applied to optimize the weights of the network. The propagation of the derivative of the error term from the last layer to the first layer is called backpropagation. A multi-layer perceptron (MLP) is presented in Figure 3.10. MLP does not have to be fully connected. Therefore the connection between the layers can be skipped (for instance layer 1 can be connected to layer 3).

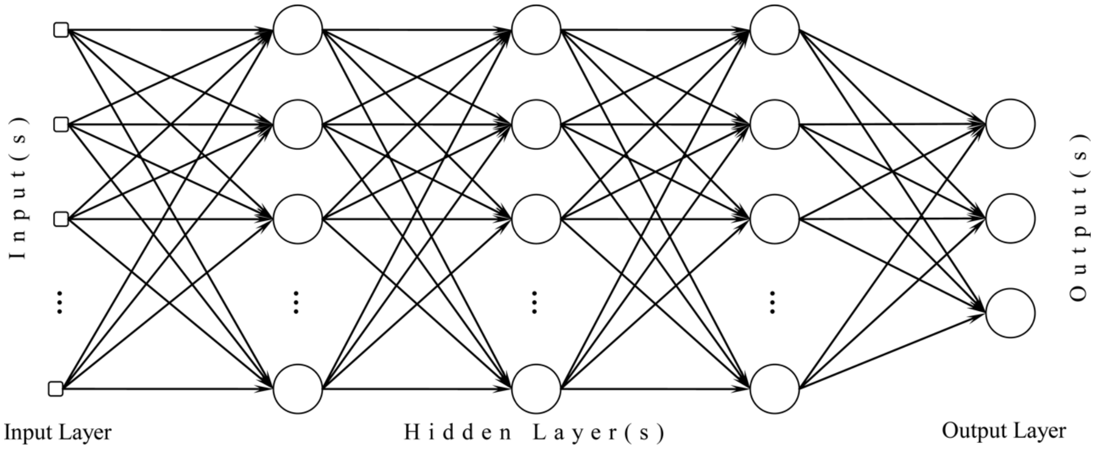


Fig. 3.10 Multi-layer perceptron or fully connected network.

The MLP is a powerful tool in the Artificial Intelligence that can be applied to a number of classification or regression problems. It has been shown that any function can be modeled with a MLP with a large enough number of neurons [59]. However, it has a large number of parameters and is not considered the most suitable option for image and time series data.

MLPs, like most other machine learning methods, require preprocessed features as input to solve a particular problem. These features are usually termed "hand-crafted features". The features are domain dependent and problem-specific expertise is required to design a feature extractor that transforms input data into the feature vectors. These features are input to the learning algorithm to solve a particular machine learning problem. The learning algorithm is highly dependent on these features, thus its performance is limited to the human knowledge about the domain.

3.2.3 Convolutional Neural Network

Convolutional neural networks (CNNs) were originally proposed in the computer vision domain for solving an image classification problem. CNNs aim to learn a non-linear mapping from the input space to the target space. For example, in the case of classifying an image, the input space is the image and the target space is its predicted class.

The major strength of the CNN comes from its representation learning capability. CNNs do not require the hand-crafted features as input. Rather, features are learned directly from data during training to yield the final prediction. For example, in the case of image classification, the image is taken as input, the network learns the features by itself during training and yields the prediction results as the output.

The CNN was first introduced by Yann LeCunn in the domain of image classification. In this pioneering work, a network was proposed for handwritten digit recognition [60]. The network takes a single channel 32×32 pixel image containing handwritten digit and classifies the image into ten classes representing the digits from 0 to 9. Despite their merits, CNNs remained unused for complex computer vision problems due to their huge memory requirements, large datasets for training and high computing power. Later on, in 2012, Alex Krizhevsky proposed a CNN model called AlexNet and applied it to the ImageNet classification challenge. The challenge comprised of 1.5 million images for training with a thousand image categories. AlexNet outperformed the previous state-of-the-art by a large margin [52]. Since then, many variants of CNNs like VGGNet [61], ResNet [62], GoogLeNet [63] have been proposed achieving ever increasing performance.

Unlike MLP, CNNs consists of different unique types of layers. The different layers of the CNNs are usually convolutional, pooling, and fully connected. A typical CNN model is presented in Figure 3.11.

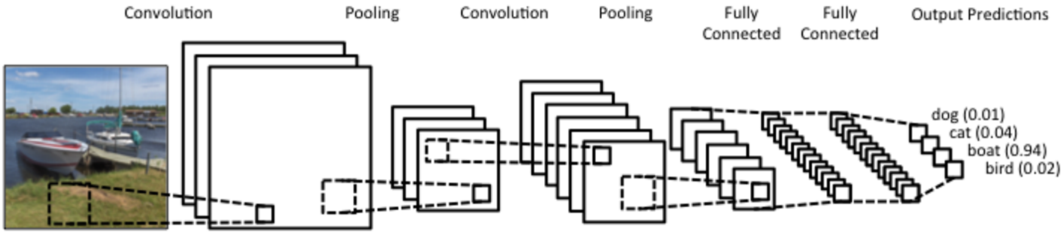


Fig. 3.11 A Convolutional Neural Network for classifying an image into four categories.

Convolutional Layer

The convolutional layer performs convolution on the input feature map x with K number of filters f and produces output feature map y. The process for two-dimensional convolution is graphically illustrated in Figure 3.12. Mathematically, the values in the output feature map can be computed as:

$$y(i,j,k) = \sum_{c=1}^{C} \sum_{P_i=-P}^{P} \sum_{P_j=-P}^{P} x(Si-S+1+P_i, Sj-S+1+P_j, c) f(P_i+P+1, P_j+P+1, c, k) + b_k, \tag{3.11}$$

where $k = 1,2,...,K$, C is the number of filters in the input feature maps, P is the amount of zero padding around the input feature map and S is a hyper-parameter called ‘stride’.

In the case of one-dimensional convolution, the filter is one dimensional and the convolution is applied to the input one-dimensional data. Figure 3.13 shows the graphical illustration of one-dimensional convolution operation.

Maxpooling Layer

Maxpooling layers reduce the spatial size/extent of the feature map to produce a smaller output feature map. The maxpooling layers do not have trainable parameters and simply

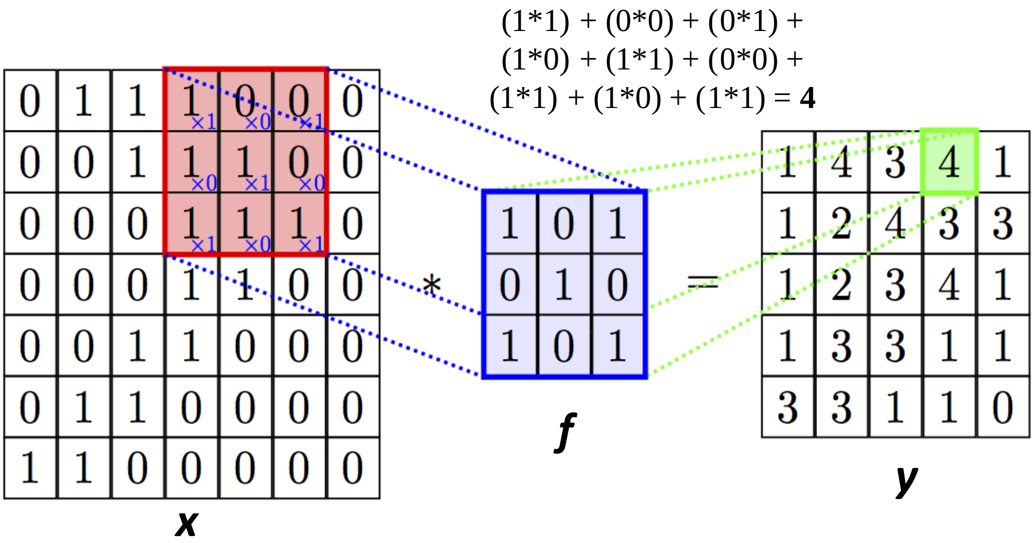


Fig. 3.12 Two-dimensional convolution of input x with filter f. The number of filters is one with no padding. Calculation of one value of output feature map, which is 4, is illustrated.

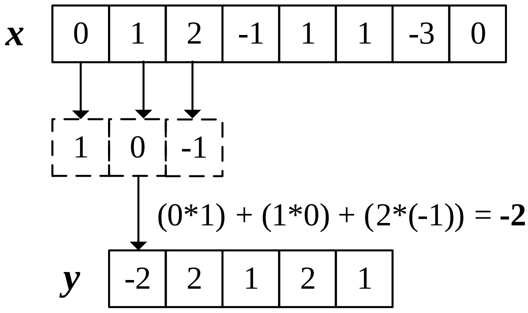
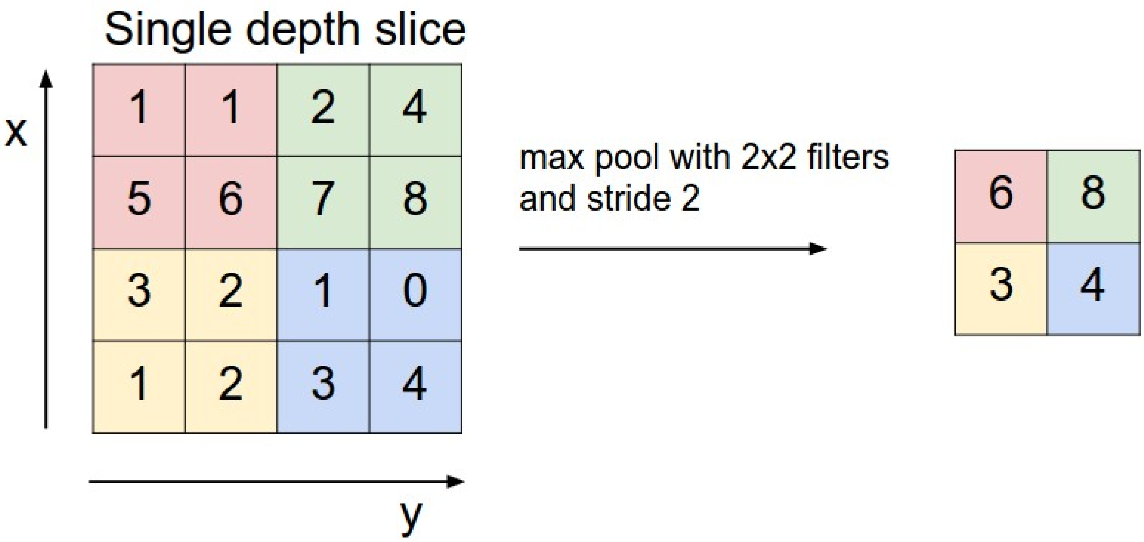


Fig. 3.13 One-dimensional convolution of input x with filter f. The calculation of a value (which is -2) in output feature map y is presented.

choose the maximum value from the input feature map under the receptive field. Mathematically, maxpooling can be expressed as:

$$y(i,j) = \max\{x(S_i - S + m_i, S_j - S + m_j) : m_i = 1,2,...M, m_j = 1,2,...M\}, \qquad (3.12)$$

where S is the stride and M is the size of the receptive field. Figure 3.14 shows graphical illustration of the maxpooling.



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Fig. 3.14 Two-dimensional maxpooling of image i with filter size of 2×2 and stride of 2. For each of the elements represented by the filter, a maximum is selected and a new element is created in the output matrix. For example, from top-left four elements $(1,1,5,6)$, maximum value 6 is selected for the output matrix.

In the case of one-dimensional data, the maxpooling layer is expressed in Equation 3.13.

$$y(i) = \max\{x(S_i - S + m_i) : m_i = 1,2,...M\}. \tag{3.13}$$

In the previous section, we have described sigmoid nonlinear function. There are some other nonlinear functions such as Rectified Linear Unit and Parametric Rectified Linear Unit used in literature. We will describe these functions in the next section.

3.2.4 Rectified Linear Unit (ReLU)

The rectified linear unit (ReLU) is a commonly used non-linear function and is preferred over sigmoid non-linear function. It has been shown that ReLU greatly accelerates the convergence rate of optimization as compared to the sigmoid function [52] and is efficient to implement. For an input x, the ReLU operation y is presented in Equation 3.14.

$$y(x) = \begin{cases} \text{x}, & \text{if } x > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.14}$$

The ReLU function is visualised in Figure 3.15.

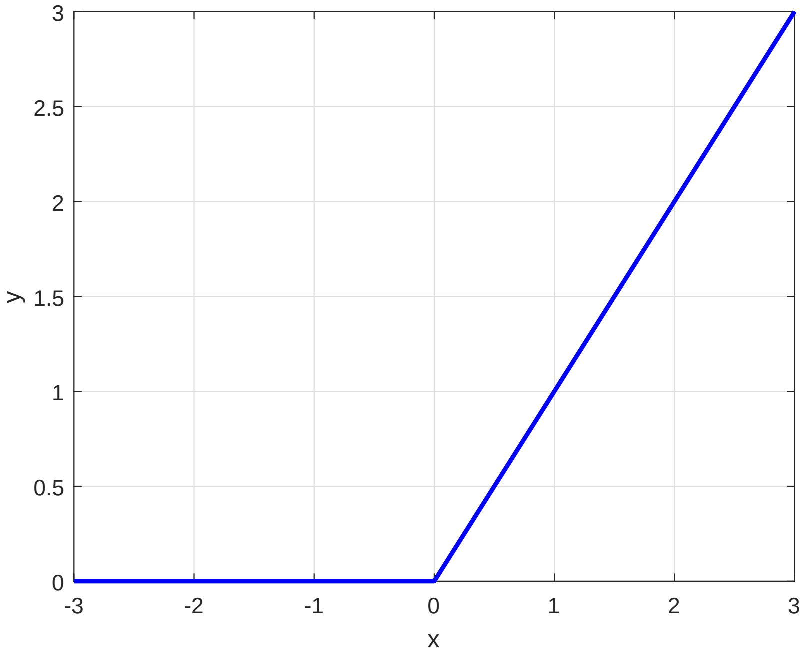


Fig. 3.15 Rectified linear unit (ReLU) for input x and output y.

3.2.5 Parametric Rectified Linear Unit (PReLU)

ReLU is a commonly used activation function. However, it might suffer from the "dying ReLU" problem. The "dying ReLU" refers to a scenario when neurons become inactive and only output 0 for any input. A parametric rectified linear unit (PReLU) was introduced to improve the performance of a neural network [64]. The PReLU adaptively learns the rectifier parameters to improve accuracy [64]. For an input x, the PReLU is defined as:

$$f(x) = \begin{cases} x, & \text{if } x > 0, \\ ax, & \text{if } x < 0, \end{cases} \tag{3.15}$$

where x is the input to the activation function and a is the coefficient controlling the slope of the negative part of the input. The coefficient a is learned during training. However, if $a = 0$, it yields the ReLU function. PReLU is graphically presented in Figure 3.16.

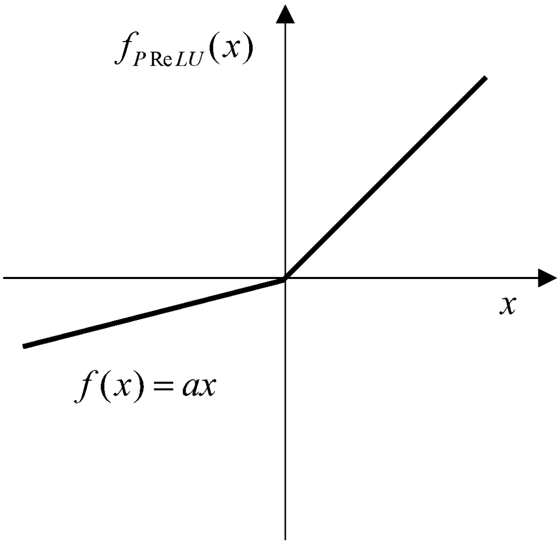


Fig. 3.16 Parametric Rectified Linear Unit (ReLU).

In the next chapter we will give an overview of the literature related to the applications of machine learning in ADHD classification.

Chapter 4

Literature Review

A number of exciting studies have explored different machine learning methods for studying fMRI data that shows the popularity of fMRI as a tool for mapping human brain functions and especially in the case of studying brain disorders [1, 6, 39, 65, 66]. In this dissertation, we have focused on functional connectivity for prediction of ADHD. In recent years, a great deal of research has been conducted on functional connectivity leading to the emergence of several methods for the analysis of functional connectivity using fMRI. These methods can be classified into two broad categories, i) model-based methods and, ii) data-driven methods [67, 68] which are described below and shown in Figure 4.1.

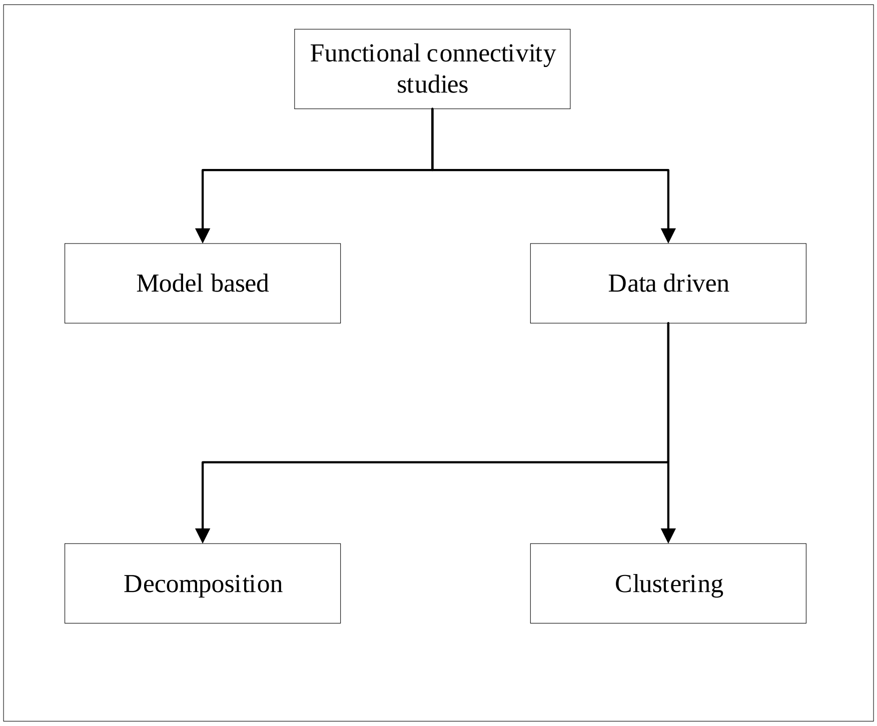


Fig. 4.1 Methods for studying functional connectivity.

4.1 Model-based Methods

Many studies have explored model-based analysis for exploring functional connectivity. These studies choose some region of interest (ROI) and determine whether other regions are functionally connected to these ROIs or not, and generate the connectivity map of the human brain. These selected ROIs are also called ‘seeds’. The generated connectivity map shows the brain regions significantly correlated with the ROI or seed [69]. As these studies require selection of seeds prior to the analysis, the studies are typically based on some strong prior neuroscience knowledge. Figure 4.2 shows an illustration of seed-based analysis.

The pioneering method used for resting-state fMRI analysis was a seed-based technique, where Biswal and colleagues [38] studied functional connectivity between regions belonging to the primary motor cortex of the human brain. After this work, several studies applied seed-based analysis on fMRI data [71–76]. Correlation analysis is a widely used method lying under this category.

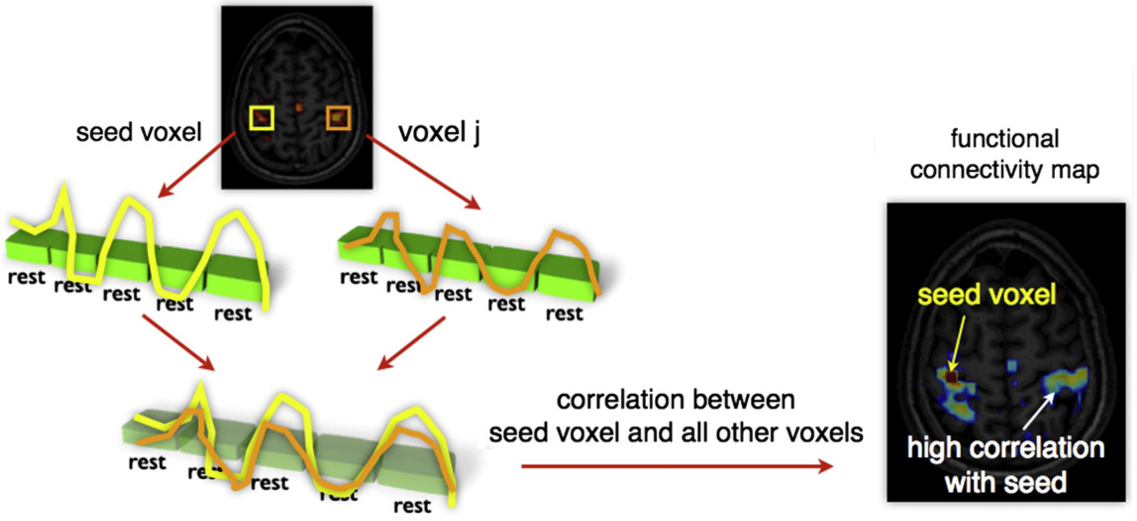

seed voxel
voxel j
rest
rest
rest
rest
rest
functional connectivity map
seed voxel
high correlation with seed
correlation between seed voxel and all other voxels

Fig. 4.2 Seed based functional connectivity analysis. To estimate the functional connectivity between a pre-selected ROI (seed) and any other brain voxel j, the time series of the seed region (or voxel) is correlated with the time series of voxel j. A significantly high level of correlation between the two regions show that both regions are functionally connected. In order to calculate the functional connectivity of the seed with all other regions, the time series of the seed is correlated with all other regions and a functional connectivity map is generated which reflects the functional connectivity of the seed with other groups of regions. (Adapted from [70]).

Correlation-based Methods

Correlation is a method for calculating functional connectivity where the regions with high correlation coefficient are considered strongly functionally connected. Dai et al. [65] segmented the brain into 351 ROIs using a template provided by [77] and calculated functional connectivity using Pearson's correlation. Bohland et al. [78] applied the Automated Anatomical Labeling (AAL) atlas [79] to segment the brain into 116 ROIs and computed functional connectivity using three correlation variants: Pearson's correlation, sparse regularized inverse covariance [80] and Patel's Kappa [81]. Eloyan et al. [82] extracted five ROIs belonging to the motor network with 264 voxels as nodes and computed functional connectivity using Pearson's correlation coefficient which was later used for classification. Similarly, Cheng et al. [83] employed Pearson's correlation and partial correlation to calculate functional connectivity of 90 brain regions extracted from the AAL template [79]. Multiple measures including Regional Homogeneity (ReHo), functional connectivity and fractional amplitude of low-frequency fluctuation (fALFF) have been employed for classification.

Correlation is an efficient and easy method for functional connectivity analysis. However, a correlation-based approach does not characterize the network structure of different brain regions, i.e. whether two brain regions belong to the same functional cluster or not [84]. Moreover, the network obtained by correlation is quite dense, which degrades the performance of a classifier [39, 84].

Model-based methods have been proven to be a powerful and effective tool for identifying brain areas which are functionally connected to a seed during the resting-state. However, these methods suffer from some major drawbacks. Firstly, model-based methods are sensitive to the initial seed and it is common that different seeds yield different functional connectivity maps [68, 85]. Secondly, functional connectivity calculated through this method is constrained by the requirement of prior neuroscience knowledge. Even with prior knowledge, one can only study the functional connectivity of regions related to the specific prior. Moreover, the seed-based method evaluates one seed at a time. In the case of fMRI, it will be more informative to simultaneously consider multiple functional connectivity patterns.

4.2 Data-driven Methods

To overcome the limitations of model-based methods, data-driven methods have been introduced in the domain of functional connectivity analysis. These methods enable exploration of functional connectivity without a need to define a priori seed information. Data-driven methods are designed to explore general connectivity patterns across the whole brain. Several methods have been proposed and can be categorized into two main categories, i) decomposition methods and ii) clustering. Decomposition methods include techniques such as principal component analysis (PCA), independent component analysis (ICA), and singular value decomposition (SVD). These methods aim to represent the original fMRI data as a linear combination of basis vectors (PCA and SVD) or statistically independent components (ICA). Clustering methods apply traditional clustering techniques to the fMRI time-series data. Both of these techniques are exploratory and estimate functional connectivity of the whole brain.

4.2.1 Decomposition Methods

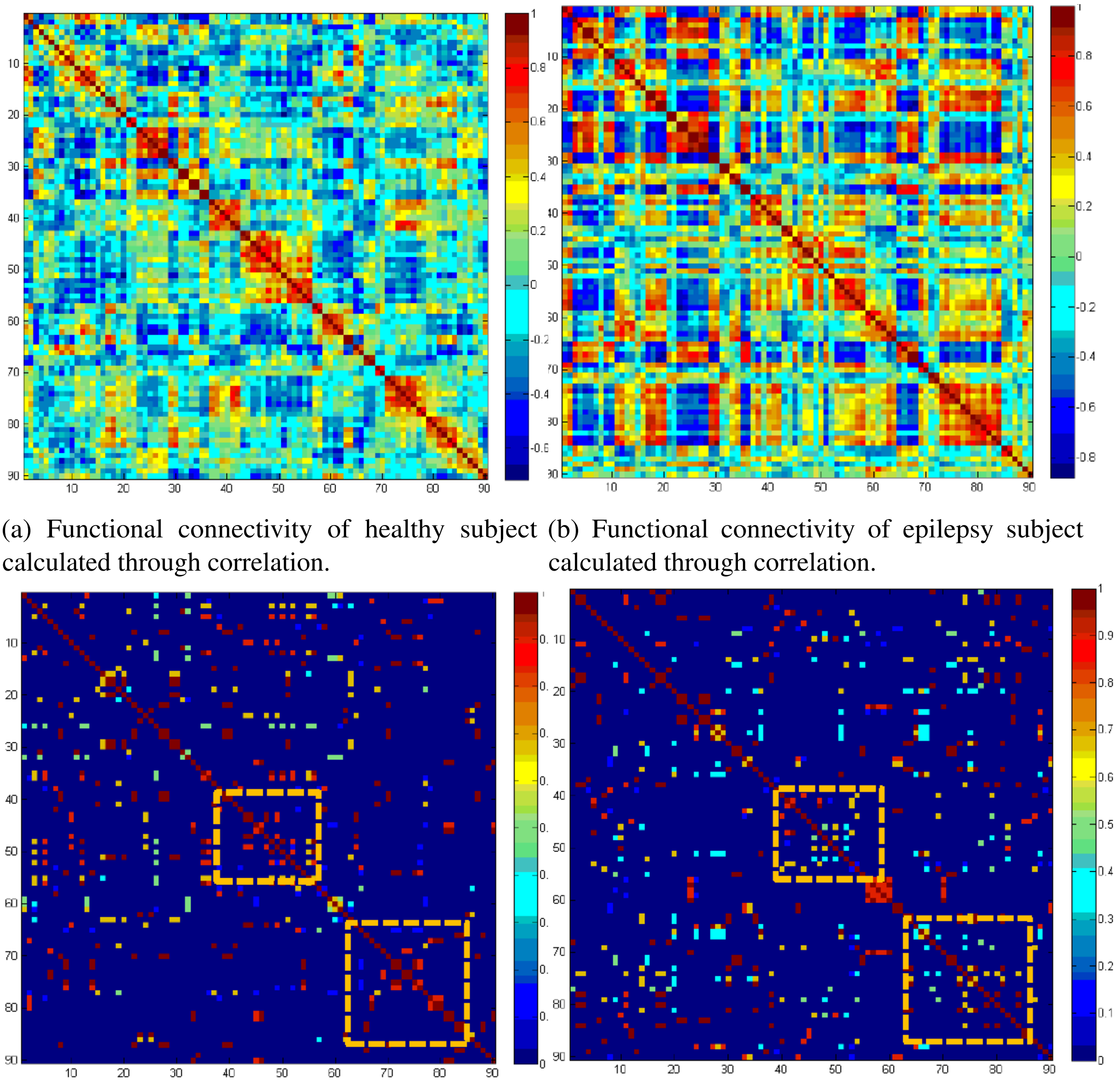
Decomposition methods such as ICA are commonly used with fMRI data. The aim of these methods is to discover the underlying structure of the data rather than requiring a priori information (seed). ICA was introduced in fMRI analysis to decompose fMRI data into spatially independent components [86]. Consequently, Garcia et al. [6] proposed an ICA based functional-anatomical discriminative region model for pattern classification of ADHD. In this study, the authors applied ICA to extract functional connectivity networks in the brain. Similarly, Tabas et al. [66] proposed a variant of ICA to characterize the differences between a healthy control group and an ADHD group. This study used twenty independent components and combined ICA and a spatial variant of Fisher's linear discriminant towards exploring the differences between the two groups. ICA-based methods are considered a natural solution for fMRI studies as these methods do not need any prior information about the spatial or temporal patterns of source signals.

ICA-based approaches have shown success in classification tasks, however, there are limitations to these methods. Firstly, independent components are often perceived as difficult to manipulate [70]. ICA is based on the assumption that components (signal sources) are independent, whether spatially or temporally, and violation of the assumption degrades performance. Moreover, the need to specify the number of independent components and a threshold value for independent components is considered a drawback [87].

4.2.2 Clustering-based Methods

Clustering is another popular approach for the evaluation of functional connectivity, where regions belonging to the same cluster are assumed to be functionally connected. Studies have shown that clustering-based approaches yield better performance than correlation-based approaches as the network obtained by clustering is sparse [1, 87]. Figure 4.3 shows functional connectivity matrices obtained from correlation and clustering methods respectively. The authors [1] demonstrate clearly that the results from the clustering method are more sparse, and give a better classification accuracy.

Zhang et al. [84] applied *k*-means clustering to calculate functional connectivity. However, in *k*-means, random initialization of clusters and prior information on the number of clusters emerge as a major drawback, as these are unknown in the case of fMRI. Hierarchical clustering can also be applied to calculate functional connectivity [2], however, the selection of the threshold value and the number of clusters are not known in advance in the case of fMRI. Other studies (e.g. [1]) have applied affinity propagation (AP) [88] clustering for the classification of brain disorders. AP clustering does not require an initial number of clusters, which is a good choice for fMRI data.



(a) Functional connectivity of healthy subject calculated through correlation.

(b) Functional connectivity of epilepsy subject calculated through correlation.

(c) Functional connectivity of healthy subject calculated through clustering.

(d) Functional connectivity of epilepsy subject calculated through clustering.

Fig. 4.3 Comparison of functional connectivity calculated through correlation and clustering. The left column presents the functional connectivity matrix of a healthy subject and the right column presents the functional connectivity of an epilepsy subject. The matrix obtained via clustering is more sparse than the correlation matrix. The sparse matrix yields better classification results [1]. (Modified from [1]).

4.3 Graph-based Methods

Graph based methods provide an alternative to model-based and data-driven methods. These methods typically work in two steps: In the first step, these methods identify a set of nodes from fMRI data, and, in the second step, these methods estimate the strength of connections or "edges" between the nodes.

The modeling of a complex system using a graph-based structure depends to a great extent on how accurately the nodes and edges represent the true system and their interactions [89]. For some domains, networks are well defined in the system and it is very straightforward to present them as graph-based models. For example, in the case of social networks, where the nodes represent a person and the edges represent their social connectivity [90], co-authored publications [91] or email traffic [92]. In the case of modeling the brain network as a graph-based structure, proper identification of nodes and edges is complex [89]. One might argue for the representation of each node as a neuron and each synaptic connection between the neurons as an edge. However, modeling billions of neurons and trillions of connections between them as nodes and edges is not feasible [93, 94]. Therefore, nodes are identified by grouping different brain voxels based on some criteria.

Modeling the brain as a network typically involves three steps: i) defining nodes of the network, ii) extracting associated time-series, and iii) defining the strength of edges between nodes. These steps are illustrated in Figure 4.4. The network nodes can be identified in different ways. Mostly, the nodes are identified as spatial ROIs, typically from a brain template (e.g. AAL atlas [79]). Alternatively, clustering based parcellation can be used to define the nodes where the identified clusters can be presented as network nodes. Once the nodes are identified, their associated time series are used to calculate the strength of the edges between all nodes [95, 96]. In general, correlation is the simplest and most common measure to calculate edges between the nodes. A number of graph based approaches have been applied to model brain networks.

Dey et al. [5] proposed a graph-based solution for the classification of ADHD. The authors employed the CC200 [77] template to identify the nodes of the network where voxels belonging to each ROI were grouped together, with each ROI was represented as

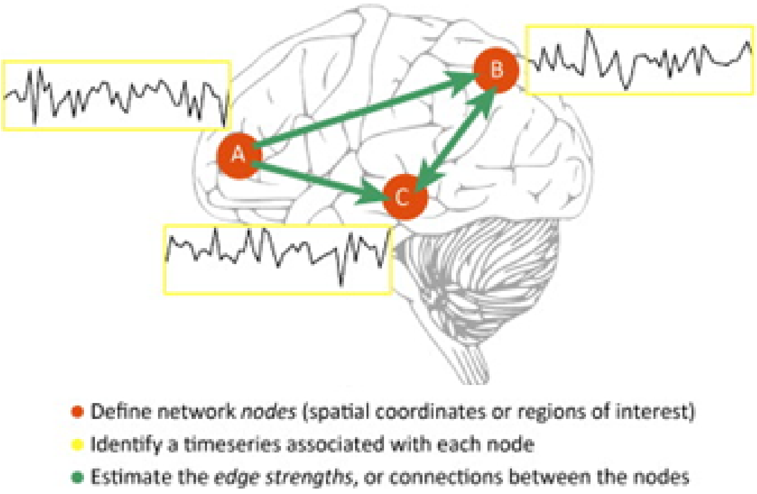

A
B
C
● Define network *nodes* (spatial coordinates or regions of interest)
● Identify a timeseries associated with each node
● Estimate the *edge strengths*, or connections between the nodes

Fig. 4.4 Illustration of modelling the brain as a network. There are three steps: i) defining network nodes (presented as red colour), ii) extracting associated time-series (yellow colour), iii) representing strength of edges (green colour).

a node of the network. Next, the network was modelled using the identified nodes and defining multiple graph basis measures (the authors termed this the "signature of a node"). Correlation was calculated and a threshold applied to construct the functional network. The threshold value was arbitrarily chosen and different values were employed for different imaging datasets. Similarly, Siqueira et al. [97] investigated different graph-based measures for the classification of ADHD. In this study, the brain was segmented into 400 ROIs using a template provided by Craddock [77]. The edge strength between all the nodes was calculated through Pearson’s correlation. Multiple graph measures were applied and finally, an SVM classifier was used to obtain final classification results.

4.4 Deep Learning-based Methods

Recently, end-to-end deep learning-based networks have been shown to outperform existing classical machine learning models in a number of domains like image classification, image segmentation and object recognition [98]. Generally speaking, an end-to-end trainable network refers to a single learning system where the predicted label of a neural network model is predicted directly from the input, with all weights learned through back-propagation. There is very limited work exploring deep learning for fMRI-based classification of neurological disorders [8, 99, 100].

The use of an artificial neural network for classification of ADHD has been explored in [99]. In this work, brain was segmented into 190 brain regions. Classical machine learning methods were applied to extract multiple features such as granger causality, a non linear extension of Granger causality [101], and PCA. A t-test was applied for feature selection and the selected features were passed into a fully connected neural network for the final classification.

Similarly, the study in [100] addressed the problem of classification of mild cognitive impairment (MCI) using fMRI data. The authors applied a deep autoencoder for dimensionality reduction of fMRI time-series signals. The representation encoded by the autoencoder was fed into a hidden Markov model to estimate the likelihood of a subject belonging to the healthy control group or the MCI group to identify its predicted label. In another study [102], authors applied a deep neural network for the classification of schizophrenia. The brain was segmented by using the AAL template [79], and Pearson's correlation was applied to extract functional connectivity. The extracted functional connectivity was used as features for a deep neural network that yielded the final classification result.

In a non-peer reviewed study [103], authors applied a CNN for classification of Alzheimer's disease using fMRI data. The study applied a two-dimensional CNN where four-dimensional fMRI data was converted into a stack of two-dimensional images and classification was evaluated on the individual two-dimensional images. The study did not incorporate temporal information, which is the most important aspect of fMRI time-series data. Being a two-dimensional CNN model, the prediction results were evaluated for individual images instead of per subject. For the prediction of a subject, the results of individual two-dimensional images were accumulated. A recent study [104] applied a three-dimensional CNN for classification of Autism Spectrum Disorder. The fMRI three-dimensional volume was downsampled and a three-dimensional CNN was applied on the downsampled data. The study did not incorporate functional connectivity which is an important characteristic in brain studies.

The Siamese network is used to compare two or more input patterns, typically images. Siamese networks are usually comprised of two neural networks that extract features from the input data and a metric measure that is used to calculate the distance or similarity between the extracted features. It was introduced in [137], where the authors designed a network to verify and match two hand signatures. In this network a Siamese network consists of two neural networks that map preprocessed images to feature sets and a cosine-based distance measure is used to find the similarity between the features. Later on, a number of studies applied Siamese networks in different domains like image patch matching [3] and face verification [138]. In [138], authors used two CNNs to map input images to a low-dimensional space and a distance-based measure to calculate the similarity between them. The two CNNs shared the same parameter set. In [3], authors used two CNNs to map the images to low-dimensional space and instead of a distance-based measure, they used a neural network to find the similarity between the images. Our proposed network in chapter 6 and 7 is inspired from the [3].

These studies highlight the importance of machine learning towards the diagnosis of brain disorders. In this dissertation, we were interested in developing novel methods for calculating functional connectivity and exploiting it for the diagnosis of ADHD. Studies have highlighted that functional connectivity calculated by clustering is better as compared to other methods such as correlation. However, clustering based functional connectivity was not explored for the ADHD-200 dataset (details of the dataset are in next chapters). We were interested to explore whether a clustering-based model of functional connectivity can improve the classification results for the ADHD-200 dataset. In the existing literature, deep learning has not been explored for the calculation of functional connectivity. We wished to investigate whether deep learning based methods can be designed to calculate functional connectivity to improve the state-of-the-art performance. In the reviewed literature, we were not able to find studies applying end-to-end deep network incorporating functional connectivity for prediction of a brain disorder. We were interested to propose an end-to-end deep learning model and evaluate whether it can improve the classification results. In the following chapters of the dissertation, we will discuss the proposed methods and the results.

Chapter 5

Integration of Non-imaging and Imaging Data for Classification

The literature discussed in the previous chapter shows encouraging results to demonstrate that machine learning techniques hold promise for the analysis of neuroimaging data. Most classical machine learning studies rely on correlation-based approaches for the calculation of functional connectivity. However, correlation-based approaches do not characterize the network structure of brain regions, i.e., whether two brain regions belong to the same functional cluster or not [84]. In addition, the network obtained by correlation is quite dense which may degrade the performance of the classifier [39, 84].

Studies have shown that a clustering-based approach is more sophisticated as compared to correlation-based approaches as the network obtained by clustering is sparse [1, 87]. Different clustering approaches can be applied to determine functional connectivity. Zhang et al. [84] applied *k*-means clustering to calculate functional connectivity. However, in *k*-means, random initialization of clusters and a priori information of the number of clusters may emerge as a major drawback, as in the case of fMRI the number of clusters is not known. Hierarchical clustering can also be applied to calculate functional connectivity [2], but the selection of threshold and number of clusters may emerge as a drawback for this method. To overcome these problems, we propose a hybrid clustering approach that determines the number of clusters from the data itself.

In this work, our motivation is to study functional connectivity alterations induced by ADHD. However, unlike previous work that relies on imaging data alone, we bring together two types of features, namely non-imaging and imaging features to form a single feature vector used for classification of individuals as ADHD or control (non-ADHD). Our framework is comprised of multiple stages. In the first stage, the functional connectivity between brain regions is determined using the Affinity Propagation (AP) clustering algorithm [88]. Instead of requiring a number of clusters in advance, AP takes a measure of similarity between data points and the initial preference for each point for being the cluster centroid. We propose a novel method to find these cluster centroids through a matrix derived from the Density Peaks (DP) algorithm by Rodriguez and Laio [105]. Next, we select discriminant features through an Elastic Net (EN), which combines variable shrinkage with a grouped selection of variables. Finally, we employ a Support Vector Machine (SVM) classifier to classify between control and ADHD. We demonstrate that the integration of non-imaging and imaging data in our framework improves performance.

The main contributions and key findings of this chapter are:

- A novel method to initialize the AP clustering algorithm by employing the Density Peaks approach.

- Demonstration of the importance of non-imaging data for classification of control vs. ADHD based on the functional connectivity between various brain regions.

- Anatomical findings of our results reveal that the Frontal and Parietal (premotor) lobes have the largest number of functional connectivity alterations for all the tested datasets.

- Experimental results outperform the previous state-of-the-art for three test datasets of the publicly available ADHD 200 data.

The publications related to this chapter are:

- Atif Riaz, Muhammad Asad, Eduardo Alonso, and Greg Slabaugh, “Fusion of fMRI and Non-Imaging Data for ADHD Classification”, Computerized Medical Imaging and Graphics Volume 65, April 2018, Pages 115-128.

- Atif Riaz, Eduardo Alonso, Greg Slabaugh, “Phenotypic Integrated Framework for Classification of ADHD using fMRI”, International Conference on Image Analysis and Recognition (ICIAR) 2016, Pages 217-225, Springer.

5.1 Data

The resting-state fMRI data used in this study is from the NeuroBureau ADHD-200 competition [106]. The data consists of resting-state functional MRI data as well as phenotypic information (non-imaging data) for each subject. There was a global competition held for classification of ADHD subjects, and the consortium has provided training and an independent test dataset for each imaging site. Eight different imaging sites contributed to the dataset. For this study we used datasets from four sites: Kennedy Krieger Institute (KKI), NeuroImage (NI), New York University Medical Center (NYU) and Peking University (Peking). The dataset is complex as well as diverse, with each site having different number of subjects, scan parameters and equipment. For all of our experiments in this chapter, we used the preprocessed data released for the competition for the four sites mentioned above. The preprocessing was performed using AFNI [107] and FSL [108] tools, using the Athena computer clusters at the Virginia Tech advance research computing center. The preprocessing steps include: removing the first four time points, slice time correction, motion correction (here the first image is taken as the reference), registration on $4 \times 4 \times 4$ voxel resolution using the Montreal Neurological Institute (MNI) space, filtration (bandpass filter range $0.009Hz < f < 0.08Hz$), and smoothing using a $6mm$ FWHM Gaussian filter. Interested readers may refer to the competition website for further details on data and preprocessing [109].

After preprocessing all the images, the brain is segmented into 90 predefined regions using the Automated Anatomical Labeling atlas [79], where voxels lying in a particular region are averaged to generate a representative time-series signal of the region. We have integrated non-imaging data (age, gender, verbal IQ, performance IQ, and Full4 IQ) for all the sites except NeuroImage, as the data was missing.

The dataset is very challenging as the imaging sites have followed different parameters for scanning. For example, in NI, the subjects were asked to keep their eyes closed. No visual stimulus was presented during the scan. For NYU, the participants were asked to close their eyes, think of nothing systematically and not fall asleep. However, a black screen was presented to them. In Peking, the participants were asked to stay still, and either keep their eyes open or closed. A black screen with a white fixation cross was displayed during the scan. Some other parameters were also not consistent across different sites.

5.2 Method

The framework proposed in this chapter consists of the following modules: functional connectivity calculation, feature selection, fusion of non-imaging data and classification. A block diagram of the methodological framework is presented in Figure 5.1 and a detailed description is given below.

5.2.1 Dataset Balancing

In classification problems, a dataset is characterized as imbalanced when the number of samples belonging to one class is smaller than the ones from other classes. As minority class is comprised of very less number of samples, classification rules that predict the small classes tend to be undiscovered or ignored. Consequently, test samples belonging to the minority class are misclassified more often than those belonging to the prevalent class [110]. In most applications, classes with the lower number of samples are typically those of higher interest [111, 112]. Therefore, the correct classification of samples in the minority class has greater importance than the majority class. For example, in an automatic medical diagnostic problem where the disease samples are typically rare as compared to the healthy subjects, the classification objective is to detect samples with disease. Hence, a classification model that provides higher detection rate is considered as the favorable model. The problem of imbalanced datasets is therefore also referred to as the small or rare class learning problem [110].

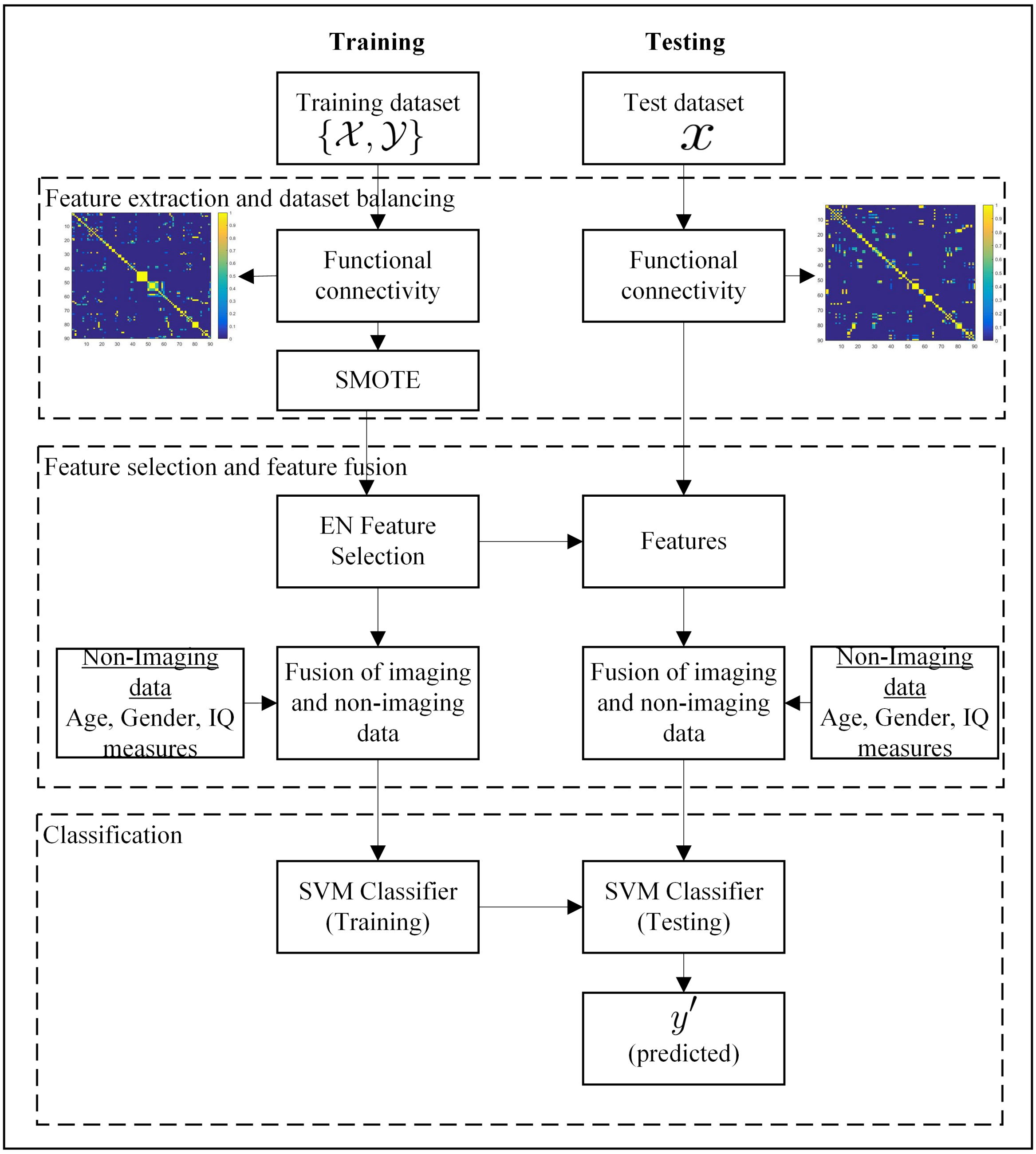

Training
Testing
Training dataset $\{\mathcal{X},\mathcal{Y}\}$
Test dataset x
Feature extraction and dataset balancing
Functional connectivity
Functional connectivity
SMOTE
Feature selection and feature fusion
EN Feature Selection
Features
Non-Imaging data Age, Gender, IQ measures
Fusion of imaging and non-imaging data
Fusion of imaging and non-imaging data
Non-Imaging data Age, Gender, IQ measures
Classification
SVM Classifier (Training)
SVM Classifier (Testing)
y' (predicted)

Fig. 5.1 Flowchart of the methodology. In the first step, functional connectivity is calculated for both training and testing datasets. For imbalanced datasets, SMOTE is applied to the training dataset only. The next step is feature selection, where discriminant features from training dataset are calculated and further used for classification. The selected features are then fused with non-imaging data. Finally, the fused feature set is presented to a SVM for classifier training and testing.

Imbalanced class distribution of a dataset poses serious problems to most of the classification algorithms which generally assume a relatively balanced distribution [111, 113–115]. Dataset imbalance has been identified as a critical problem in domains such as anomaly detection [116, 117], fault diagnosis [118, 119], email forwarding [120], face detection [121], manufacturing plants [115], text classification [122] and especially medical diagnosis [123].

Generally in the machine learning domain, the issue of class imbalance has been addressed in two ways. One way is to assign different costs to the individual classes during training [124]. The other way is to re-sample the training dataset. For data re-sampling, one approach is to randomly down-sample the majority, or over-sample the minority classes to create a balanced training dataset. However, there is a chance that these strategies may yield suboptimal performance [125]. Therefore, instead of these strategies, we apply Synthetic Minority Over-sampling Technique (SMOTE) [126] to create a balanced training dataset. It has been shown previously that SMOTE has improved performance when compared to other approaches like re-sampling or modifying the loss ratios [126]. In SMOTE, the minority class is over-sampled by creating 'synthetic' samples instead of over sampling with replacement. Each minority class is over-sampled by generating synthetic examples along the line segments joining any of the k minority class nearest neighbors. Consider $I_A \in I$, where I is the set of individual subjects and I_A represents the minority subjects. For each individual subject $\boldsymbol{x}_i \in I_A$, k-nearest neighbors of $\boldsymbol{x}_i$ are calculated. A random subject $\hat{\boldsymbol{x}}_i$ is chosen from these neighbors and an additional minority subject is synthesized as

$$\boldsymbol{x}_s = \boldsymbol{x}_i + (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i) \times r, \tag{5.1}$$

where $\boldsymbol{x}_s$ is a synthetic subject and r is a random number such that $r \in [0,1]$. In our work, we applied SMOTE as shown in Figure 5.1. SMOTE is applied only on the training data and is not required for the testing dataset. By applying SMOTE to the minority class in the training dataset creates new synthetic samples by interpolating features between two samples in the minority class. Fundamentally, this assumes the convex hull of the feature space of the minority class doesn't contain samples from the majority class. When this assumption is incorrect, noisy synthetic samples may be generated. However, the SVM contains slack variables that tolerate misclassified samples.

5.2.2 Functional Connectivity

Functional connectivity can be defined as the temporal correlation between spatially separate brain regions and can be estimated by calculating the correlation of temporal signals [5, 97], as well as clustering [39]. To overcome the limitations of some popular clustering algorithms such as k-means, as discussed earlier, we propose a hybrid framework which employs Affinity Propagation (AP) clustering [88] and the Density Peaks (DP) algorithm [105] for functional connectivity estimation. Specifically, we employ AP clustering for grouping the brain regions into clusters. AP clustering takes real-valued similarities between brain regions as the input, where the similarity $s(i, j)$ indicates how well the region j is suited for the centroid of the region i. Typically, negative Euclidean distance is employed as the similarity measure [88]. One of the most appealing properties of AP clustering is that it does not require a number of clusters in advance. Rather it takes a real-valued number $s(i,i)$ as input for each region i such that the regions with larger values of $s(i,i)$ are more likely to be selected as centroids. These values are referred to as ‘preferences’ [88]. AP clustering is a message passing algorithm where each data point is simultaneously considered as potential centroid as well as being part of any cluster. Messages are passed between all data points until robust clusters and their centroids emerge. Two kind of messages are passed between data points, namely responsibility and availability messages where each message is associated with a different kind of competition. The responsibility message $r(i, j)$ is sent from the region i to a potential centroid candidate j, reflecting the accumulated strength of how well-suited region j is to serve as a cluster centroid for region i, taking into consideration all other potential centroids for region i. The availability message $a(i, j)$ is sent from a candidate centroid j to region i, which reflects the accumulated strength of how well suited it would be for region i to select region j as its centroid, considering the support from all other regions that shows that region j should be a centroid. Availability messages for all regions are initialized as

$$a(i,j) = 0, \tag{5.2}$$

and responsibility can be calculated as

$$r(i,j) = S(i,j) - \max_{j' \neq j}\{a(i,j') + S(i,j')\}. \tag{5.3}$$

where S in Equation 5.3 is the similarity measure between brain regions as discussed above. For any two regions i and j with temporal dimensions $k = \{1,2,...t\}$, the similarity measure S is initialized as

$$S(i,j) = -\sqrt{\sum_{k=1}^{t} \frac{(i_k - j_k)^2}{\sigma_k^2}}, \tag{5.4}$$

where i_k is the k_{th} time point of region i and σ_k is the standard deviation.

For the initial iteration, with availabilities being zero, responsibility $r(i,j)$ is set to the input similarity $S(i,j)$ between region i and region j as its centroid minus the largest of the similarities between region i and other candidate centroids. In later iterations, when some regions are associated with other centroids, their availabilities will drop to negative values using Equation 5.5. These negative availabilities will effectively remove the corresponding candidate centroids from the competition. With the responsibility updates, the candidate centroid competes for the ownership of a region. The availability update below combines evidence from data whether each candidate centroid would effectively emerge as a good centroid

$$a(i,j) = \min\{0, r(j,j) + \sum_{i', i' \neq \{i,j\}} \max\{0, r(i',j)\}\}, \tag{5.5}$$

The "self-availability" $a(j,j)$ is updated differently as

$$a(j,j) = \sum_{i', i' \neq \{j\}} \max\{0, r(i',j)\}, \tag{5.6}$$

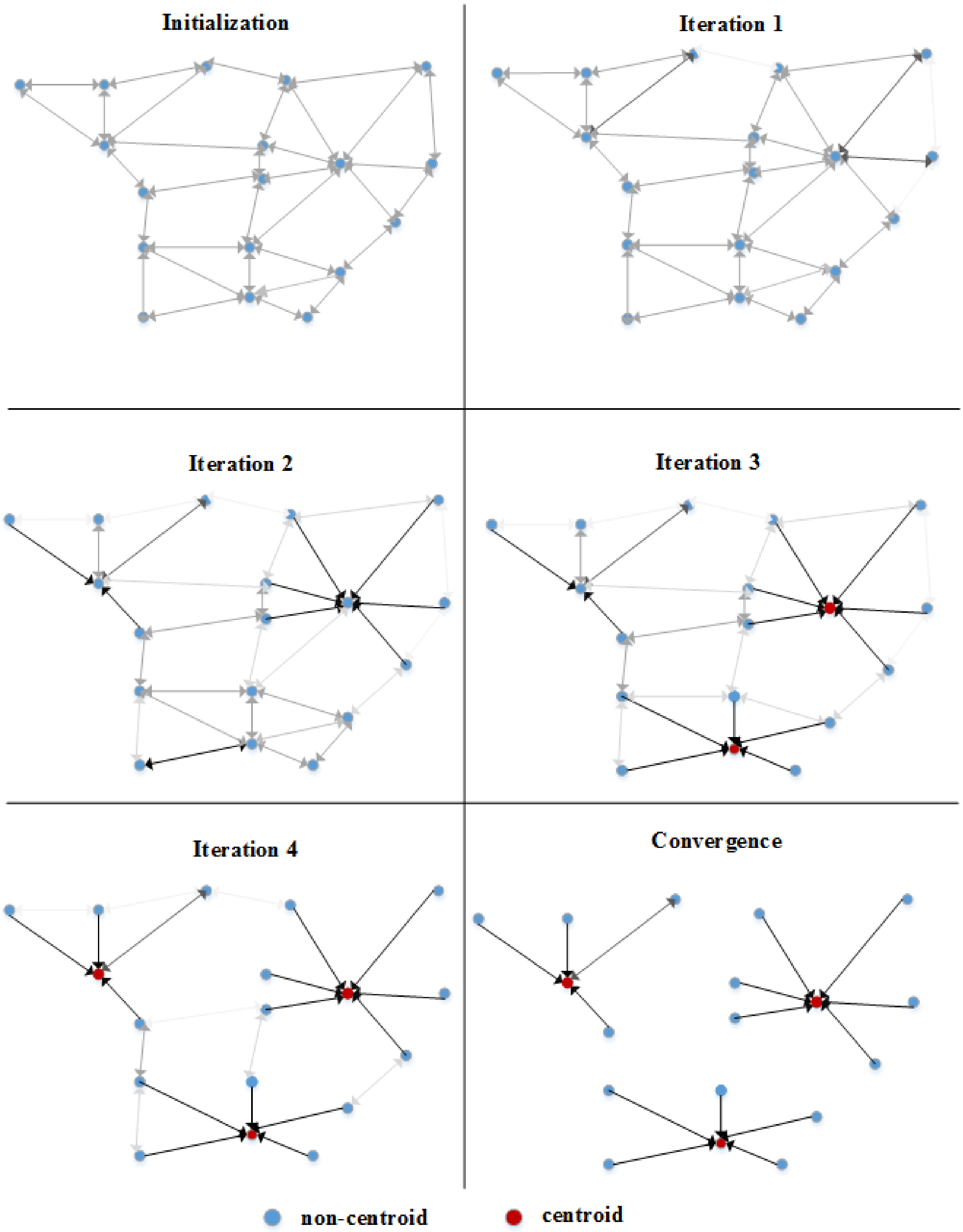


Fig. 5.2 Illustration of AP clustering for two-dimensional data points, where negative Euclidean distance was used to measure similarity. The colour of each point represents the (current) evidence that it is a cluster center (centroid). The darkness of the arrow from point *i* to point *j* represents the strength of the message that point *i* belongs to centroid point *j*. Initially, the strength of messages is weak and there are no clusters. After some iterations, the strength of the messages increases and finally, robust clusters emerge.

The working of AP clustering for two-dimensional points is illustrated in Figure 5.2. Initially, the strength of messages is the same between all the points. After some iterations, the strength of messages for certain points increases and, at the same time it is decreased for some other points. The points with higher strength of messages are potential points of a cluster. After some iterations, clusters and their centroids emerge based on the strength of the messages.

AP clustering does not need a prior guess on the number of clusters, rather it requires a preference value p assigned to each region as the initial probability of being a cluster centroid. The number of identified clusters is influenced by the preference value, but also emerges from the message passing procedure [39, 88]. As a common practice, all data points are considered equally suitable as centroids, thus the preference value is set to a common value. The number of clusters produced is affected by this value. The shared value could be the median of the similarities (moderate number of clusters produced) or their minimum (a small number of clusters produced) [88]. However, instead of initializing with a common value, we propose a novel data-driven method to initialize the preference value. We propose to estimate this initial strength for each region as being a cluster centroid by using the Density Peaks (DP) algorithm [105]. The DP algorithm states that the cluster center can be identified as the points that have a higher local density within its neighbour points and are at a larger distance from other higher density points. The density ρ_i of a region i is defined as [105]

$$\rho_i = \sum_{j=1}^{N} f(d_{i,j} - d_c), \tag{5.7}$$

where d_c is a cut off distance, $d_{i,j} = -S(i,j)$ and f is

$$f(x) = \begin{cases} 1, & \text{if } x < 0, \\ 0, & \text{otherwise}, \end{cases} \tag{5.8}$$

δ_i is defined as the minimum distance between the region i and any other region with higher density, which is calculated as

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{i,j}. \tag{5.9}$$

The measure $\rho_i \delta_i$ approximates the strength of a region being a centroid [105]. We use the measure $\rho_i \delta_i$ for each region to scale for N regions and use it as preference for each respective region. Consider $\gamma_i = \rho_i \delta_i$, we initialize the preference value as

$$p(i) = \frac{\gamma_i - \min\{\gamma_1, \ldots \gamma_N\}}{\max\{\gamma_1 \ldots \gamma_N\} - \min\{\gamma_1 \ldots \gamma_N\}} \times (N-1) + c, \tag{5.10}$$

where N is the number of brain regions $(N = 90)$, c is empirically chosen so that when γ_i is minimal, the preference value for the region is initialized as $N/6$, which is a small non-zero number that gives enough local support for initialization of the AP clustering algorithm.

After initializing p, the availability and responsibility messages are updated iteratively. When updating these messages in each iteration, a damping update is applied to each message to avoid possible numerical oscillations. For a particular iteration m, the damping update is applied as

$$a_m(i,j) = (1-\lambda)a_m(i,j) + (\lambda)a_{m-1}(i,j), \tag{5.11}$$

$$r_m(i,j) = (1-\lambda)r_m(i,j) + (\lambda)r_{m-1}(i,j), \tag{5.12}$$

where we initialize $\lambda = 0.5$ as suggested by [88]. Message passing iterations were terminated based upon either i) the maximum number of iterations (I) reached or ii) the centroids remained unchanged for C consecutive iterations. In this work, we use $I = 1500$ and $C = 100$, to allow convergence. We then combine the availability and responsibility messages during iterations to determine the centroids and their points. For any region i, we find the region j

that maximizes $a(i,j)+r(i,j)$ and identify the association of region i as

$$Association(i) = \begin{cases} \text{centroid} \,, & \text{if } i = j \,, \\ i \text{ is a member of centroid } j, & \text{otherwise.} \end{cases} \tag{5.13}$$

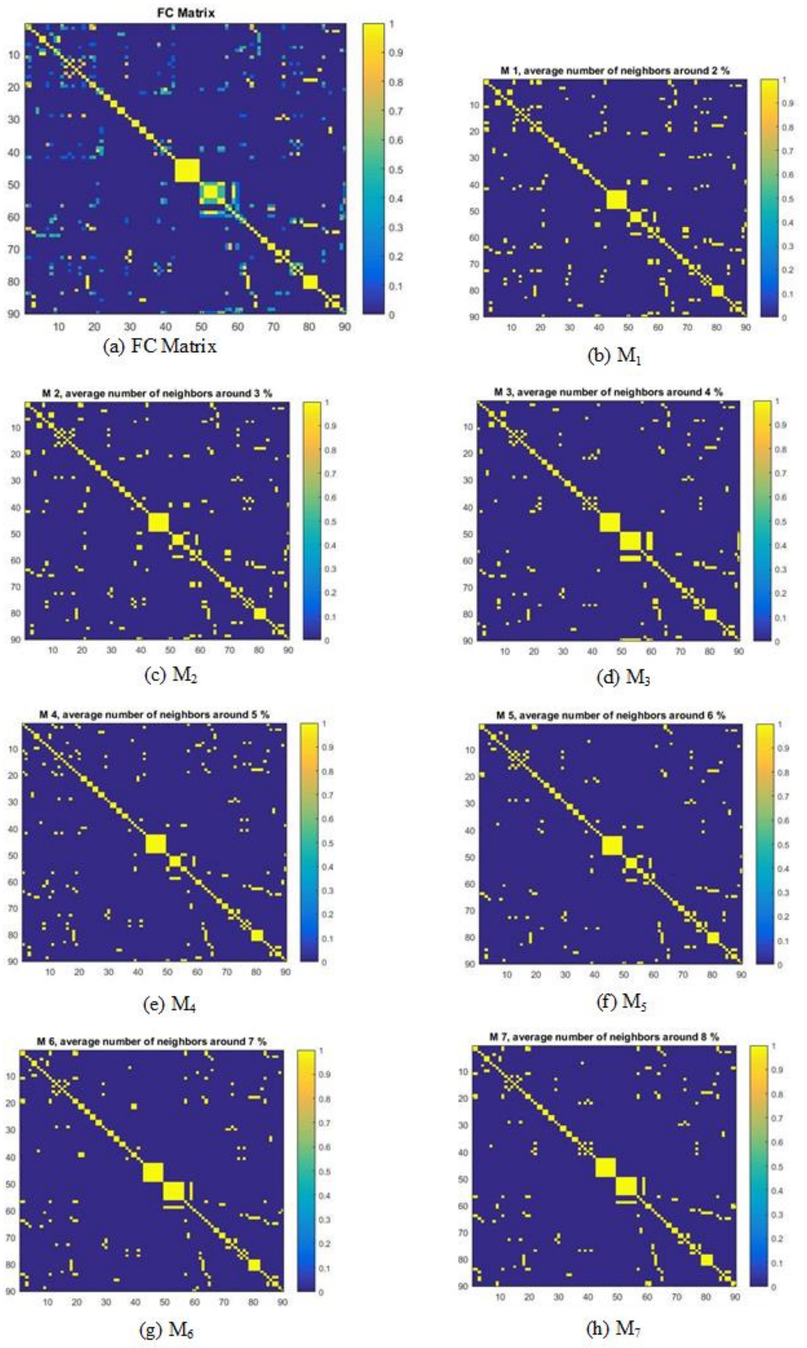
From the AP clustering algorithm results, we construct a matrix M as

$$M(i,j) = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are in the same cluster,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.14}$$

The cutoff distance d_c in Equation 5.7 impacts clustering by varying the preference value computed in Equation 5.10, yielding different clustering results. [105] proposed this cutoff distance to be around 2%. The optimal number of clusters is data dependent, and critically, not known in advance. Rather than fixing a set number of clusters (as in popular clustering algorithms like k-means), we instead select the number of clusters in a data-driven fashion by adjusting this cutoff distance. For a given cutoff distance, the clustering algorithm will produce a clustering of the data. We apply the clustering algorithm multiple times to produce a total of K matrices (each matrix denoted as M), one for each clustering. To achieve this, the cutoff distance is varied sequentially, between 2% and 8% inclusive, of the neighbours to produce multiple clusterings. After these multiple runs of clustering, we calculate a functional connectivity (FC) matrix as

$$FC(i,j) = \frac{1}{K}\sum_{l=1}^{K} M_l(i,j), \tag{5.15}$$

where $K = 7$. The FC and M matrices are visualized in Figure 5.3. The FC matrix represents the functional connectivity, such that each entry in $FC(i,j)$ may be considered as an estimate of the probability that the i^{th} and j^{th} regions belong to the same functional connectivity. The functional connectivity matrix is employed further in feature selection as described in the next section.



(a) FC Matrix

(b) M_1

(c) M_2

(d) M_3

(e) M_4

(f) M_5

(g) M_6

(h) M_7

Fig. 5.3 Visualization of FC and M matrices. M_1 to M_7 are binary matrices calculated using Equation 5.14, and FC is calculated using Equation 5.15. The FC matrix represents the functional connectivity, where values closer to one represent high functional connectivity between corresponding regions and values closer to zero represent no or very low functional connectivity.

5.2.3 Discriminant Feature Selection

The characteristics of fMRI data make feature selection an important step in classification problems [67]. The dimensionality of functional connectivity is typically very large even if functional connectivity is evaluated between the defined region of interests (ROIs) instead of all voxels. The number of functional connectivity alterations related to a particular disorder is very small as compared to all brain connections. If all the functional connectivity values are presented to the classifier as features, it may not yield good classification performance as i) it may cause overfitting, ii) it may provide substantial irrelevant information for the classification task. Therefore, it is considered very important to incorporate a good feature selection strategy to identify possible discriminant functional connectivity features for classification [67].

The functional connectivity matrix from Equation 5.15 has a dimensionality of 4005 $(90 \times 89/2)$ unique features. The high dimension of the matrix may degrade the performance of a classifier (the well known "curse of dimensionality" problem [127]). Also, a small number of functional connectivity features might be altered by ADHD as compared to all functional connectivity features. We are interested in identifying only those altered features, therefore, there is a need to select the discriminant features.

The FC matrix constructed in the earlier step represents the functional connectivity of all brain regions and may contain highly correlated features, as they may belong to the brain networks. We investigate Elastic Net (EN) feature selection [50] for extraction of the discriminant features. The most appealing property of EN is that it encourages grouped selection of features which makes it well suitable for this domain. EN is an embedded-based feature selection algorithm that takes advantages of both LASSO and ridge regressions by combining their penalties in one single solution. Similar to LASSO regression, the L_1 penalty is employed to enable variable selection and continuous shrinkage, and similar to the ridge regression, the L_2 penalty is employed to encourage grouped selection of features. If $\boldsymbol{y}$ is the label vector for subjects, $y_i \in \{l_1, l_2, ...l_n\}$, $l_k \in \{1,2\}$ for $k = \{1,2,...n\}$. This is a binary classification problem, so there are two labels (ADHD or control). Consider

$\boldsymbol{X} = \{FC_1, FC_2, ...FC_n\}$, the cost function to be minimized by the Elastic Net is

$$L(\lambda_1, \lambda_2, \beta) = ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||^2, \tag{5.16}$$

where

$$||\boldsymbol{\beta}||_1 = \sum_{j=1}^{n} |\beta_j|, \tag{5.17}$$

and

$$||\boldsymbol{\beta}||^2 = \sum_{j=1}^{n} (\beta_j)^2, \tag{5.18}$$

where λ_1 and λ_2 are weights of the terms forming the penalty function, and $\boldsymbol{\beta}$ coefficients are calculated through model fitting. If we denote α as

$$\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \tag{5.19}$$

then Equation 5.16 can be written as

$$L(\alpha, \beta) = ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \alpha ||\boldsymbol{\beta}||_1 + (1 - \alpha) ||\boldsymbol{\beta}||^2, \tag{5.20}$$

where $\alpha \varepsilon [0, 1]$ and the function $\alpha ||\boldsymbol{\beta}||_1 + (1 - \alpha) ||\boldsymbol{\beta}||^2$ is called the elastic net penalty which is a combination of ridge and LASSO regressions. The parameter α controls both combinations where $\alpha = 1$ represents LASSO regression and α close to 0 represents ridge regression. Typically, multiple iterations of EN are run in a cross validation setup and the mean-squared error is recorded for each iteration. As a result, the fixed number of features or the features with minimum error are returned. In this work, we use $\alpha = 0.1$ as we are interested in selecting grouped features from a sparse FC matrix. Multiple iterations of EN are run until i) max iterations ($iter = 100$) is reached or ii) all $\boldsymbol{\beta}$ coefficients converge to zero. By minimizing the cost function L in Equation 5.20, we extract the features with non-zero $\boldsymbol{\beta}$ coefficients relating to minimum cross validation error employing the training set. We did not select a fixed number of features from the EN as i) the optimum number of features is not known in the case of fMRI and ii) also as our method was applied on different data sets it was not possible to fix the number of selected features.

Next, we fused together the EN selected features with non-imaging features (age,gender and IQ levels of each subject) to construct a combined feature set for training the classifier. It should be noted that the EN feature selection was applied to the imaging features only and was not applied to the non-imaging data. The combined feature set is employed for classification, as described in the next subsection.

5.2.4 Classification

The final step in our study is classification, where we employ a classifier to evaluate the discriminative ability of the selected features from the previous steps. For this work, we evaluate performance using a Support Vector Machine (SVM) [128] classifier. SVM is a popular machine learning classification algorithm which has resulted in good performance in various neuroimaging studies (e.g., [39, 129–131]). SVM is well suited for tasks where the number of features is large compared to the number of training samples [132]. Given that for this problem we have a large number of features compared to the number of available subjects, SVM is a reasonable choice. During the training phase of the classifier, the labelled training data is presented. In this phase, SVM seeks an optimum boundary with a maximum separating margin between the two classes (healthy control and ADHD). The boundary is defined by a linear combination of the predictor variables. The learned SVM model is then employed in the testing phase by presenting unseen testing data (without labels of subjects). The SVM classifier predicts the label (control or ADHD) for each test subject. Consider $\boldsymbol{y}$ as the label vector for subjects, $y_i \varepsilon (l_1, l_2, ...l_n)$, $l_k \varepsilon \{1,2\}$ for $i = \{1,2,...n\}$ and $\boldsymbol{X} = \{x_1, x_2, ...x_m\}$ as our combined feature vector. The decision function of SVM is given by [133]

$$f(x) = sign\left(\sum_{i=1}^{N}(y_i \lambda_i^{*} \Phi(\boldsymbol{x}, \boldsymbol{x}_i)) + b^{*}\right), \tag{5.21}$$

where $b^* \varepsilon R$, Φ is a kernel function, and λ_i^* is constrained as follows: $0 \leq \lambda_i^* \leq C_1$ for $y_i = 1$ and $0 \leq \lambda_i^* \leq C_2$ for $y_i = 2$ where C_1 and C_2 are penalties for class 1 and 2 respectively. We set $C_1 = 1$ and $C_2 = 1$. For all our experiments, we used Matlab (R2016a) implementation of SVM with a linear kernel.

5.3 Experiments and Results

The proposed framework was evaluated on the dataset provided by the ADHD-200 consortium [106] which contains four categories of subjects: controls, ADHD-combined, ADHD-hyperactive/ impulsive, and ADHD-inattentive. Here we present the task as a binary classification problem, control vs. ADHD, by combining all ADHD subtypes in one category. The number of subjects in the training dataset of each imaging site is presented in Table 5.1. We conducted experiments on the i) training dataset alone and the ii) training and test datasets. For the evaluation of the ADHD-200 consortium dataset, we selected the features from the training data for each individual site using ElasticNet and the selected features were integrated with the non-imaging data for training the SVM classifier. The non-imaging features explored in our work are comprised of age, gender, verbal IQ, performance IQ and full4 IQ. Datasets from two imaging sites (Peking and KKI) were highly imbalanced with the majority class being the control subjects. To avoid imbalance learning in our model, we applied SMOTE on the Peking and KKI datasets as described earlier. It should be noted that the data generated by SMOTE was employed only for training the classifier and not for testing. Also, the parameters of our framework were held constant for all the imaging sites datasets which includes parameters for SMOTE and SVM.

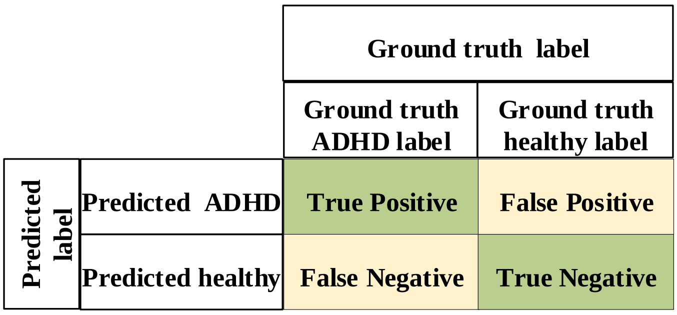
5.3.1 Results on the Training Dataset

In order to evaluate the training dataset, we employed leave-one-out (LOO) cross-validation on the individual imaging sites. Results are presented in Figure 5.5. Let TP, TN, FP and FN denote true positive, true negative, false positive and false negative respectively, as illustrated in Figure 5.4. Sensitivity, specificity and accuracy are defined as $sensitivity = TP/(TP+FN)$,

	Train dataset		Test dataset	
Imaging site	Healthy controls	ADHD	Healthy controls	ADHD
NYU	98	118	12	29
NI	23	25	14	11
Peking	61	24	24	27
KKI	61	22	23	28

Table 5.1 Total Number of control and ADHD subjects for four imaging sites in the training datasets, namely, Kennedy Krieger Institute (KKI), NeuroImage (NI), New York University Medical Center (NYU) and Peking University (Peking).

$specificity = TN/(TN+FP)$ and $accuracy = (TP+TN)/(TP+TN+FP+FN)$. The highest accuracy of 86.7% was achieved on the KKI dataset.



		Ground truth label	
		Ground truth ADHD label	Ground truth healthy label
Predicted label	Predicted ADHD	True Positive	False Positive
	Predicted healthy	False Negative	True Negative

Fig. 5.4 Illustration of true positive, false positive, false negative and true negative. For instance, in the case true positive, an ADHD subject is correctly diagnosed and in the case of false positive, a healthy subject is incorrectly diagnosed as ADHD.

The ADHD-200 consortium did not provide classification results for the training dataset. In state-of-the-art work [5], the authors applied LOO validation on the training dataset to evaluate classification performance. For fair comparison, we also applied LOO validation and the results are presented in 5.2. As shown in the table, our methodology has improved results as compared to Dey et al. [5] in three imaging sites (We could not compare the results for NYU as it was not provided by [5]). We also computed our results without non-imaging data and results are given in Table 5.3. The results show that, except for the KKI dataset, our method gives best performance when compared to the state-of-the-art model.

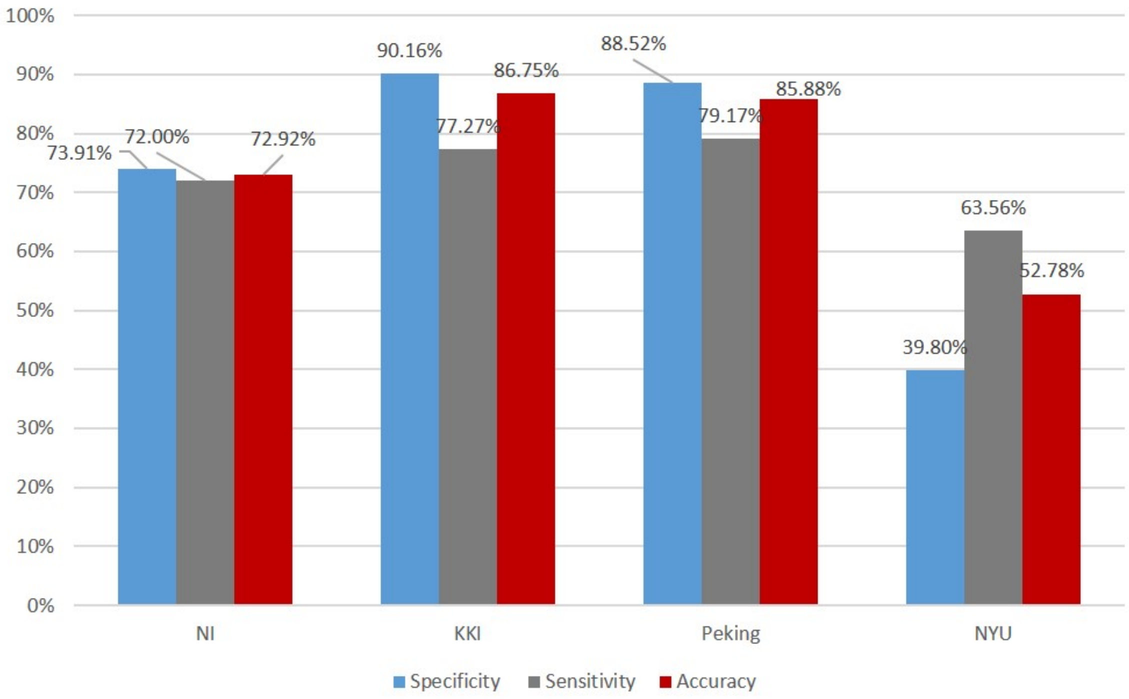


Fig. 5.5 Results on the training dataset. Classification Accuracy, Sensitivity and Specificity attained for the four imaging sites namely KKI, NI, NYU and Peking. Highest classification accuracy of 86.75% was achieved on the KKI dataset.

	Dey et al.[5] Results			Our methodology		
	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
KKI	100%	9.5%	75.6%	90.1%	77.2%	**86.7%**
NI	68.1%	58.8%	64.1%	73.9%	72.0%	**72.9%**
NYU	–	–	–	39.8%	63.5%	**52.7%**
Peking	96.6%	21.1%	61.2%	88.5%	79.1%	**85.8%**

Table 5.2 Comparison of leave-one-out (LOO) results on the training dataset. Our proposed method was able to achieve higher classification accuracy in three datasets as compared to Dey et al.[5].

Results on the Test Dataset

In this experiment, our framework was trained on the training dataset provided for each imaging site. The trained SVM classifier was tested with the independent test data provided for each individual site. We evaluate our results by comparing them against the competition team results reported by NITRC. We compare our results with the highest classification

Name	Accuracy of Dey et al.[5]	Accuracy of fused imaging + non-imaging data	Accuracy without non-imaging data
KKI	75.6%	**86.7%**	67.4%
NI	64.1%	–	**72.9%**
NYU	–	**52.7%**	25.4%
Peking	61.2%	**85.8%**	85.3%

Table 5.3 Comparison of leave-one-out (LOO) results of Dey et al.[5] with our methodology. We calculated our results with i) fusing imaging + non-imaging data and ii) without non-imaging data. (Non-imaging data for NI was not available).

accuracy achieved by teams for individual imaging sites (data from [6]). The results are presented in Table 5.4. It can be seen that our model achieves best accuracy in all the datasets except NI. Lower accuracy for NI may be due to the fewer number of available subjects in the dataset.

Name	Average accuracy	Highest accuracy	Our accuracy	Number of imaging features
Peking	51.0%	58%	**64.7%**	733
KKI	43.1%	81%	**81.8%**	820
NYU	32.3%	56%	**60.9%**	230
NI	56.9%	–	44.0%	346

Table 5.4 Comparison of our results with average results of competition teams and highest accuracy achieved for individual sites. The highest accuracy for NI was not reported by [6]. Our proposed method achieved higher accuracy than the average accuracy of the competition teams for three imaging sites.

In order to explore the impact of non-imaging data towards classification performance in our framework, we computed the performance by comparing the results of fusing non-imaging data with imaging data and without integrating non-imaging data. The results are presented in Table 5.5. It can be seen from the results that the integration of non-imaging data provides better classification results for Peking and NYU as compared to results without non-imaging data.

In order to evaluate the generalization capability of our method we computed the cross-site validation accuracy results. We trained our model on the combined training data set of three

Name	Accuracy with fused imaging + non-imaging data	Accuracy without non-imaging data
Peking	**64.7%**	58.8%
KKI	**81.8%**	**81.8%**
NYU	**60.9%**	24.3%

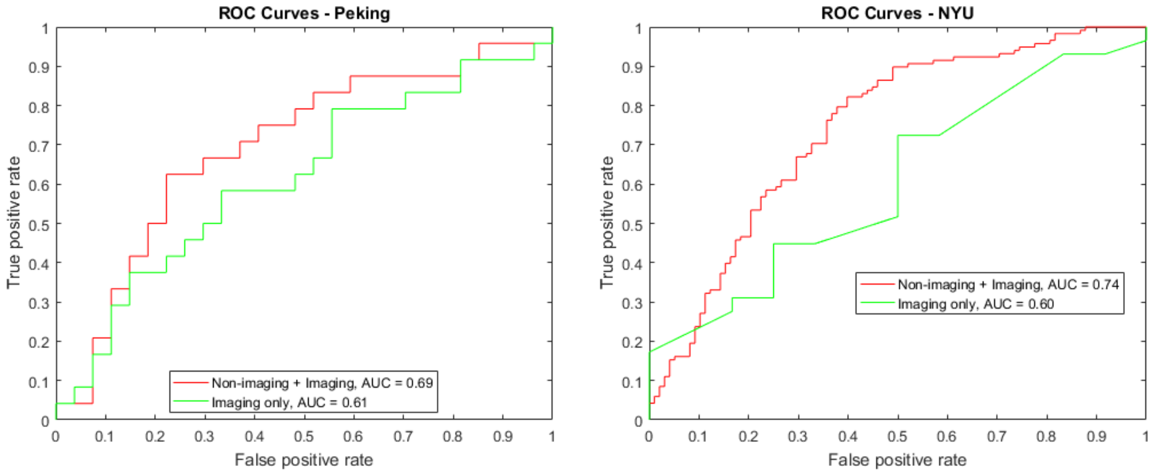
Table 5.5 Comparison of the accuracy results with fusing imaging and non-imaging data and without non-imaging data. The results show that fusing non-imaging data with imaging data provides better accuracy for two imaging sites (Peking and NYU).

imaging sites (KKI, PI and NYU). We did not evaluate NI for this experiment because some non-imaging data was not available for this imaging site. The trained framework was evaluated on each individual imaging site and results are presented in Table 5.6. This is a challenging experiment as the ADHD-200 data set is very heterogeneous. However, the results show that our method was able to attain a comparable accuracy to that attained by training on individual imaging site. One interesting observation is that our method was able to achieve same classification results (81.8%) for all experiments as can be seen in Table 5.5 and Table 5.6. It appears that the SVM was able to find optimum support vectors from the imbalanced imaging data. Therefore, dataset balancing and fusing non-imaging data have no effect on accuracy.

Test data set	Accuracy when trained on each individual imaging site	Accuracy when trained on a combined training data set
Peking	**64.7%**	60.7%
KKI	**81.8%**	**81.8%**
NYU	**60.9%**	56.1%

Table 5.6 Comparison of accuracies of i) trained and tested on each individual imaging site ii) trained once by combining the three training datasets and tested individually for three imaging sites.

Next, we calculated ROC curves for: i) imaging data only and ii) fusing imaging and non-imaging data for Peking and NYU datasets and the results are presented in Figure 5.6. It is clear from the Area Under the Curve (AUC) values that our model yields better results



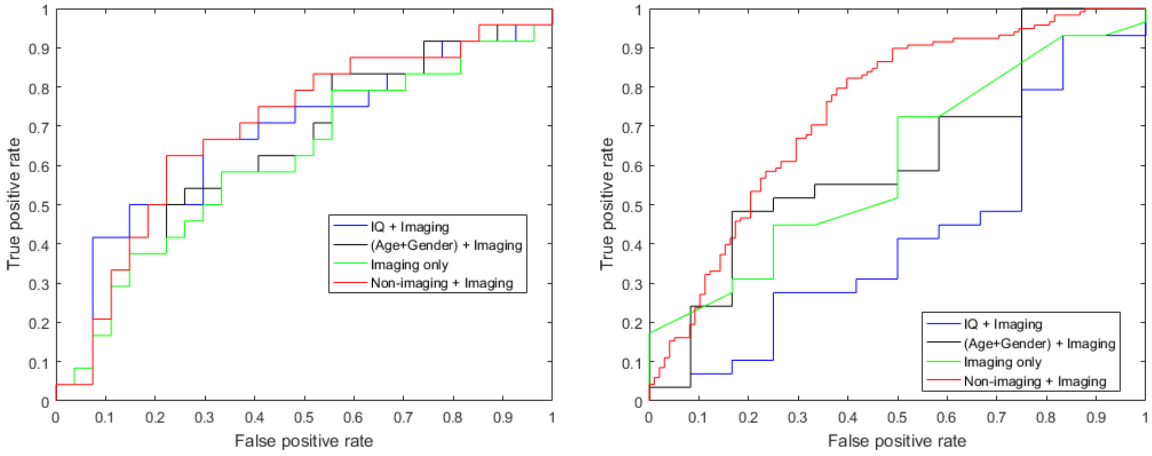
(a) Peking dataset

(b) NYU dataset

Fig. 5.6 ROC curves for Peking and NYU for: i) fusing non-imaging and imaging and ii) imaging only. For Peking, AUC with imaging data is 0.61 and with non-imaging + imaging it is 0.69, and for NYU, AUC is increased from 0.60 to 0.74 with the fusion of non-imaging data which suggests that fusion of non-imaging data yields better performance.

for the fusion of imaging and non-imaging measures (for Peking, AUC for imaging data only=0.61 and AUC for imaging + non-imaging data=0.69, and for NYU, AUC for imaging data only=0.60 and for imaging + non-imaging data=0.74). In order to study the impact of different non-imaging measures towards classification, we calculated ROC curves for Peking and NYU datasets by categorizing non-imaging data in two groups: i) IQ levels and ii) age and gender. The results are presented in Figure 5.7. The ROC curves in the figure compare the results of combining these non-imaging measures with imaging data. The ROC curves for non-imaging + imaging for both imaging sites show better performance as compared to other curves for both imaging sites which suggests that fusion of all the non-imaging measures yield better performance overall.

Finally, in order to evaluate our proposed novel methodology to initialize the AP clustering algorithm as discussed in the previous section, we computed and compared our results with standard AP clustering results. The comparison is presented in Table 5.7. It should be noted that in this comparison all other parameters are held same in both experiments. The accuracy achieved by our proposed methodology is higher when compared to the accuracy achieved by AP clustering for all four imaging sites.



(a) Peking dataset

(b) NYU dataset

Fig. 5.7 ROC curves for different non-imaging measures for Peking and NYU. For both datasets, the ROC curves for i) IQ + Imaging ii) (Age+ Gender) + Imaging iii) All Non-imaging measures + Imaging and iv) Imaging only, are shown. For both imaging sites, ROC curves for non-imaging + imaging (shown in red colour) show better performance as compared to all other three curves, which shows that fusion of all non-imaging measures yields better performance for both datasets.

Name	AP clustering			Proposed methodology		
	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
Peking	81.4%	33.3%	58.8%	92.6%	33.3%	**64.7%**
KKI	87.5%	33.3%	72.7%	75.0%	100.0%	**81.8%**
NYU	41.6%	62.0%	56.1%	41.6%	68.9%	**60.9%**
NI	7.1%	63.6%	32.0%	42.8%	45.4	**44.0%**

Table 5.7 Comparison of our proposed methodology with the AP clustering method for the four imaging sites. The accuracy achieved by our proposed methodology is higher as compared to accuracy achieved by AP clustering for all four imaging sites.

5.4 Anatomical Analysis

Finally, we performed an anatomical analysis of selected features from our framework for all four imaging sites. Selected features for each individual imaging site in our framework represent the altered functional connectivity between control and ADHD subjects. We discuss our findings in terms of i) hemispheric analysis and ii) lobe analysis, which are explained below.

5.4.1 Hemispheric Analysis

The human brain is segmented into the left hemisphere and the right hemisphere. We analysed our selected features with respect to both hemispheres and results are presented in Figure 5.8. For the analysis, each region was mapped into a particular hemisphere. Figure 5.8 suggests that for all four imaging sites, the inter hemispheric functional connectivity is altered the most as compared to individual hemispheres. For Peking and KKI, inter hemispheric alterations constitute 49.7% and 49.3% respectively. The number of alterations belonging to left and right hemispheres is quite close to each other. The results suggest that the functional connectivity between the two hemispheres might be impaired by ADHD.

5.4.2 Lobe Analysis

Next, we discuss our findings in groups of brain lobes as suggested by Salvador et al. [2]. The study identified six brain lobes namely: (i) Medial temporal lobe, (ii) Subcortical lobe, and the four standard Neocortical lobes which are (iii) Occipital lobe, (iv) Frontal lobe, (v) Temporal lobe, and (vi) Parietal (pre) motor lobe. We studied intra lobe alterations for each imaging site by mapping the brain regions to a particular lobe and the results are presented in Figure 5.9. The results in Figure 5.9 suggest that in all four imaging sites, the Frontal lobe is affected the most as compared to the other lobes, followed by the Parietal (pre) motor lobe.

Similarly, we studied functional connectivity alterations in terms of inter lobe alterations for all lobes in individual imaging sites and the results are presented in Figure 5.10. The results suggest that the functional connectivity of the Frontal lobe and the Parietal (pre)

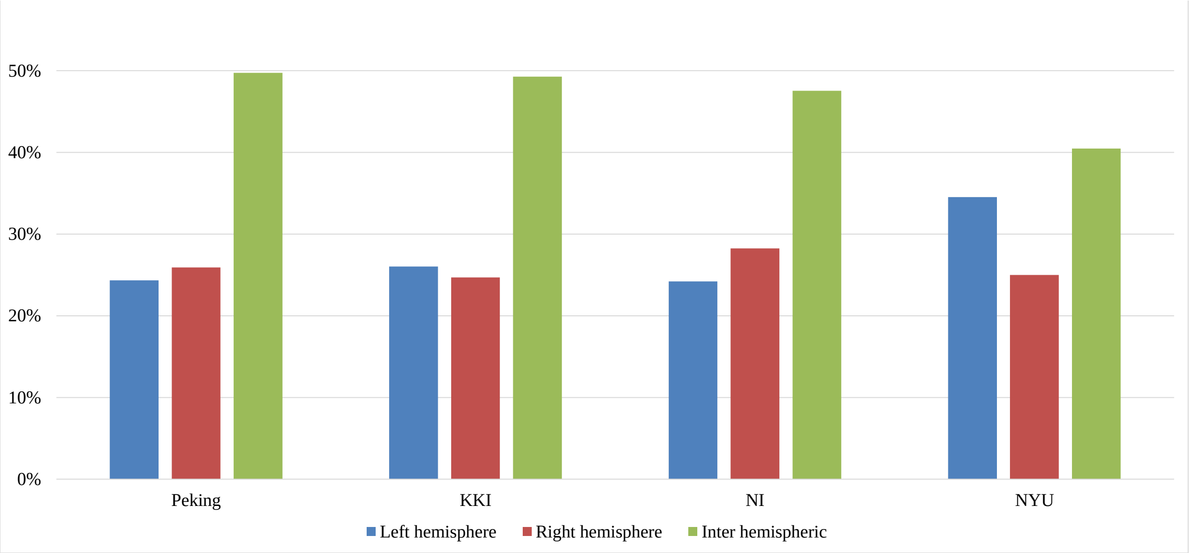
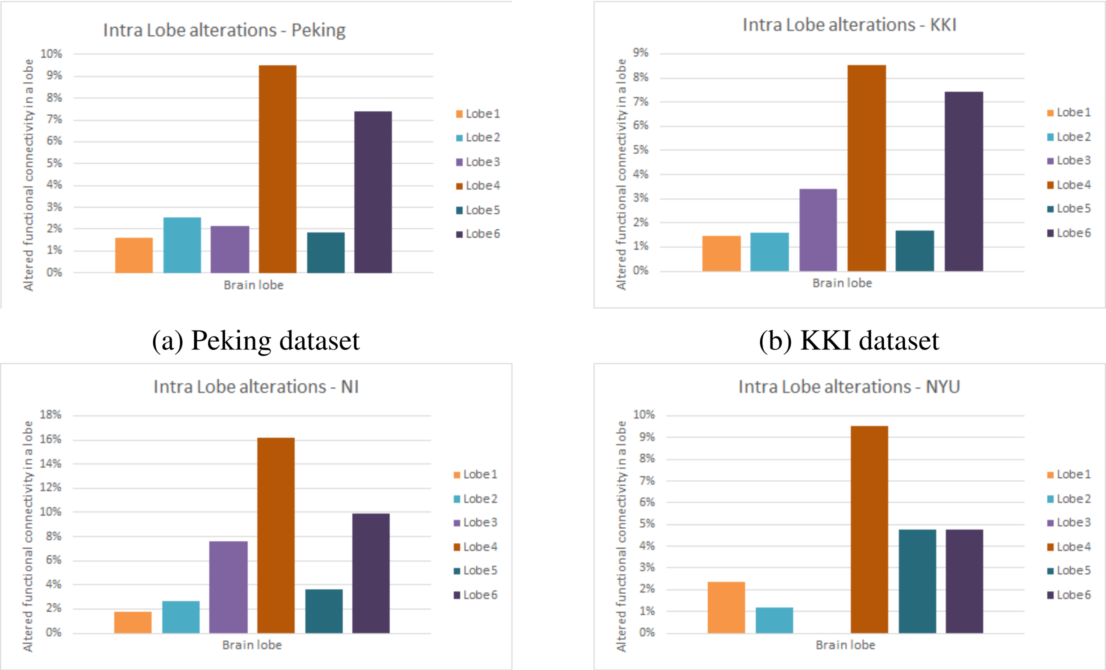


Fig. 5.8 Functional connectivity alterations with respect to brain hemispheres. The results show that for all four imaging sites, the majority of functional connectivity alterations belong to inter hemispheric brain connections.

motor lobe is the most affected. Results of the inter and intra lobe alterations from Figures 5.9 and 5.10 suggest that in ADHD, Frontal and Parietal (pre) motor lobes are affected the most, in terms of inter and intra lobe functional connectivity alterations. The Frontal lobe is associated with a number of critical brain functions such as attention, executive functions (involved with purposeful, goal-directed behaviour), memory, affect and mood [134]. With the alterations in the Frontal lobe, these associated brain functions might be impaired in ADHD subjects. Parietal (pre) motor lobe is known to be associated with movement intention and motor awareness [135]. With the alterations in Parietal (pre) motor lobe, abnormal body activities might be observed.



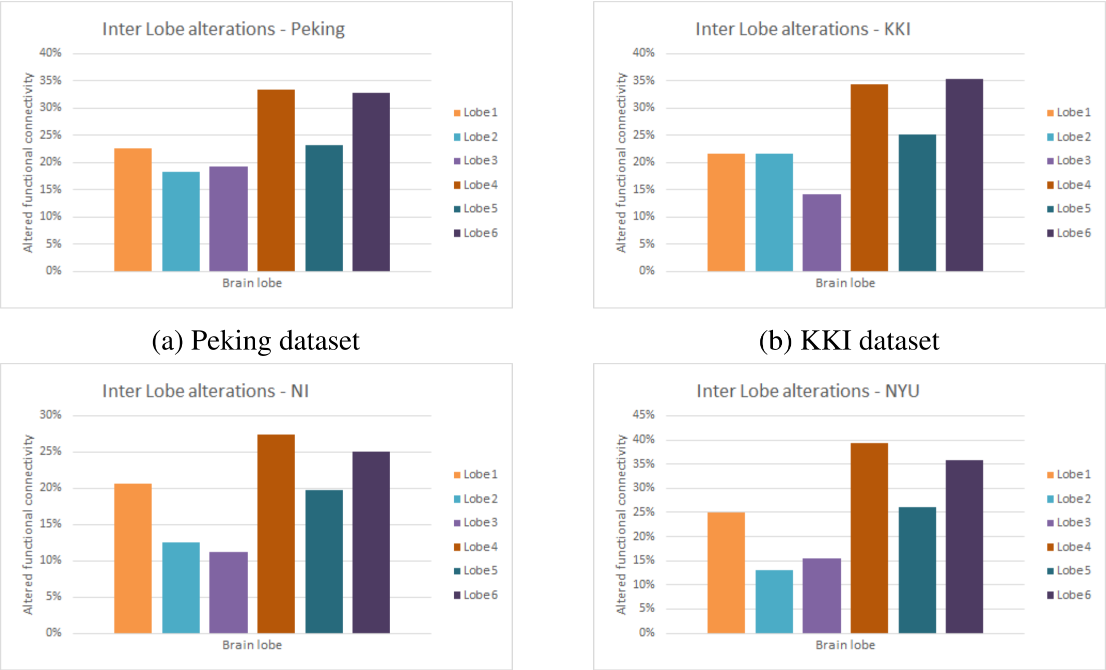
(a) Peking dataset

(b) KKI dataset

(c) NI dataset

(d) NYU dataset

Fig. 5.9 Functional connectivity alterations in terms of intra lobe alterations. Brain lobe groups are segmented by Salvador et al. [2] which are: (Lobe 1) Medial temporal lobe, (Lobe 2) Subcortical lobe, (Lobe 3) Occipital lobe, (Lobe 4) Frontal lobe, (Lobe 5) Temporal lobe, and (Lobe 6) Parietal (pre) motor lobe. For all four imaging sites, the Frontal lobe is affected the most as compared to other lobes.



(a) Peking dataset

(b) KKI dataset

(c) NI dataset

(d) NYU dataset

Fig. 5.10 Functional connectivity alterations in terms of inter lobe alterations. Brain lobe groups are segmented by [2] which are: (Lobe 1) Medial temporal lobe, (Lobe 2) Subcortical lobe, (Lobe 3) Occipital lobe, (Lobe 4) Frontal lobe, (Lobe 5) Temporal lobe, and (Lobe 6) Parietal (pre) motor lobe. For all imaging sites, the Frontal and Parietal (pre) motor lobes are affected the most.

5.5 Conclusions

In this chapter, we have addressed the problem of identification of discriminant features between control and ADHD subjects for classification based on fMRI data. Classification of neuroimaging data is considered a difficult task due to its high dimensionality. We have proposed a machine learning framework for this problem and evaluated our method on four training and test datasets provided by the ADHD-200 consortium [106]. Our framework introduced a novel method for estimation of functional connectivity between brain regions. Functional connectivity between brain regions was determined using the Affinity Propagation (AP) clustering algorithm [88], which does not require a number of clusters in advance. Instead, AP takes a measure of similarity between data points and the initial preference for each point for being the cluster centroid. We propose a novel method to find these cluster centroids through a matrix derived from the Density Peaks (DP) algorithm by Rodriguez and Laio [105]. The number of clusters and the centroids emerge from the data itself. The brain is a complex network where a number of brain regions may show coherent activity. Therefore, discriminant features may be highly correlated with each other. Here, we employed Elastic Net for feature selection that encourages grouped feature selection.

In this work, we also evaluated the importance of non-imaging data by fusing it with the selected features. Our results show that Elastic Net based feature selection integrated with non-imaging data provides an important feature selection strategy. Our selected features and SVM classifier were able to outperform the state-of-the-art in classification accuracy on data from three institutions. Our results also show that in ADHD, inter hemispheric functional connectivity is altered the most as compared to alterations belonging to the individual hemispheres, which further suggests that in ADHD, coordination between the lobes is affected. Our results indicate that the Frontal and Parietal (pre) motor lobes are impaired the most by ADHD.

To conclude, the framework described in this chapter outperformed state-of-the-art methods as described earlier. However, the proposed framework is based on classical machine learning methods. In recent years, deep learning has been shown to outperform classical machine learning methods in a number of domains like image classification, medical image segmentation [136] and object recognition [98]. Therefore, we were interested in exploring whether a deep learning method could be used for functional connectivity analysis and for the classification of a disorder like ADHD. To address this, in the next chapter, we explore deep learning based method for functional connectivity analysis.

Chapter 6

Convolutional Neural Network-based Functional Connectivity

In the previous chapter, we have proposed a classical machine learning framework for classification of ADHD. The method outperforms previous state-of-the-art methods, however, like many modern machine learning techniques, it relies on conventional distance measures as a basic step towards the calculation of functional connectivity. Such measures may not be able to capture the latent characteristics of time-series signals. To overcome this shortcoming, in this chapter, we present a novel convolutional neural network model, FCNet, that calculates functional connectivity directly from fMRI time-series signals. The FCNet consists of a convolutional neural network that computes features from time-series signals and a fully connected network that computes the similarity between the extracted features in a Siamese network architecture. The functional connectivity computed using FCNet is combined with phenotypic information and used to classify individuals as healthy controls or subject with neurological disorders. Experimental results on the publicly available ADHD-200 dataset demonstrate that this innovative framework can improve classification accuracy. This indicates that the features learnt from the FCNet have superior discriminative power.

Following publication is related to this chapter:

- Atif Riaz, Muhammad Asad, S M Masudur Rahman Al Arif, Eduardo Alonso, Danai Dima, Philip Corr and Greg Slabaugh, “FCNet: A Convolutional Neural Network for

Calculating Functional Connectivity from functional MRI", 1st International Workshop on Connectomics in NeuroImaging (CNI), MICCAI 2017, Proceedings (Vol. 10511, p. 70). Springer.

6.1 Overview

Several methods have been developed for extracting functional connectivity (FC) from temporal resting state fMRI data such as correlation measures [102], clustering [7] and graph measures [5]. Most existing techniques, including modern machine learning methods like clustering, rely on conventional distance-based measures for calculating the strength of similarity between brain region signals. These measures may not be able to capture the inherent characteristics of time-series signals.

Convolutional neural networks (CNNs) have been shown to outperform existing hand-crafted features-based methods in a number of domains like image classification, image segmentation and object recognition [98]. The strength of a CNN comes from its representation learning capabilities, where the most discriminative features are learned during training. A CNN is composed of multiple modules, where each module learns the representation from one lower level to a higher, more abstract level. To our knowledge, CNNs have not been investigated to determine FC of brain regions. In this work, our motivation is to construct FC patterns from fMRI data by exploiting the representation learning capability of a CNN. Particularly, we are interested in determining whether a CNN can capture the latent characteristics of brain signals. Compared with other methods, our approach calculates FC directly from time-series signals pairs, naturally preserving the inherent characteristics of time-series signal in the constructed FC.

For training, FCNet requires pairs of fMRI signals and a real value indicating the degree of FC. Training data (comprising pair of fMRI signals and their degree of FC) is produced using a generator that selects pairs of time-series signals that are considered functionally connected, and those that are not. This data is used to train a Siamese network [137] to predict FC from an input signal pair. We demonstrate the expressive power of the features extracted from the FCNet in a classification framework that classifies individuals as healthy control or disorder subjects.

The proposed framework has several stages and is illustrated in Figure 6.1. The first stage is to train the proposed FCNet using the data generated by a data generator (Figure 6.1(a)). The FCNet learns to infer FC between brain regions. Once the FCNet is trained, the next step is to use FCs to distinguish healthy control and disorder subjects. This is accomplished by the classification pathways (Figure 6.1(b) and 6.1(c)). During training, the fMRI signals from a training subject are fed into the trained FCNet, which generates a FC map of the brain regions. After the FCNet computes functional connectivity, the remaining processing follows a similar approach as the last chapter: an Elastic Net [50] for variable shrinkage and feature selection and SVM for classification.

The contributions of this chapter are:

- A novel CNN model for estimation of functional connectivity from fMRI signals.
- A learnable similarity measure for calculation of functional connectivity.
- Improved classification accuracy over the state-of-the-art on the ADHD-200 dataset.

In the next section, we describe the details of the method.

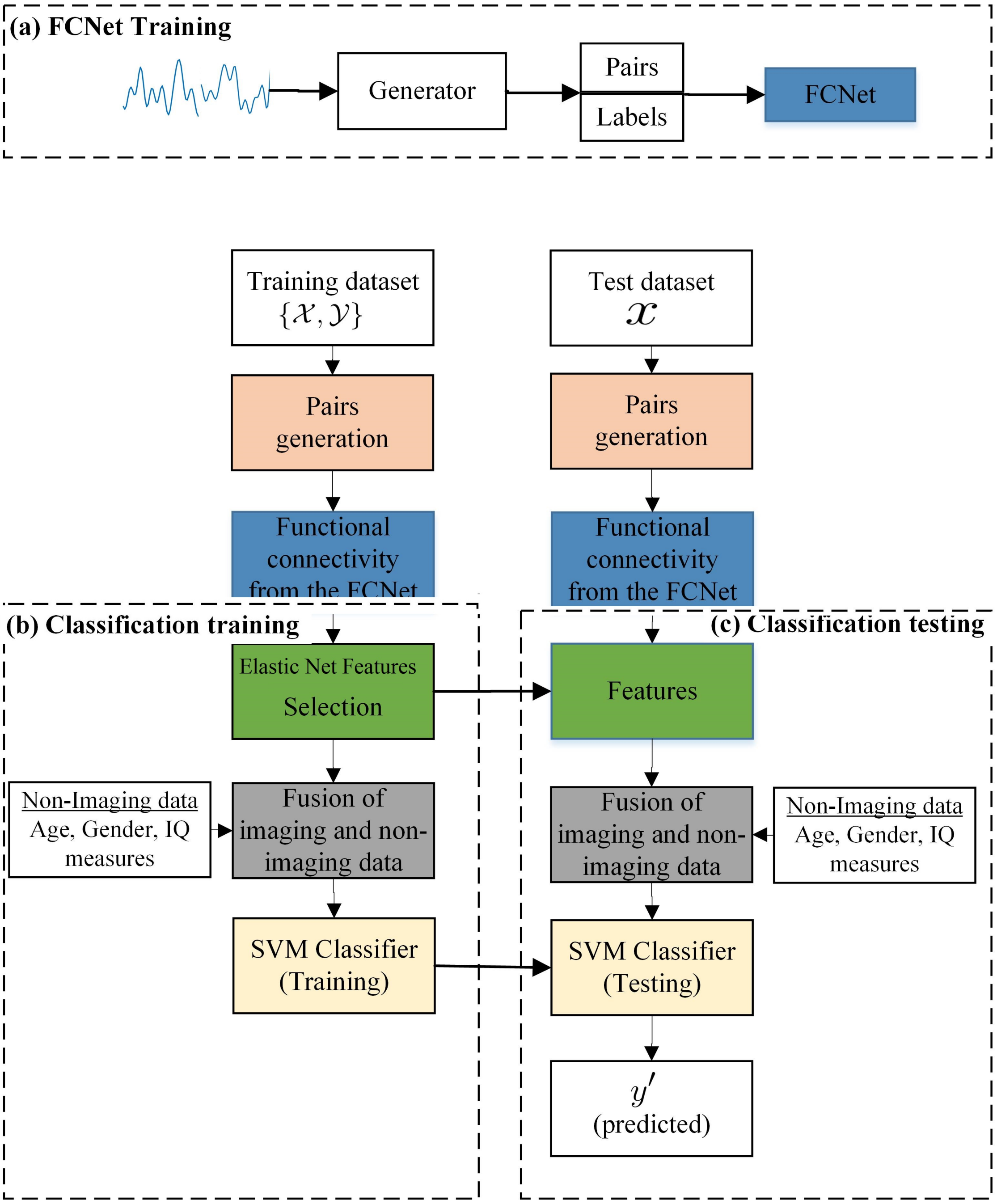

(a) FCNet Training
Generator
Pairs
Labels
FCNet
Training dataset
$\{\mathcal{X},\mathcal{Y}\}$
Test dataset
$\mathcal{X}$
Pairs generation
Pairs generation
Functional connectivity from the FCNet
Functional connectivity from the FCNet
(b) Classification training
(c) Classification testing
Elastic Net Features Selection
Features
Non-Imaging data
Age, Gender, IQ measures
Fusion of imaging and non-imaging data
Fusion of imaging and non-imaging data
Non-Imaging data
Age, Gender, IQ measures
SVM Classifier (Training)
SVM Classifier (Testing)
y'
(predicted)

Fig. 6.1 Flowchart of the proposed method. In (a), the FCNet is trained from the data generated by the generator. In the training pipeline (b), functional connectivity (FC) is generated through the FCNet. Next, discriminant features are selected and fused with the non-imaging data, then employed to train a SVM classifier. The testing pipeline is shown in (c). After FC is calculated, features are selected and fused with the non-imaging data. A trained SVM is employed for classification.

	NYU	NI	Peking
Slices	33	37	33
TR (ms)	2000	1960	2000
TE (ms)	15	40	30
Thickness (mm)	4.0	3.0	3.5
FoV read (mm)	240	224	200
FoV phase (%)	80	100	100
Flip angle (degree)	90	80	90

Table 6.1 Scan parameters per imaging sites.

6.2 Method

6.2.1 Data and Preprocessing

The resting state fMRI data used in this study is from the NeuroBureau ADHD-200 competition [106] as in the previous chapter. For the evaluation of network proposed in this chapter, we used datasets from three imaging sites: NeuroImage (NI), New York University Medical Center (NYU), and Peking University (Peking). The deep learning methodology employed in this work requires a fixed length of input signal and can not accept input with different number of input lengths. To decide for a fixed number of input length, we selected the imaging site with highest number of subjects. The site with maximum number of subjects was NYU with 222 subjects and the length of time-series signals was 172. Therefore, we designed our network to accept input length of 172. We discarded the imaging sites with length of time-series smaller than this number. Also, the time-series of length greater than 172 were truncated to make fixed length of input signals. All the imaging sites have a different number of subjects.

The scan parameters and the equipment used were not necessarily consistent across different imaging sites. Scan parameters used by different imaging sites are presented in Table 6.1.

For all our experiments, we used preprocessed data released for the competition [109].

6.2.2 Functional Connectivity through FCNet

In this work, we propose a Siamese network-based [137] deep CNN for the calculation of FC. Our proposed method calculates FC directly from the time-series signals instead of relying on conventional similarity measures like correlation or distance-based measures. In the 1900s, Siamese networks were first proposed for signature verification [137]. After this pioneering study, Siamese network have been applied to face verification [138, 139], local patch descriptor learning [140, 141], ground-to-aerial image matching [142], stereo matching [143] and target tracking [144]. In this chapter, we propose a Siamese network architecture, called FCNet, to learn a robust and generic representation of fMRI time-series signals, and to calculate similarity from these representations. Similarity serves as the functional connectivity measure in our work. FCNet is a deep network architecture for jointly learning to extract features from the individual regional time-series signals and a learnable similarity network that calculates the similarity between the pairs. The architecture of the proposed Siamese network is presented in Figure 6.2.

In the figure, x_1 and x_2 is a pair of fMRI time-series signals. Both the time-series signals are fed into the feature extractor networks. Here, w is the shared parameter that is subject to learning during the training phase and the feature extractor networks map the original signals x_1 and x_2 into $f_w(x_1)$ and $f_w(x_2)$, respectively. The similarity measure network calculates the degree of similarity between the mapped features $f_w(x_1)$ and $f_w(x_2)$, which is presented as functional connectivity between the pair of regions x_1 and x_2. The FCNet is presented in Figure 6.3 and individual networks are specified detailed below.

The feature extractor network:

In order to map the fMRI time-series signals to a low dimensional space and hence to formulate a learned similarity metric, we design two identical convolutional neural networks with a shared parameter set (Figure 6.2), feature extractor network. The main advantage of using CNNs is that they can learn local features and can build a robust and abstract representation of the data [98]. The feature extractor network is a multi-layer, trainable and non-linear system that can operate at the time-series level and learn abstract representations in

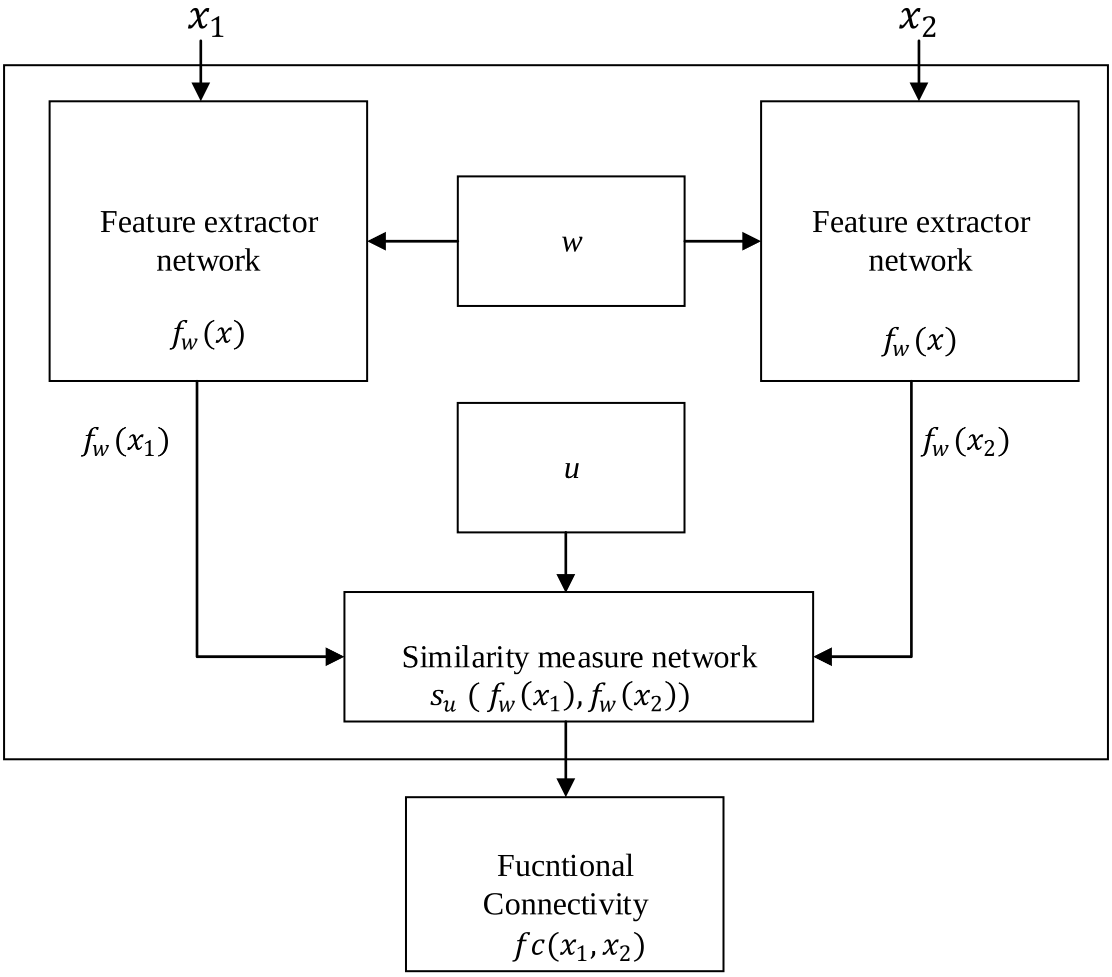

x_1
x_2
Feature extractor network
$f_w(x)$
w
Feature extractor network
$f_w(x)$
$f_w(x_1)$
u
$f_w(x_2)$
Similarity measure network
$s_u\ (f_w(x_1), f_w(x_2))$
Fucntional Connectivity
$fc(x_1, x_2)$

Fig. 6.2 The architecture of the FCNet. x_1 and x_2 is a pair of time-series signals and w is the shared parameters between the two feature extractor networks. The feature extractor networks map the time-series signal to the abstracted features that are passed into the similarity measure network. The similarity measure network calculates the similarity between these abstracted features, which is presented as functional connectivity. u is the set of parameters of the similarity network. Both u and w are learned during training.

an integrated manner. The feature extractor network is trained end-to-end to map time-series to outputs. The network architecture is inspired by [3], which was originally designed for images. We have adopted the network for one-dimensional time series data. We introduced batch normalization layers. Here, we use a Leaky Rectified Linear Unit (ReLU) as the non-linearity function, due to its faster convergence over ReLU [145]. The network is

Layer Name	Kernel size	Stride	Number of filters
Convolutional 1	3	1	32
Convolutional 2	3	1	64
Convolutional 3	3	1	96
Convolutional 4	3	1	64
Convolutional 5	3	1	64

Table 6.2 The details of the convolutional layers of Figure 6.3.

designed to accept a time-series signal of length 172. The details of the network architecture are given in Figure 6.3(b). All pooling layers pool temporally with pool length of 2 and stride of 2. The details of the convolutional layers are presented in Table 6.2. The last fully connected layer in the network has 32 nodes.

The similarity measure network:

The network is designed to calculate similarity between the two time-series signals. The output of the network represents the functional connectivity between two brain regions, as the functionally connected brain regions will show a high similarity value and vice versa. This network employs a neural network to learn the FC between *pairs* of extracted features from two brain regions. This is in contrast to conventional methods that use hand-crafted computations like correlation or distance-based measures. The input to this network is the abstracted features extracted from two regions. The network computes their FC, which relates to the similarity between the two regions. The network is comprised of three fully connected layers where the last layer is connected to a softmax classifier with dense connections. The number of nodes in the three layers is 32, 32 and 2, respectively. The network is presented in Figure 6.3(e). Next, we describe architectural considerations and training.

Two-stream architecture with shared parameters:

In our work, the intention is to learn a similar feature extraction rule from both brain regions of a pair. In other words, in order to calculate the similarity (which can be viewed as functional connectivity) between different pairs of brain regions, the brain regions must undergo the same feature extraction processing. It can be realized by employing the two feature extractor networks (two-stream architecture) with the constraint that both networks share the same set of parameters. The set of parameters for both streams is presented as w in Figure 6.2. The parameter set w is learned during the training phase. During this phase, updates are applied to the shared parameters. The approach is similar to a Siamese network [137], which is used to measure the similarity between two images.

Data generator for training FCNet:

For training FCNet, we require similar (functionally connected) and dissimilar (not functionally connected) regions with corresponding labels (one and zero respectively). We develop a generator to generate pairs of brain regions using support from Affinity Propagation [88] clustering for labelling training pairs. The Affinity Propagation algorithm groups regions into clusters based on their temporal activity. Therefore, regions with similar functional activity are grouped in one cluster regardless of their spatial distance or locality. We make pairs for regions that lie in the same cluster and assign them label one (functionally connected). For unconnected pairs (regions that are not functionally connected), we randomly pick regions that do not belong to the same cluster and label the pair zero. The data generator generates balanced numbers of connected and unconnected pairs. The procedure is detailed in Algorithm 1.

Training of FCNet:

Teh FCNet is trained on pair-wise signals with labels generated from the generator as described above. The FCNet is trained end-to-end using a two-stream architecture minimizing the cross-entropy loss

$$L_{fc} = -\frac{1}{n}\sum_{1}^{n}[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)], \tag{6.1}$$

where n is the number of training samples, y_i is the label of pairs (1 for functionally connected and 0 for unconnected regions) and $\hat{y}_i$ is the prediction by the softmax layer.

Algorithm 1: Data generation for training of the FCNet.

```
Input: X
X is the subjects in training data.
nReg is the number of regions = 90.
Output: (Pairs, Labels)
Pairs contains pairs of the time-series signals of brain regions and Labels contains 1
or 0. Pairs and Labels are used for training of FCNet.
for each $x$ in X do
    $c \leftarrow$ cluster($x$) % clustering results in $c$
    for $i \leftarrow 1$ to nReg do
        count $\leftarrow 0$
        for each $j$ in $(1 \rightarrow nReg)$ such that $c(x_i) = c(x_j)$ and $i \neq j$ do
            AddToPairs(($x_i$,$x_j$), Pairs)
            AddToLabels(1, Labels)
            count $\leftarrow$ count + 1
        end

        for $k \leftarrow 1$ to count do
            $r \leftarrow$ RandomSelectRegion($x$) such that c($x_i$) $\neq$ c($r$)
            AddToPairs(($x_i$,$r$), Pairs)
            AddToLabels(0, Labels)
        end
    end
end
return (Pairs,Labels)
```

To evaluate FC through the FCNet, regions belonging to each subject are grouped into pairs (for 90 regions belonging to a subject 4005 unique pairs are created). The pairs are passed to the trained FCNet, which computes FC for each pair.

6.2.3 Feature Selection and Classification

The high dimensionality of fMRI data makes feature selection as important step in classification problems. Typically, the dimension of the features (functional connectivity) are of the order of thousands and the number of available subjects is of the order of hundreds. The number of altered functional connectivity belonging to a particular disorder is very small as compared to all functional connectivity values. If all the features are presented to a classifier for prediction, it will introduce the well known “curse of dimensionality" problem, yielding poor classification performance. Overfitting is another problem that may be introduced if all features are employed. Therefore, a good feature selection strategy is considered as an important step in a classification problem [67]. There are a number of different feature selection strategies lying under filter, wrapper and embedded-based feature selection that can be applied to select the most discriminant features from functional connectivity.

In this work, we have applied an Elastic Net (EN) [50] based feature selection to select discriminant features from functional connectivity. The details of the feature selection step are the same as discussed in the previous chapter. The number of features selected by EN for the Peking, NYU and NI are 523, 96 and 205 respectively.

Similar to the previous chapter, non-imaging features (age, gender and IQ levels) of the subjects are concatenated with the EN selected features to construct a combined feature set for classification. EN feature selection is only applied to the imaging features. The combined set of imaging and non-imaging features is employed for classification.

The final step in the proposed framework is to classify the features coming from the previous steps into control and healthy classes. Similar to the previous chapter, we evaluate a support vector machine (SVM) classifier [133]. The SVM is considered as a popular choice for classification in a number of neuroimaging studies [131, 146–148]. SVM is considered

	NI	Peking	NYU
Average accuracy [106]	56.9%	51.0%	35.1%
Highest accuracy [6]	–	58%	56%
Clustering method [7]	44%	65%	61%
Correlation	52.0 %	52.9%	56.1%
Proposed method	**64.0%**	**68.6%**	**63.4%**

Table 6.3 Comparison of FCNet with the average results of competition teams, highest accuracy achieved for individual site, correlation based FC and clustering based results from the previous chapter [7]. The highest accuracy for NI was not quoted by [6].

to be well suited to deal with problems where the number of features is large as compared to the number of training samples [132].

During the training phase, the SVM is presented with data along with the labels of the data (labels for healthy control and ADHD subjects). During training, the SVM seeks an optimum and maximum separating boundary between the two classes (healthy controls and ADHD). The learned SVM model is then employed for the testing phase where it is presented with the testing data (testing data is not used in the training phase). The SVM classifier predicts the label (healthy control or ADHD) for each subject. Here, we use the Matlab (R2016) implementation of SVM with a linear kernel to evaluate our results.

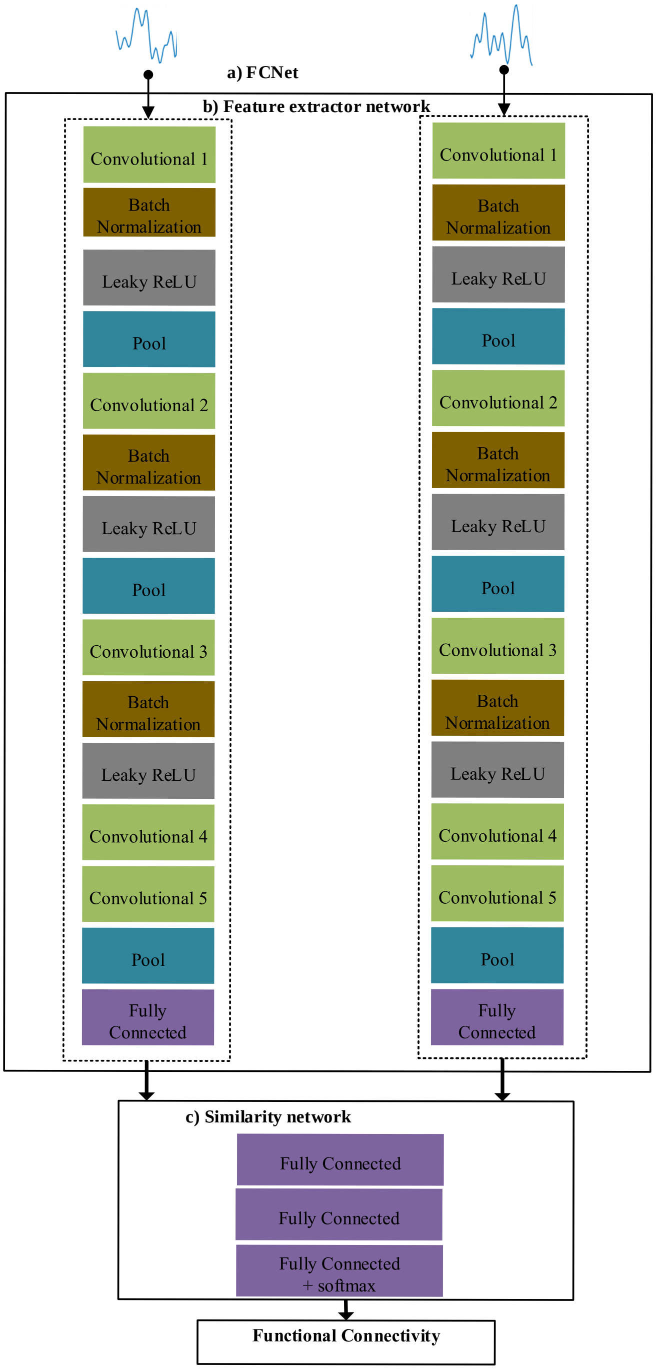


Fig. 6.3 Detailed architecture of the FCNet. The architecture is inspired by [3]. We have modified the architecture for one dimensional data. (a) FCNet with a coupled feature extractor network (one network for each brain region) and the similarity network, which measures the degree of similarity between two regions. (b) The feature extractor network which includes multiple layers namely Convolutional, Batch Normalization, Pooling (pool), Fully Connected and Leaky-ReLU. (c) The similarity measure network.

6.3 Experiments and Results

The proposed framework is evaluated on a dataset provided by the ADHD-200 consortium and contains four categories of subjects: controls, ADHD combined, ADHD hyperactive-impulsive and ADHD inattentive. Here we combine all ADHD subtypes in one category due to two factors, i) we want to investigate classification between healthy control and ADHD, and ii) there are very less number of samples in the ADHD sub categories.

In many biomedical domains specifically fMRI, scarcity of data emerges as one of the most challenging tasks. To address this issue, we combine all subjects from the training datasets of the different imaging sites and FCNet is trained on this combined training dataset. However, feature selection and classification is evaluated on individual imaging datasets. The FCNet is developed in Python with the Tensorflow library. Total pairs of brain regions to train FCNet are 303596. Number of epochs are 30 with a batch size 100, and Adam optimizer is used to optimized the weights of the network.

In order to evaluate the ADHD-200 consortium dataset, our model is trained and tested on each individual imaging sites and the processing pipeline of training and testing is presented in Figure 6.1. In the training phase (presented in Figure 6.1 (b)), for each subject, pairs are generated for all time-series of brain regions. There are 4005 unique pairs for 90 brain regions. The trained FCNet is employed to calculate functional connectivity from these pairs. Discriminant features are selected by the EN from functional connectivity and non-imaging features are fused with imaging features to create a final feature set. The non-imaging features explored in this work are the same as in the previous chapter, and are comprised of age, gender, verbal IQ, performance IQ, and full4 IQ. In the final step, these fused features are presented to the SVM classifier for training. Once our model is trained for an individual imaging site, it is evaluated on the testing dataset of that imaging site. The testing pipeline is presented in Figure 6.1(c) where unseen test data from individual imaging sites is presented to the model (without a label). At the end of the testing pipeline, the trained SVM predicts the label of the data. The pipeline is applied for each individual site (training and testing our model on each imaging site individually) and the results are presented in Table 6.3.

Phenotypic information	Method	NI	Peking	NYU
Not used	Previous chapter	44%	58.8%	24.3%
	Proposed method	60.0%	62.7%	58.5%
Used	Previous chapter	–	65%	61%
	Proposed method	**64.0%**	**68.6%**	**63.4%**

Table 6.4 Comparison of the proposed method with state-of-the-art results (previous chapter): The results suggest that FCNet outperforms state-of-the-art classification accuracy. The phenotypic information relates to the non-imaging features (Age, gender and IQ levels).

The results show that our method outperforms the average accuracy results of competition teams (data from the competition website [106]), highest accuracy for any individual site (from [6]) and correlation-based FC results. For correlation based results, FC is calculated through correlation and the rest of the processing pipeline is the same as our method. It is worth noting that the parameters of our framework are held constant for all the imaging datasets. Our method also performed well in comparison with the results of the previous chapter. In order to compare with the results of the previous chapter, we compare and present the results in Table 6.4, which shows that the FCNet performs well in all of the three imaging sites as compared to the results of the previous chapter.

In order to study the generalization capability of our model, we performed an experiment to calculate cross-site validation accuracy results. In this experiment, we trained our model on the combined training data set of all three imaging sites (NI, Peking and NYU). For this experiment, we trained the model without using non-imaging features because non-imaging data of IQ levels was not available for the NI dataset. Once the model was trained on the combined dataset, it was evaluated on each individual imaging site and the results are presented in Table 6.5. The results show that our model was able to achieve comparable performance to that attained by training on the individual site. The model trained on the combined dataset was able to achieve high accuracy for the NYU dataset. This may be due to the fact that the NYU dataset has a higher number of subjects as compared to other imaging sites.

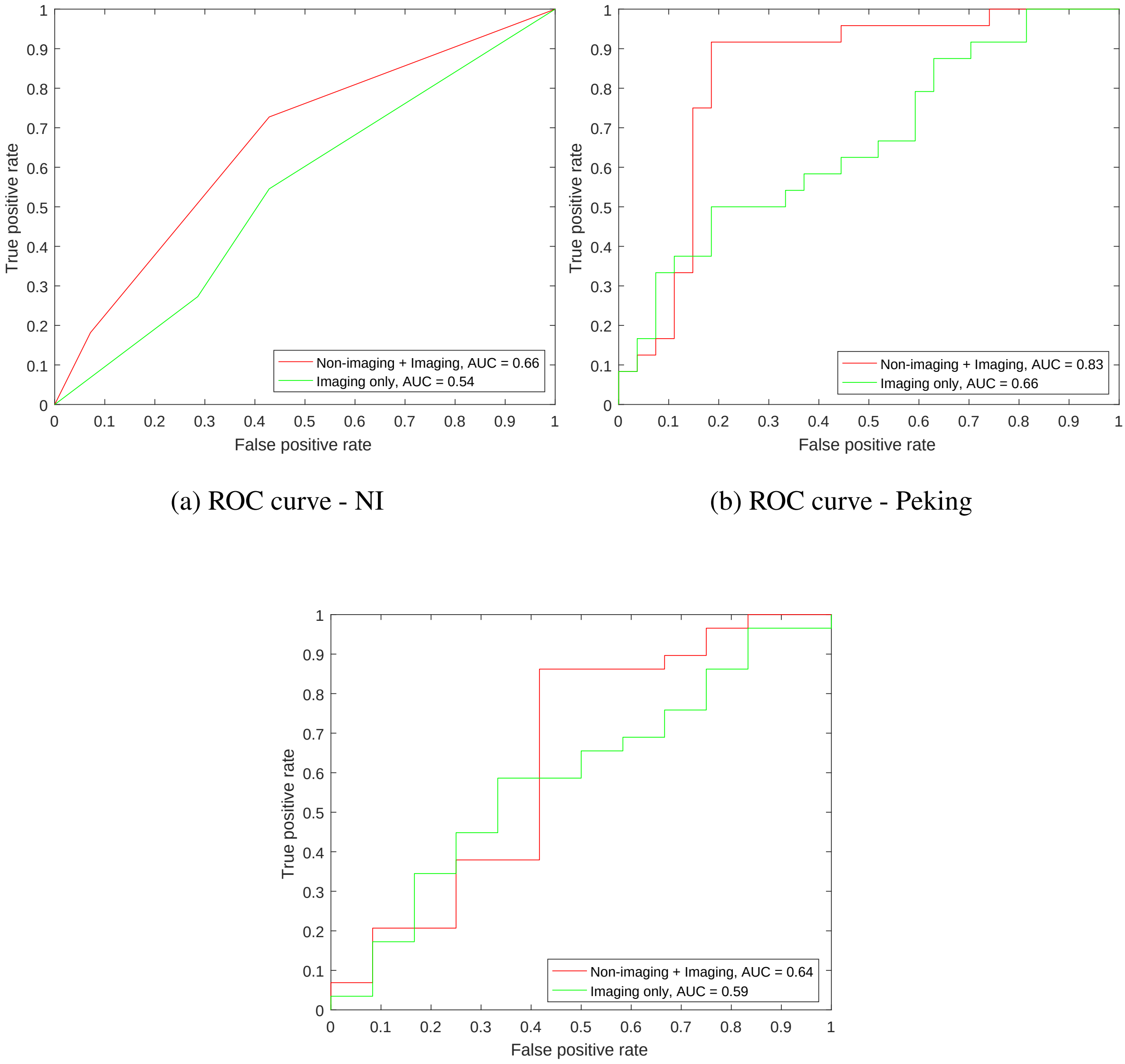
Test data set	Accuracy when trained on each individual imaging site	Accuracy when trained on the combined training data set
NI	**60.0%**	56.0%
Peking	**62.7%**	60.7%
NYU	58.5%	**70.7%**

Table 6.5 Comparison of accuracies of i) trained and tested on each individual imaging site ii) trained once on the combined training data set of three imaging sites (NI, Peking and NYU) and tested individually on the three imaging sites.

Next, in order to study the impact of non-imaging features on the results, we calculated the ROC curves for i) imaging data only and ii) fusing imaging and non-imaging data for all three imaging sites. Please note that for the NI dataset, non-imaging data of IQ levels was not available. The results for all three imaging sites are presented in Figure 6.4. For all the imaging sites, Area Under the Curve (AUC) values with non-imaging + imaging results are higher than the AUC values with imaging features only. The results show that the fusion of non-imaging features yields better performance.

In order to explore the impact of different non-imaging features towards classification results, we calculated the ROC curves for the Peking and NYU datasets by categorizing the non-imaging features in two categories: i) age and gender, and ii) IQ levels. We could not perform this analysis for the NI dataset as non-imaging feature of IQ levels was not available for the NI dataset. The results for both sites are presented in Figure 6.5. The ROC curves in the figure compare the results of combining these non-imaging features with the imaging data. The ROC curves for combined non-imaging + imaging features shows better performance as compared to other curves for both sides which shows that fusion of all the non-imaging features yields better performance.

Next, we analyzed our selected features with respect to the brain hemispheres. The human brain is segmented in two hemispheres: the right hemisphere and left hemisphere. For this analysis, regions belonging to each selected feature (functional connectivity) were mapped into the particular hemisphere. The result of this analysis is presented in Figure 6.6. The figure shows that for all imaging sites, functional connectivity belonging to the inter



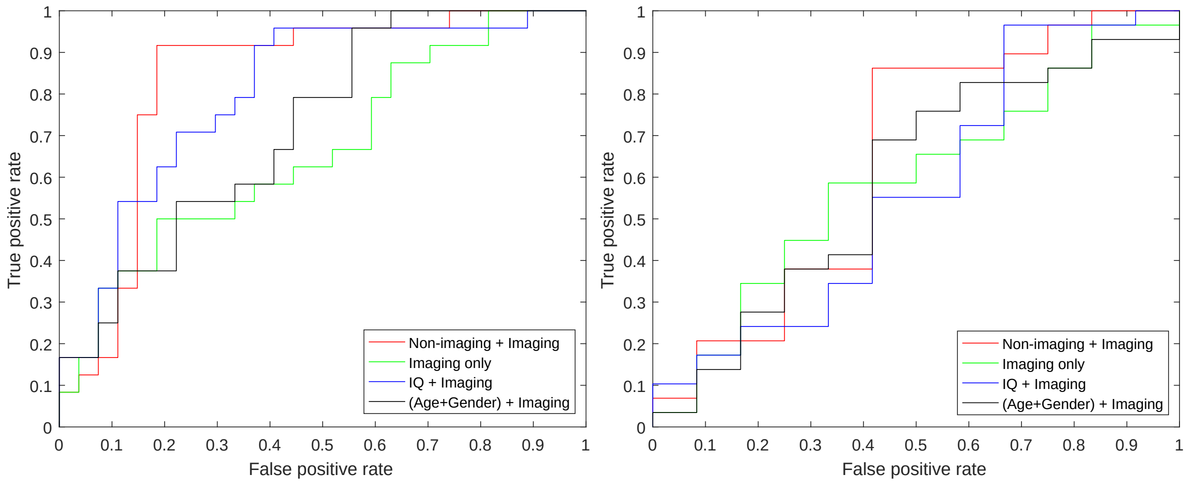
(a) ROC curve - NI

(b) ROC curve - Peking

(c) ROC curve - NYU

Fig. 6.4 ROC curves for NI, Peking and NYU for i) fusing non-imaging and imaging data, and ii) imaging data only. It is clear that the area under the curve (AUC) values for all three imaging sites is higher for non-imging + imaging features. AUC values for the Peking dataset shows the highest difference i.e. the largest AUC: for imaging only data is 0.66 and for non-imaging + imaging data is 0.83.

hemispheric regions is altered the most as compared to the individual hemisphere. For all imaging sites, the inter hemispheric alterations constitute around 50%. The findings suggest



(a) ROC curves - Peking

(b) ROC curves - NYU

Fig. 6.5 ROC curves for different non-imaging features for the Peking and NYU imaging sites. For both datasets, the ROC curves for i) all Non-imaging measures + Imaging, ii) imaging only, iii) IQ + imaging and iv) (Age + Gender) + Imaging, are shown. For both imaging sites, ROC curves for non-imaging + imaging (shown by red colour) show better performance as compared to all other three curves, which shows that fusion of all non-imaging measures yields better performance for both datasets.

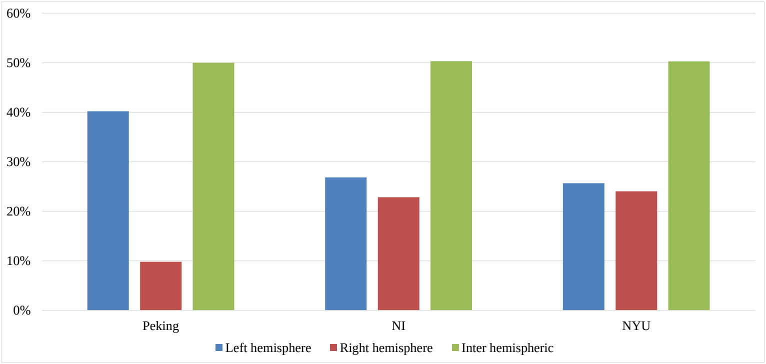
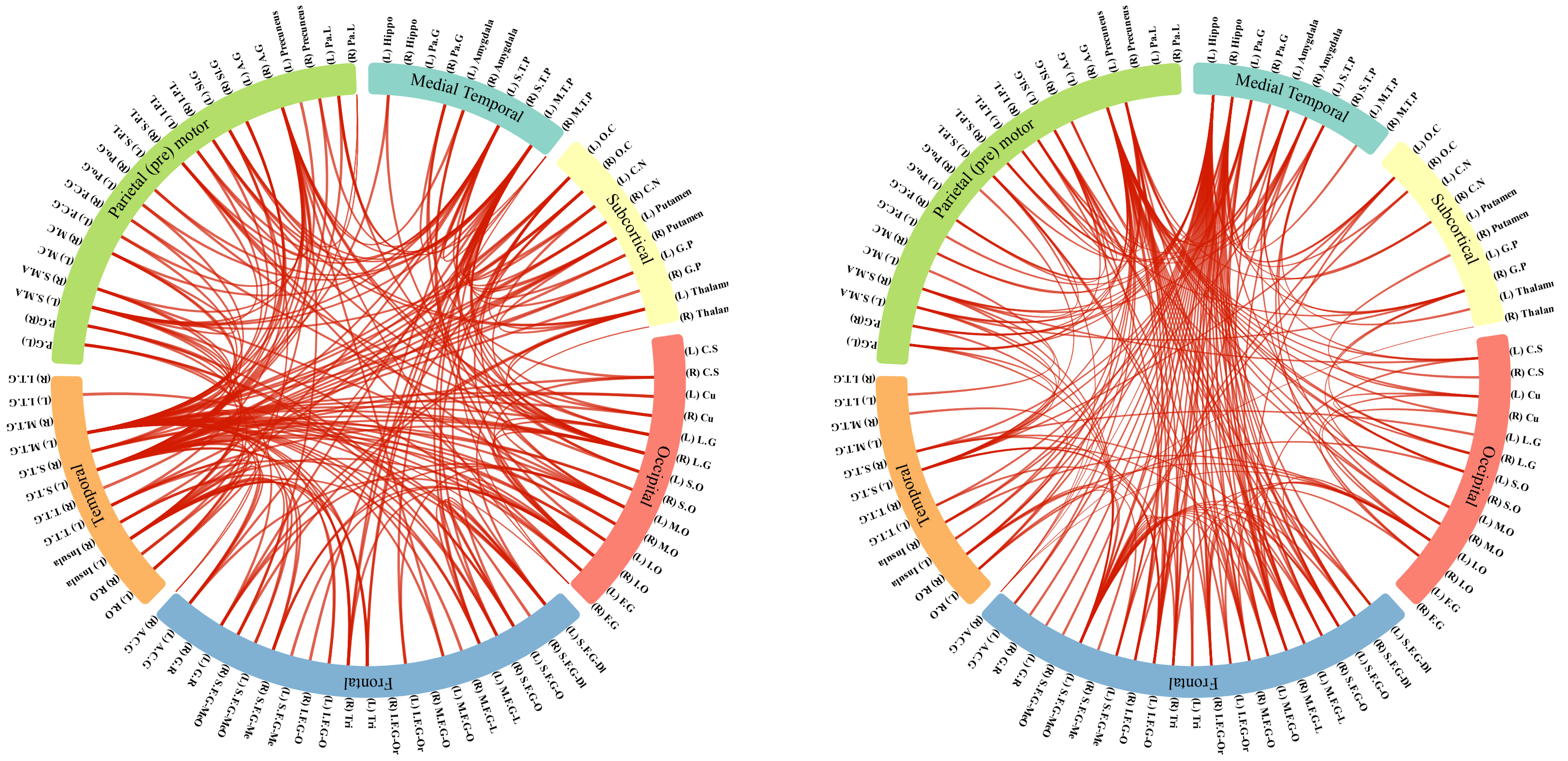


Fig. 6.6 Functional connectivity alterations with respect to brain hemispheres. The results show that for all imaging sites, the majority of functional connectivity alterations belong to inter hemispheric brain connections.

that functional connectivity between the two hemispheres is altered the most in the case of ADHD.

Finally, in order to study functional connectivity differences between the healthy control group and the ADHD group, we visualize their respective functional connectivity patterns using the Peking dataset and present the results in Fig 6.7.



(a) FC patterns of the healthy control group.

(b) FC patterns of the ADHD group.

Fig. 6.7 Comparison of mean functional connectivity (FC) of the healthy control group (a) and the ADHD group (b) for the Peking dataset. For the sake of clarity, only the top 200 connections (based upon their connectivity strength) from both groups are presented. The FC patterns show alterations.

6.4 Conclusion

In this chapter, we have proposed a deep learning method to address the problem of identification of discriminant features between healthy control and ADHD subjects. The high dimensionality of fMRI data makes this problem a challenging task. Machine learning is probably the best available tool to address such a hard task. However, most machine learning techniques rely on conventional distance measures as a basic step towards the calculation of functional connectivity. Such measures may not be able to capture the latent characteristics of time-series signals. Recently, the Convolutional Neural Networks have emerged as a powerful deep learning model which has shown to outperform existing hand-crafted features extraction methods in a number of domains. In this chapter, we have presented a novel Convolutional Neural Network model, FCNet, that takes pre-processed fMRI time series signals as input and calculates functional connectivity.

The FCNet is comprised of a feature extractor network that extracts features from time-series signals and a learnable similarity measure network that calculates the similarity between regions. The FCNet is an end-to-end trainable network. Input to the FCNet is a pair of time-series signals and it yields the functional connectivity of the brain regions belonging to those time-series signals. The next step is the selection of discriminant features from the calculated functional connectivity. Similar to the previous chapter, we use an Elastic Net to select discriminant features. The main advantage of using Elastic Net is that it encourages grouped selection of features, which is most suitable in the case of fMRI as functional connectivity may contain correlated features that belong to brain functional networks. The selected features are combined with non-imaging features to make a final feature set. Finally, an SVM classifier is used to classify individuals as healthy controls or neurological disorder subjects. Experimental results on the publicly available ADHD-200 dataset demonstrate that this innovative framework can improve classification accuracy, which indicates that the features learned from FCNet have superior discriminative power. Our results also suggest that in ADHD, inter hemispheric functional connectivity is altered the most as compared to alterations belonging to the individual hemispheres, which indicates that in ADHD coordination between the lobes is affected. Our results highlight that the Frontal lobe is impaired the most in the case of ADHD.

Although this method provides better accuracy than state-of-the-art methods, still it relies on classical machine learning methods such as Elastic Net and SVM for final prediction. In recent years, end-to-end trainable methods have been shown to outperform classical machine learning methods in a number of domains like image classification, optical character recognition [149], image segmentation and object recognition [98]. We were interested in exploring whether an end-to-end deep learning model can yield better performance as compared to the classical machine learning models in the domain of fMRI. Towards this objective, we explore a deep learning model in the next chapter that takes preprocessed fMRI time-series signals as input and provides the prediction label as its output.

Chapter 7

End-to-end Deep Learning for Classification of ADHD using fMRI

In the previous chapter, we proposed a deep learning method to calculate functional connectivity. The framework was dependent on classical machine learning methods of feature selection and classification, namely, Elastic Net and Support Vector Machine classifier. In this chapter, we propose an end-to-end trainable model that utilizes the representation learning capability of deep learning to classify ADHD from preprocessed fMRI time-series data. Our aim is to apply deep learning techniques to (1) automatically classify a subject as ADHD or healthy control directly from fMRI time-series signals, and (2) evaluate the importance of functional connectivity in an end-to-end deep neural network. The proposed model is comprised of three networks, namely (1) a feature extractor, (2) a functional connectivity network, and (3) a classification network. The model takes preprocessed fMRI time-series signals as input and outputs the predicted labels, and is trained end-to-end using back-propagation. Our results suggest that functional connectivity serves as an important biomarker towards classification of ADHD. Experimental results on the publicly available ADHD-200 dataset demonstrate that this innovative model outperforms previous state-of-the-art. Results suggest that the frontal lobe contains the most discriminative power towards classification of ADHD.

In this work, we propose a deep learning based model, namely, DeepFMRI for prediction of ADHD. The DeepFMRI consists of an end-to-end trainable network that takes fMRI time-series signals as input and produces predicted label as its output. The proposed architecture incorporates a functional connectivity network which is designed to capture pair-wise region connectivity. The second component is a classifier that takes as input all pairs of functional connectivity, and produces a final prediction. The contributions of this chapter include:

- a deep learning architecture, trained end-to-end, for the classification of ADHD.
- demonstration of the importance of functional connectivity for improved results.
- improved classification accuracy on the ADHD-200 dataset.

Following publications are related to this chapter:

- Atif Riaz, Muhammad Asad, Eduardo Alonso, Greg Slabaugh, “DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI", Journal of Neuroscience Methods.
- Atif Riaz, Muhammad Asad, S M Masudur Rahman Al Arif, Eduardo Alonso, Danai Dima, Philip Corr and Greg Slabaugh, “Deep FMRI: An end-to-end deep network for classification of fMRI data”, IEEE International Symposium on Biomedical Imaging (ISBI), 2018.

7.1 Data and Preprocessing

The resting state fMRI data used in this study is from the NeuroBureau ADHD-200 competition [106] as in previous chapters. For the evaluation of network proposed in this chapter, we used datasets from three imaging sites: NeuroImage (NI), New York University Medical Center (NYU), and Peking University (Peking). The details of the data are same as in the Chapter 6.

7.2 Method

7.2.1 End-to-end Model

In this chapter, we propose an end-to-end deep learning model for the classification of ADHD that takes fMRI time-series signals as input and predicts a label (1 for ADHD subject and 0 for healthy control) as output. The proposed work is built on the top of the FCNet [8] as presented in the previous chapter. FCNet is used to extract functional connectivity from fMRI time-series signals, however it combines deep learning and classical machine learning and is not trained end-to-end. For ease of understanding, our proposed architecture can be divided into three modules: 1) feature extractor network, 2) functional connectivity network, and 3) classification network. The feature extractor network is applied to a pre-processed time-series signal of an individual brain region and it produces an abstracted feature as its output. These features are learned during training. The functional connectivity network takes the abstracted features as input and produces the strength of similarity between any two brain regions. Finally, the classification network produces the final prediction label based on the functional connectivity values of all brain regions. We describe the details of each individual network below.

The feature extractor network

This convolutional neural network (CNN) extracts features from *individual* brain region's preprocessed time-series signals and is comprised of multiple layers, namely, conolutional layer, pooling layer and batch normalization layer. CNNs are popular for their capability to learn features and their ability to build robust and abstract representations of data [98].

The network parameters and architecture are inspired by [3], which was originally designed for images. We have adopted the network and parameters for one-dimensional time series data for this specific problem. The network is designed to accept signals of length 172 as input and produces an abstract representation (vector of size 32). The network architecture is presented in Figure 7.1(a) and is comprised of multiple layers (presented in Figure 7.1(d)).

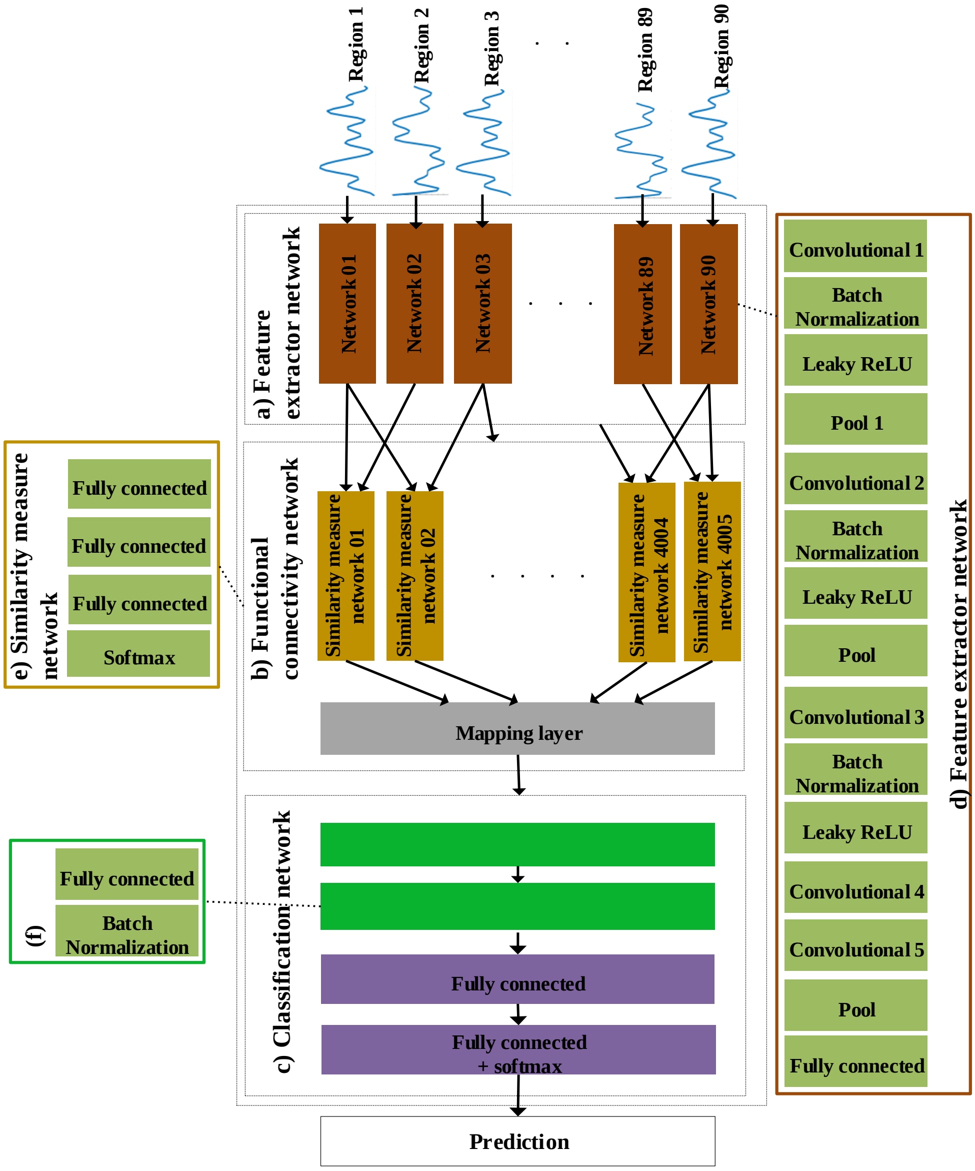


Fig. 7.1 The architecture of the proposed end-to-end model: a) represents a set of 90 feature extractor networks, where each network is applied on each individual region *R*. All networks share the same parameter set. b) represents a functional connectivity network comprising of a set of 4005 similarity measure networks. Each network's input contains abstracted features of two brain regions. All networks share the same parameter set. c) is the classification network comprising of fully connected layers and a softmax layer. d) represents the layers in the feature extractor network. Similarly, e) represents the layer architecture of the similarity measure network, and f) represents the layers of an individual block in the classification network (the two blocks in the classification network have the same architecture but they do not share parameters).

All convolutional layers are one dimensional with a kernel size of 3 and the numbers of filters are 32, 64, 96, 64, 64, respectively, as presented in Figure 7.1. The stride is 1 for all convolutional layers. All max-pooling layers pool temporally with pool length of 2. The last fully connected layer in the network has 32 nodes. The time-series data of all brain regions are passed through the feature extractor network and the output of the network is used as input for the functional connectivity network described in the next section.

The functional connectivity network

The functional connectivity network determines the functional connectivity between brain regions and is presented in Figure 7.1(b). The network is comprised of multiple similarity measure networks where the architecture of each network is the same as the similarity measure network of the FCNet as discussed in the previous chapter. The network is presented in Figure 7.1(e). This Siamese-inspired similarity measure network determines the similarity between *pairs* of extracted features from two brain regions. Here, the calculated similarity measure serves as the degree of functional connectivity between two regions. Each similarity measure network operates on two brain regions, where the input to each network are the abstracted features of the two brain regions from the feature extractor network. The neural network learns to identify functionally connected regions using a non-linear function. This function is learned from the data and is more specific to this particular problem compared to hand-crafted features calculated through generic measures like correlation. The similarity measure network is comprised of three fully connected layers, where the last layer is connected to a softmax layer with dense connections. These layers are presented in Figure 7.1(e). The output of the similarity measure network is a length two vector, and can be interpreted as the probability the two regions are functionally connected, and the complement of the probability.

The outputs of the similarity measure network are fed into a mapping layer using the following operation:

$$M(i) = w_1 v_1^i + w_2 v_2^i, \tag{7.1}$$

where v_1^i and v_2^i are the scalar outputs of the i^{th} similarity measure network, w_1 and w_2 are the weights such that $w_1 + w_2 = 1$. We use $w_1 = 1$ and $w_2 = 0$ that enforces to pass the functional connectivity to the classification network (v_1^i contains functional connectivity). The output of this network can be assumed to be the functional connectivity mapping of all the brain regions, and uses deep learning-based features from the feature extractor network.

Instead of initializing the weights of the feature extractor network and of the similarity measure network randomly, we use the weights of the pre-trained FCNet [8] from previous chapter. The output of this network is fed into the classification network, which is presented in the next section.

Classification network

This neural network produces the final classification results. The input to this network is the output of the mapping layer features (M) representing the functional connectivity of all brain regions. The network produces prediction labels (healthy control or ADHD) as its output.

The network is comprised of two batch normalization layers and four fully connected layers where the last layer is connected to a softmax classifier with fully connected layers. The network parameters are optimized during the training phase of the model. The network is presented in Figure 7.1(c). The number of nodes in the fully connected layers are 100, 50, 50 and 2 respectively, where the final softmax layer yields the output of the model.

Next, we describe the architectural considerations and the training of our proposed model.

Shared parameters architecture

The architecture of the feature extractor network and the similarity measure network is inspired by FCNet, as presented in the previous chapter. However, the FCNet architecture cannot be applied directly to construct an end-to-end network as it is designed to calculate functional connectivity. In the architecture of the proposed work, the same feature extraction steps are applied to individual brain regions, and all pairs of brain regions are passed through the same similarity measure network. This is implemented by employing n_f feature extractor networks and n_s similarity measure networks. Each feature extractor network is applied to an individual brain region ($n_f = 90$), converting individual time-series data into an abstract representation. All the feature extractor networks share the same parameters and updates are applied to these shared parameters during training. The similarity measure network is applied to all combinations of pairs of brain regions, so $n_s = 4005$ $(n_f \times (n_f - 1)/2)$. All the similarity measure networks are implemented with the constraint that the networks share the same parameters and updates are applied to these shared parameters. The approach is similar to a Siamese network [137].

7.3 Experimental Settings and Results

In this section, we evaluate the effectiveness of the DeepFMRI for ADHD classification employing resting-state fMRI and by comparing our results with those of state-of-the-art methods in the literature.

7.3.1 Experimental Settings

The proposed model is evaluated on the ADHD-200 dataset. This publicly available dataset was contributed by different imaging sites. Each imaging site provided separate training and testing datasets. For the evaluation of our method on individual sites, we train our end-to-end model on the training dataset of each imaging site and test it on the corresponding test dataset of that individual site. There are four categories of subjects in the dataset: healthy control, ADHD combined, ADHD hyperactive-impulsive and ADHD inattentive. Here, we combine all ADHD types in one category as we are interested to investigate classification between healthy control and ADHD only.

The proposed model was created in Python using the Tensorflow library. The network is trained end-to-end. The Adam optimizer [150] is used to optimize the network and the number of epochs are 50. After 50 epochs, the training loss converges and becomes stable. For the initialization of the feature extractor and similarity measure networks, we use weights from the pre-trained FCNet [8], and these weights are updated through fine-tuning. The training time for Peking, NYU and NI imaging sites was approximately 1 hour, 5 hours and 1 hour, respectively.

The full deep network is trained the end-to-end model with the following cross-entropy loss:

$$L = -\frac{1}{n}\sum_{1}^{n}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)], \tag{7.2}$$

where n is the number of training samples, y_i is the ground truth label of the subject (1 for ADHD subject and 0 for healthy control) and $\hat{y}_i$ is the prediction.

As the feature extraction and similarity measure networks are initialized with a pre-trained FCNet, we employ different learning rates for i) the feature extraction and similarity measure networks (10^{-5}), and ii) for the classification network (10^{-4}).

7.3.2 Comparison Methods

To validate the effectiveness of the DeepFMRI, we compare it with different network architectures and state-of-the-art method, an end-to-end network without functional connectivity, clustering based method and FCNet from the previous chapters and a correlation method. We describe overview of these methods in next sections.

End-to-end Model without Functional Connectivity

A number of studies have shown that functional connectivity plays a key role in the cognitive processes of the brain [70]. Recently, studies have shown that altered functional connectivity can serve as an important biomarker towards the identification and classification of different brain disorders [1, 5, 39, 40, 151–153]. Inspired by such findings, we have integrated a functional connectivity network in the proposed architecture. The functional connectivity network calculates functional connectivity measure in the DeepFMRI model as discussed earlier.

In order to evaluate the importance of functional connectivity in our work towards the classification of ADHD, we have evaluated our end-to-end network without the functional connectivity network. The model without the functional connectivity network is presented in Figure 7.2.

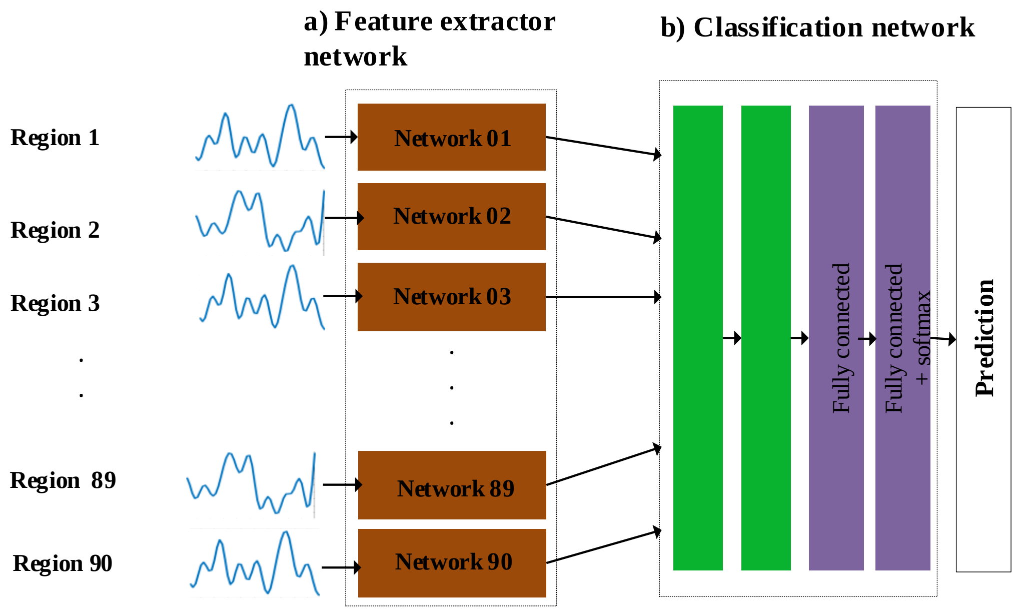


Fig. 7.2 The end-to-end model without the functional connectivity network. a) represents a set of 90 feature extractor networks where each network is applied to each individual region *R*. b) is the classification network.

In this model, the abstracted features calculated through the feature extraction network are merged and passed directly to the classification network and there is no functional connectivity network. Due to the exclusion of the functional connectivity network, there are fewer overall parameters than in the proposed model. The weights and parameters of the feature extraction network are the same as in the proposed network.

FCNet

The FCNet method [8], as detailed in Chapter 6, uses a CNN to extract functional connectivity from the pre-processed fMRI signals. An Elastic Net [50] is applied to extract the discriminant features from the calculated functional connectivity and finally an SVM classifier is applied to evaluate the classification results. This is the first method that applies a CNN on time-series signals, incorporating functional connectivity for the classification of ADHD.

Correlation Method

Correlation is a popular method for calculating functional connectivity. In order to compare the DeepFMRI with correlation, we performed correlation on pre-processed fMRI signals to calculate functional connectivity between the brain regions. We applied an Elastic Net based feature selection to extract discriminant features. Finally, an SVM classifier was applied for classification.

Clustering Method

A clustering-based approach for calculating functional connectivity of brain regions is used in [7] and discussed earlier in Chapter 5. Clustering is considered a more sophisticated technique than correlation-based techniques for calculating functional connectivity [40] as the network obtained by clustering is sparse [1, 87]. In this work, functional connectivity is calculated through clustering. An Elastic Net [50] is applied to functional connectivity to extract discriminant features. Finally an SVM classifier is utilized to classify healthy vs ADHD subjects.

7.3.3 Feature Importance of Functional Connectivity

A common criticism of deep networks is that they are a ‘black box’, mapping inputs to outputs and lacking interpretability. In a clinical context, it is of keen interest to not just produce diagnoses, but also draw some insights from the network itself, particularly looking for differences between healthy control and patient groups to characterise the neurological condition. A key advantage of the proposed method is that due to the functional connectivity network, once the model is trained, we can analyse the functional connectivity of brain regions for patients and control, leading to interpretable results. As a demonstration, we carried out an experiment to rank the contribution of individual functional connectivity values towards prediction of a particular class label (in our case, class labels are healthy control and ADHD). This weighted rank can be viewed as feature importance of individual functional connectivity towards predicting a class label.

In our end-to-end network, the final prediction is calculated through the classification network. The classification network is comprised of multiple layers where it gets the functional connectivity from the mapping layer as input and produces the final prediction of the network (i.e. control or ADHD) through a softmax layer. During the training step, the network optimizes the parameters with respect to the individual class label. The network back-propagates the error from the last layer to the mapping layer (reminiscent of functional connectivity in our network) during the training phase. Thus, the learned weights of this network carry important information towards determining the importance of functional connectivity for each of the 4005 pairs of brain regions.

Specifically, we are interested in exploring the weights assigned by the classification network to the mapping layer M in Equation 7.1. Deep neural networks have been applied to visualize feature importance on images [154] and videos [155]. To explore the importance of features assigned by the classification network, we carried out work similar to [102]. The main idea of the approach is: given a learned neural network and a class of interest, we trace back to the original input by a backward pass with which we can determine how each input entity affects the final detection score for a specific class. In our model, we have two classes (healthy control and ADHD) and we trace back to the mapping layer values to find out how each mapping layer value affects the prediction of a particular class.

Given a particular output value of the mapping layer M_0, a class c and the class score function $S_c(M)$, we would like to rank the elements of M_0 based upon their influence on the score $S_c(M_0)$. Consider the linear score model for the class c:

$$S_c(M) = w_c M + b_c, \tag{7.3}$$

where M is the one-dimensional vector, calculated from Equation 7.1 and is reminiscent of the functional connectivity in our network, w_c is the weight and b_c is the bias of the model. Here, it is clear that the magnitude of the elements of the weight vector w_c specifies the importance of the corresponding element of M for the class c.

In the case of a deep neural network, the class score is a non-linear function of input values, so the above assumption cannot be applied directly. However, given a vector M_0, we can approximate S_c with a linear function in the neighbourhood of M_0 by a first-order Taylor expansion [154]:

$$S_c(M) \approx wM + b, \tag{7.4}$$

where w is the derivative of S_c with respect to the vector M at the point M_0:

$$w = \frac{\partial S_c}{\partial M}|_{M_0}. \tag{7.5}$$

Another justification of the network-learned weight using the class score derivative from Equation 7.5 is that the magnitude of the derivative indicates which elements need to be changed the least to affect the class score the most. One can expect such elements to be more discriminative for a particular class. The derivative w in Equation 7.5 is calculated through back-propagation during the training of the network. We define the feature importance of a node i at layer d as:

$$f_c^d(i) = \sum_{l=L-1}^{d} \sum_k w_c^{(l,l+1)} f_c^{(l+1)}(k), \tag{7.6}$$

where L is the total number of layers in our classification network, k is the number of nodes and f_c^L is the output of the classification network. We define I as the feature importance map for the class c, where each element is given by:

$$I_c(x) = f_c^M(x). \tag{7.7}$$

I_c defines the feature importance of a particular class c.

7.3.4 Results

We evaluate the proposed network with data from three imaging sites (NYU, NI and Peking) from the ADHD-200 dataset [106]. The number of training subjects in each site is $226, 48$ and 85 respectively. ADHD-200 has provided separate train and test datasets for individual imaging site. The proposed end-to-end model is trained on the training dataset of each imaging site and the corresponding test dataset of the individual site is used for testing. Please note, the data used to test the method is completely independent from the data used to train. In order to evaluate the performance of the DeepFMRI, we have evaluated and compared results with state-of-the-art methods as described in the previous section. The comparison results are presented in Table 7.1. The results show that the method proposed in this chapter outperforms the average accuracy results of the competition teams (data from the competition website [106]), the highest accuracy of competition for any individual site (from [6]), correlation-based functional connectivity results and clustering based results. The method also performs well in comparison with the FCNet method [8] explained in Chapter 6. The highest accuracy achieved with our method is for the NYU dataset with a classification accuracy of 73.1%. The classification accuracy for the NI and Peking datasets are 67.9% and 62.7%.

Table 5.1 highlights that the distribution of healthy control and ADHD classes in train and test splits are different. However, in order to achieve better performance by any classifier, training and testing data should follow a similar class-distribution. The performance of any classifier depends on the distribution of the training data. If the majority class is changed for the testing data, the classifier performance would drop significantly. For the calculation of the baseline classifier accuracy, it can be assumed that a simple classifier would predict the majority class of the training dataset for all testing subjects. In the case of Peking, the majority class in the training dataset is healthy control, so the baseline accuracy for Peking on testing dataset is 47.1% $(24/(24+27))$. Similarly for NYU, with ADHD as the majority class in the training data set, baseline accuracy is 70.7% $(29/(29+12))$, and for NI, with ADHD as the majority class in the training dataset is 44.0% $(11/(11+14))$. The baseline accuracy for three imaging sites is presented in Table 7.1, where DeepFMRI performs much better than baseline accuracy for Peking and NYU and slightly better for NYU.

The results show that the DeepFMRI shows the improved results for NI and NYU and the classification accuracy is highest in all three imaging sites. For the Peking, results for both the FCNet [8] and DeepFMRI are the same.

	NI	Peking	NYU
Average accuracy [106]	56.9%	51.0%	35.1%
Highest accuracy [6]	–	58%	56%
Correlation method	52.0%	52.9%	56.1%
Clustering method [7] (Chapter 5)	44%	58.8%	24.3%
FCNet [8] (Chapter 6)	60.0%	**62.7%**	58.5%
DeepFMRI	**67.9%**	**62.7%**	**73.1%**

Table 7.1 Comparison of the DeepFMRI with the average results of competition teams, highest accuracy achieved for individual sites, correlation method, clustering-based results [7] and the FCNet method [8]. The highest accuracy for NI was not quoted by [6].

One interesting point about the ADHD dataset is that the studies employing the dataset were not able to achieve high classification accuracy. The average and highest accuracy achieved by competing studies is presented in Table 7.1 where the accuracy results are comparable to 50% chance. One possible reason for lower accuracy could be the heterogeneous nature of the data and the scan parameters as discussed in chapter 4 making the dataset difficult to train any single machine learning model.

Next, we are interested in studying the performance of the individual networks of the DeepFMRI.

7.4 Discussion

In this section, we discuss the performance comparison of networks of our proposed method and analyse the features learned by the method.

7.4.1 Performance Comparison

Based on the results in Table 7.1, the proposed end-to-end method comprising the feature extractor, functional connectivity and the classification network to classify ADHD presents better performance than state-of-the-art methods. Although it would be helpful to conduct a statistical significance test, unfortunately, we could not conduct such a test due to the very small number of available subjects in the imaging sites. However, from a methodological point of view, we are mainly interested in investigating which subnetwork is causing the performance enhancement. To this end, we additionally performed some experiments by replacing different combinations of the networks.

Comparison Methods

For comparison, we conducted additional experiments, namely, the effect of functional connectivity, the end-to-end model without the classification network, clustering + classification networks and correlation + classification networks, which are detailed below and the results are presented in Figure 7.4.

Effect of Functional Connectivity

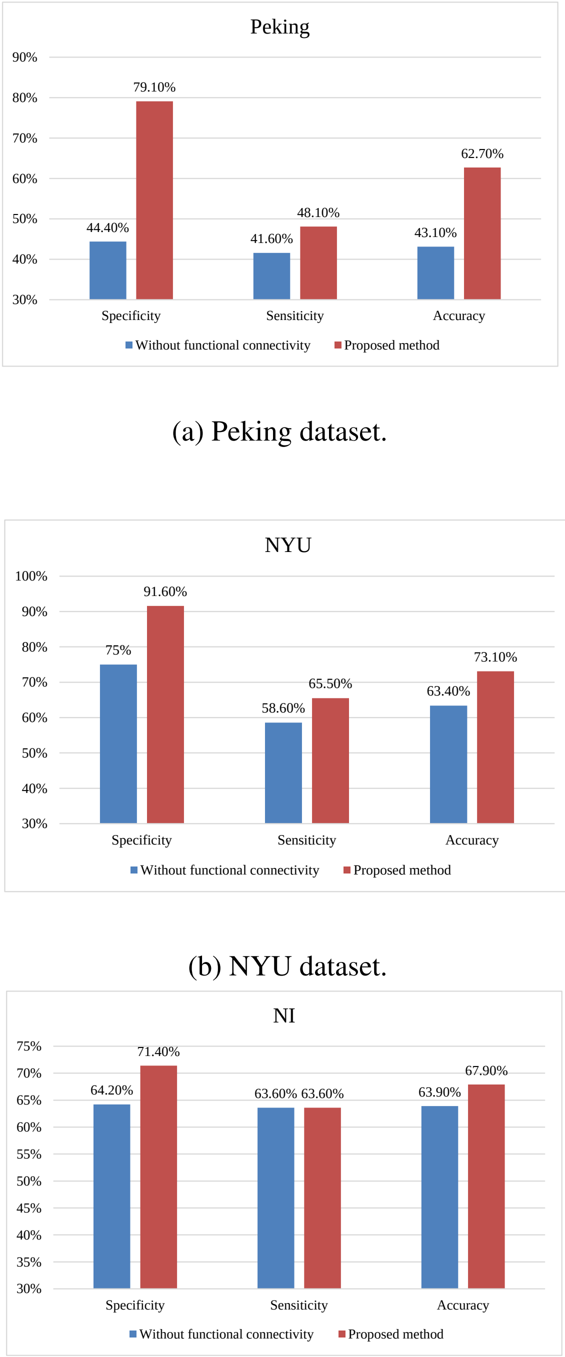
In this experiment, we were interested in exploring whether functional connectivity plays an important role towards the classification of ADHD or not. Towards this end, we evaluated an end-to-end model without the functional connectivity network (as discussed in the previous section and presented in Figure 7.2). Specifically, we were interested in comparing the performance of the end-to-end model with and without functional connectivity and the results are presented in Figure 7.3.

It is important to note that for the end-to-end model without functional connectivity, the number of parameters are less as compared to the end-to-end model with the functional connectivity network. The number of trainable parameters for the end-to-end model with functional connectivity is $502,751$ vs $386,665$ for the end-to-end model without functional connectivity. However, the end-to-end model with functional connectivity yields better performance than the model without functional connectivity. These findings show that functional connectivity serves as an important biomarker towards classification of ADHD.

End-to-end Model without Classification Network

In this experiment, we are interested in determining the importance of the classification network towards diagnosis. Therefore, we use the pre-trained feature extractor and the functional connectivity network to calculate functional connectivity. The proposed classification network is not used in this experiment. The classification network serves two functions in our proposed work: i) it assigns feature importance to each feature and ii) based on those feature importance values, it performs the classification. Therefore, to compensate this network, we need to incorporate feature selection and classification separately. Towards this, an Elastic Net was applied to extract discriminant features from functional connectivity and finally, an SVM classifier was applied to evaluate the classification accuracy as discussed in Chapter 6.

Next, we want to evaluate the importance of the functional connectivity network in our network. To explore this, we designed the network without the feature extractor and the functional connectivity network. Towards this, we carried out two experiments where we replaced the feature extractor and the functional connectivity network with i) clustering and ii) correlation. Both correlation and clustering are used for functional connectivity calculation in these methods. The methods are described below.



(a) Peking dataset.

(b) NYU dataset.

(c) NI datatset.

Fig. 7.3 Comparison of the performance of i) DeepFMRI and ii) the model without the functional connectivity network for three imaging sites. The proposed model shows better performance as compared to the model without functional connectivity.

Clustering + Classification Network

In this experiment, we apply clustering to calculate functional connectivity between the brain regions as proposed by [7, 40] and detailed in Chapter 5. As discussed earlier, the feature extractor and the functional connectivity networks are not used here. The functional connectivity through clustering is passed into the proposed classification network to evaluate the performance of the network.

Correlation + Classification Network

Correlation is a popular method to calculate functional connectivity between brain regions. We employ correlation to calculate functional connectivity. Similar to the previous experiment, a classification network was employed on the calculated functional connectivity.

Comparison Results

We performed the comparison of these four methods and the results are presented in Figure 7.4. From the results, it is apparent that the DeepFMRI outperforms all other evaluated methods or combinations. Comparison of 'clustering + classification network' and 'correlation + classification network' supports the findings of [7] that clustering is a better method to calculate functional connectivity as compared to correlation-based techniques. However, our proposed end-to-end model yields better performance.

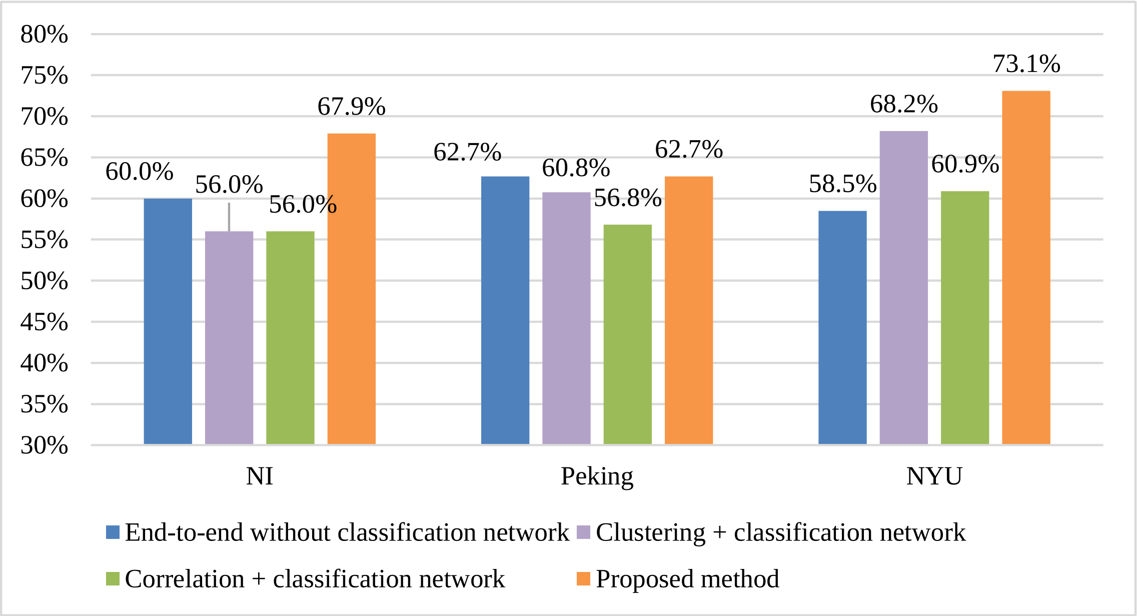


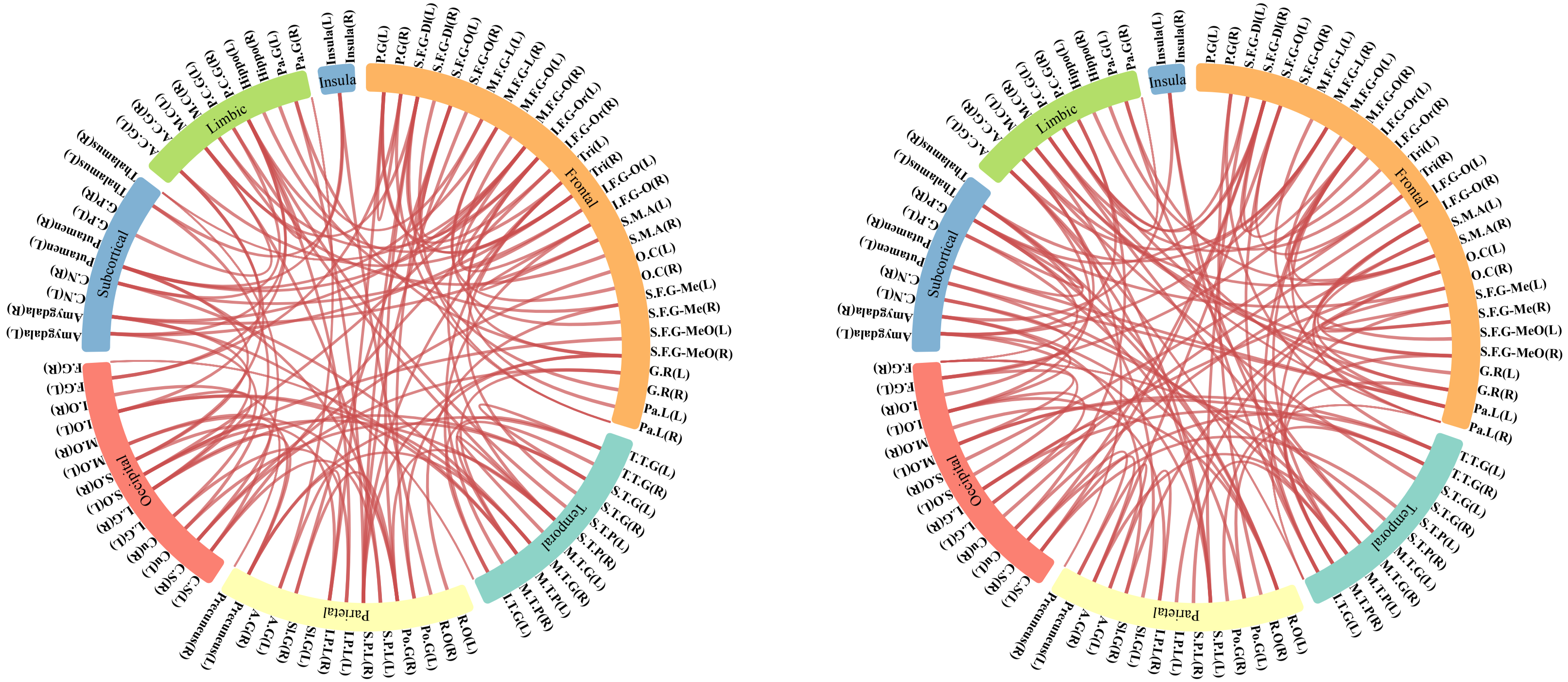
Fig. 7.4 Comparisons of classification accuracy of different methods. The results suggest that the DeepFMRI method outperforms all other evaluated methods. The DeepFMRI is able to achieve the highest accuracy on all three imaging sites where it outperforms in NI and NYU imaging dataset.

7.4.2 Analysis of Learned Feature Importance of Functional Connectivity

The feature importance map (I_c) from Equation 7.7 is a 4005 dimensional vector where each value corresponds to the importance of the respective functional connectivity value towards prediction of a particular class. We were interested in exploring the learnt feature importance values. Towards this goal, we have selected feature importance values from the NYU dataset as, i) NYU has the largest number of subjects compared to other imaging sites, and ii) NYU has highest classification accuracy.

We have plotted the feature importance map for both the healthy and ADHD classes in Figure 7.5. For the sake of clarity, we have plotted the top 100 feature maps for both classes. The figure highlights some of the differences in feature importance learned by our method for both classes. Our method assigns different weights to an individual feature with respect to its importance towards prediction of a subject. This is in contrast to the most classical machine learning methods [7, 40], which typically employ a feature selection algorithm that assigns a single weight to a functional connectivity regardless of the class.

Next, we have plotted the feature importance values on the brain map. The visualization of the healthy and the ADHD classes are shown in Figure 7.6 and Figure 7.7, respectively. The figures show that in most of the cases, the importance value assigned by our network to a particular functional connectivity is different for both classes.



(a) $I_{healthy}$.

(b) I_{ADHD}.

Fig. 7.5 Visualization of the learned feature importance map for a) healthy and, b) ADHD classes for the NYU dataset. For the sake of clarity, only top 100 values for an individual class are visualized. The visualization shows the differences in the feature maps of both classes.

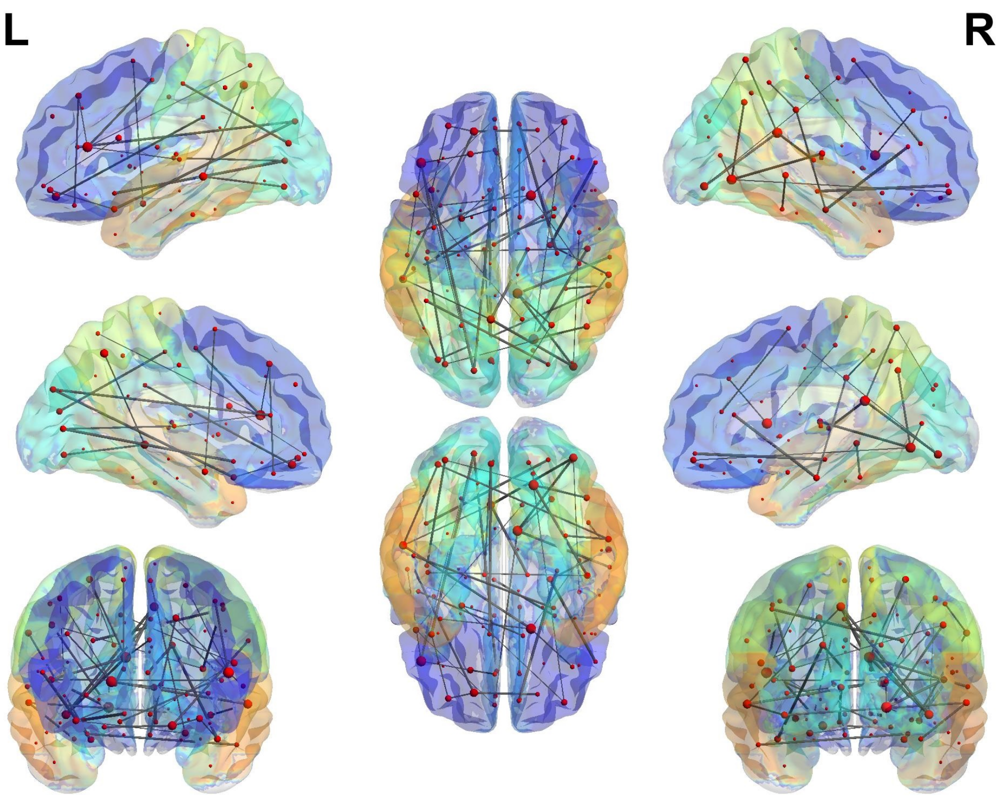


Fig. 7.6 Visualization of the learned feature importance map for the healthy class on the brain volume. For the sake of clarity, only top 50 values are visualized. The size of a node relates to the number of edges (functional connections) connected to the particular brain region and the width of an edge highlights the strength of a particular feature's importance. (Data visualized through the BrainNet viewer software [4]).

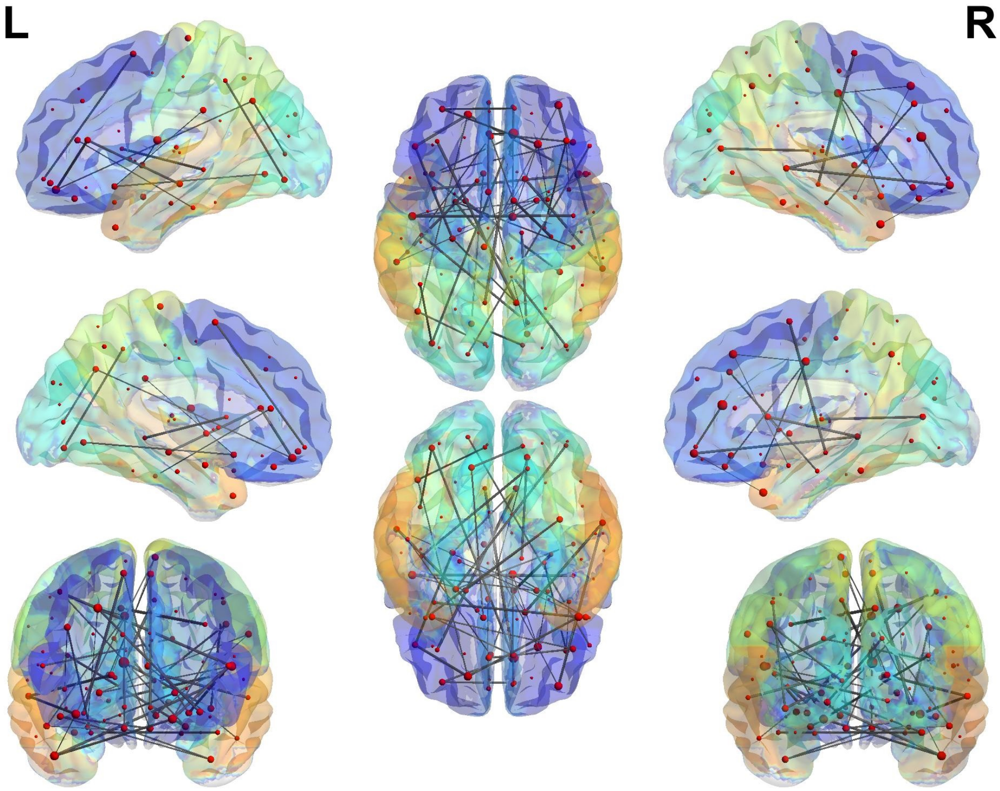


Fig. 7.7 Visualization of the learned feature importance map for the ADHD class on the brain volume. For the sake of clarity, only top 50 values are visualized. The size of a node relates to the number of edges (functional connections) connected to the particular brain region and the width of an edge highlights the strength of a particular feature's importance. Visualization through the BrainNet viewer software [4].

We performed an experiment for the quantitative analysis of feature maps of both classes. Our motivation was to compare the top 100 feature maps of both classes. The top 100 feature maps values were extracted from the healthy class and a lookup was performed in the ADHD feature maps. The result is presented in Figure 7.8. The figure shows that out of the top 100 feature maps of the healthy class, less than 10% fall in the top 500 feature maps in the ADHD class. Similarly, we extracted top 100 feature maps from the ADHD class and computed the lookup in the healthy class and the results are presented in Figure 7.9. As in the previous inference, out of the top 100 feature maps of the ADHD class, less than 10% fall in the top 500 feature maps in the healthy class. Our findings suggest that the altered functional connectivity between healthy control and ADHD may relate to functional brain network differences. In particular, the DeepFMRI appears to weight different brain network structures depending on the particular class (control or ADHD).

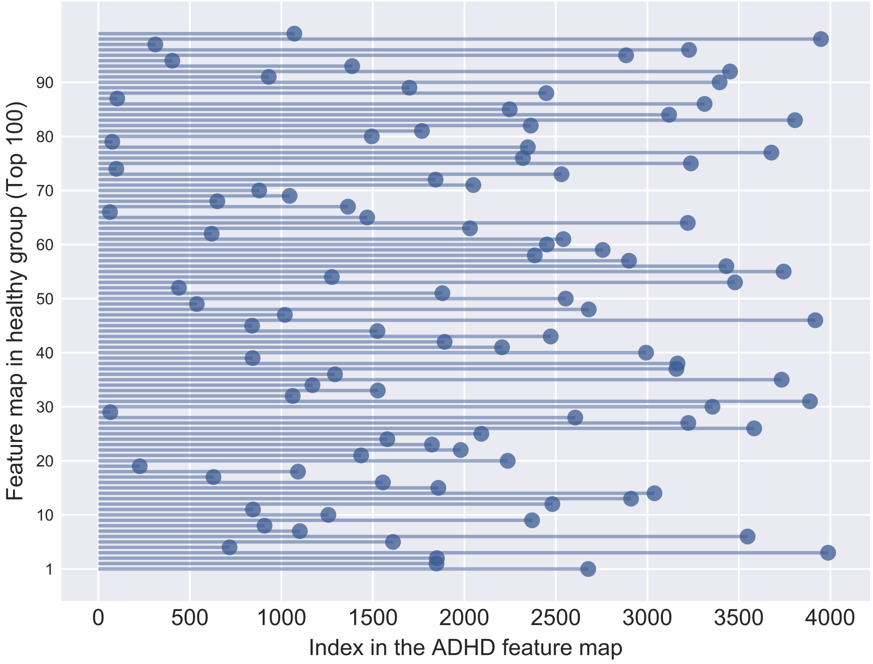


Fig. 7.8 Plot of matching the top 100 healthy feature maps in the ADHD feature maps. The y-axis represents the top 100 feature maps in the healthy group and the x-axis represents the index of a particular healthy feature map in the ADHD feature map. The figure shows that out of top 100 feature maps of the healthy class, less than 10% fall in the top 500 feature maps in the ADHD class.

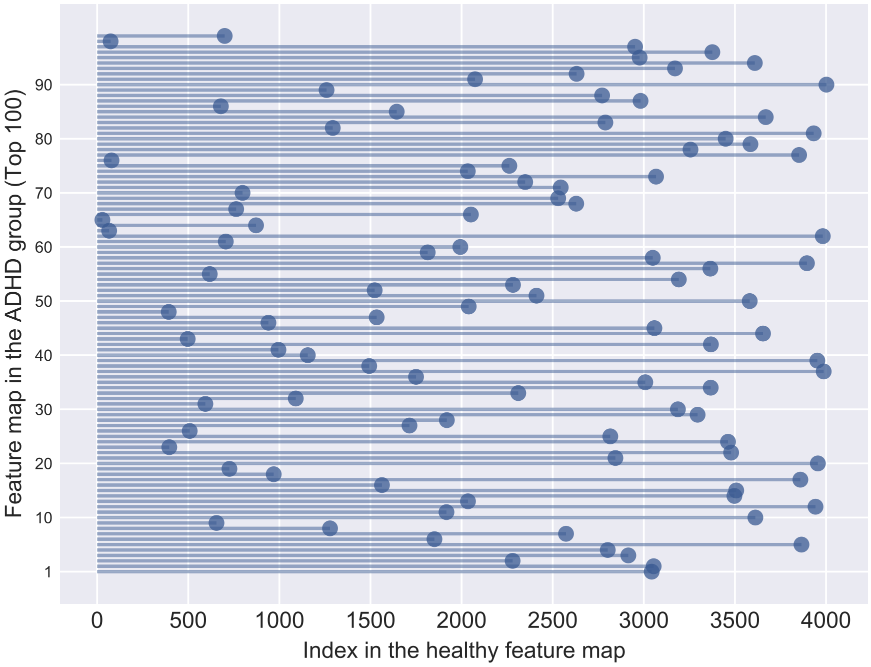


Fig. 7.9 Plot of matching the top 100 ADHD feature maps in the healthy feature maps. The y-axis represents the top 100 feature maps in the ADHD group and the x-axis represents the index of a particular ADHD feature map in the healthy feature map. The figure shows that out of the top 100 feature maps of the ADHD class, less than 10% fall in the top 500 feature maps in the healthy class.

Finally, we are interested in analyzing the learned feature importance map for both classes with respect to the inter-lobe and intra-lobe distribution. We have categorized the learned feature importance map with respect to their respective lobes and the results are visualized in Figure 7.10. The results suggest that for both classes, the frontal lobe carries a higher number of discriminant features in terms of both inter and intra-lobe features. The figure shows a different distribution for all the lobes in both classes. The distribution is highlighted by the different shape of an individual lobe when comparing the two classes. The frontal lobe is known to be involved in cognitive processes [134], including attention, planning, sequential organization and self-monitoring of actions, affect and mood, memory, self-awareness and personality [134]. The alterations in the frontal lobe may cause abnormal behaviours in these functions. Volumetric alterations in ADHD have been reported [156]. Studies have also shown connectivity alterations in frontal, temporal, and occipital cortices locally as well as with the rest of the brain in individuals with ADHD [157]. Our findings about the frontal lobe alterations in ADHD support the results found in earlier studies [158–160].

7.5 Conclusions

In this chapter, we have proposed an innovative end-to-end deep neural network for classification of ADHD from fMRI data. The DeepFMRI takes pre-processed time-series signals of fMRI as input and learns to predict the classification label. We were interested in studying whether the classification task in fMRI can be solved by an end-to-end network. As far as we know, this is the first attempt to apply an end-to-end network incorporating functional connectivity for classification of a neurological disorder.

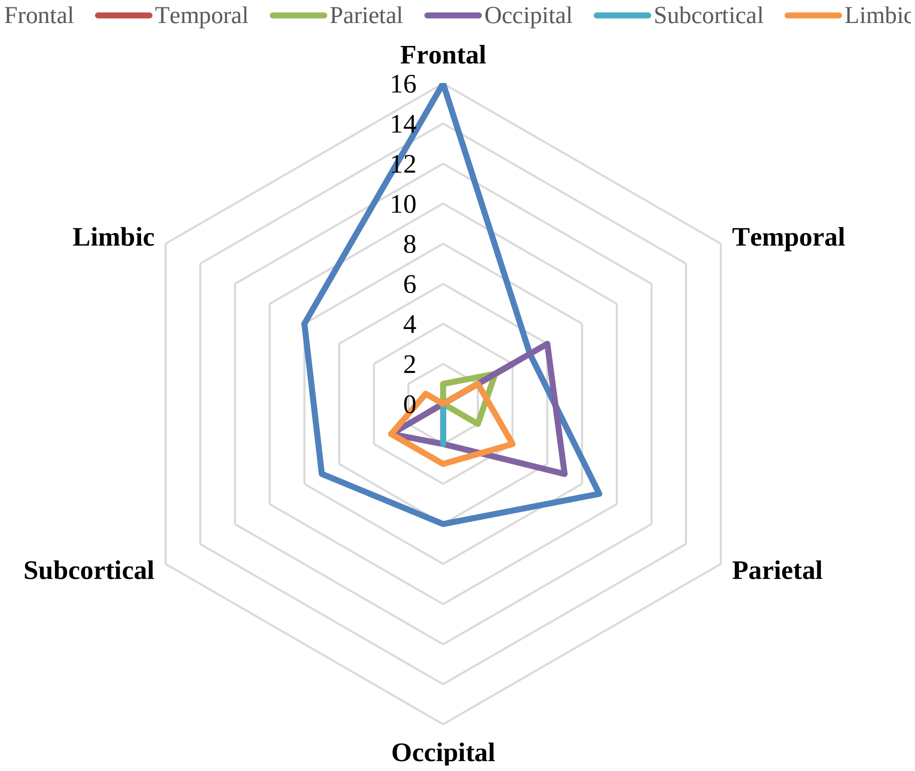
We have evaluated the importance of functional connectivity in the proposed end-to-end network. Findings show that despite the large number of parameters in our end-to-end network, it performs better as compared to an end-to-end network without functional connectivity with comparatively less number of trainable parameters. This result strengthens the argument that functional connectivity is an important biomarker towards the identification of a neurological disorder. Experimental results on the ADHD-200 dataset demonstrate that utilizing such model outperforms the current state-of-the-art.

Our proposed model is able to associate different weights to an individual functional connectivity with respect to its importance in predicting a class label (healthy control and ADHD), unlike most of the feature selection strategies in classical machine learning. The proposed method in this chapter appears to assign weight to different brain networks with respect to a particular class (healthy control or ADHD).

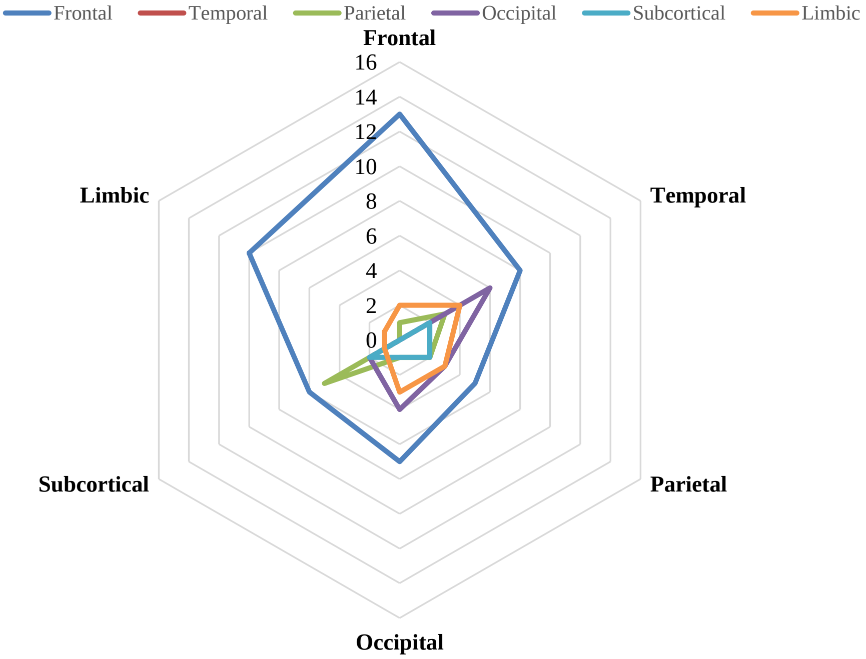
Our results suggest that the frontal lobe carries most discriminant power in classifying ADHD. The frontal lobe is known to be associated with cognitive functions like attention, memory, planning and mood. Our findings about the frontal lobe anomalies in ADHD support earlier studies.

One interesting extension of the this work could be to incorporate non-imaging features in the DeepFMRI. The model would need to be modified by adding a sub-network that will take the non-imaging features as input and incorporate them in the existing network. Such an extension is also discussed in section 8.3.

In the next chapter, we provide the conclusions, limitations and future dimensions of the dissertation.



(a) Healthy class.



(b) ADHD class.

Fig. 7.10 Distribution of the top 100 features maps in the healthy and the ADHD classes.

Chapter 8

Conclusions

In this dissertation, we addressed the problem of classification of ADHD subjects using their brain resting state functional MRI (fMRI) data. The problem is particularly of importance due to the widespread impact of ADHD on the global child population and the lack of biological measures to diagnose it. Approximately 5-10% of the children all over the world are diagnosed with ADHD. This motivated us to propose a solution for the automatic ADHD diagnosis problem. The central idea of our approach is to exploit brain functional connectivity differences between healthy and ADHD subjects. We explored different methods to calculate functional connectivity and evaluated the differences between healthy and ADHD functional connectivity patterns.

Our first approach for solving this problem used the integration of non-imaging features with imaging features. We proposed an affinity propagation clustering-based method of calculating functional connectivity of brain regions. We have employed a novel way to initialise its parameter (preference value). An Elastic Net based feature selection method was explored to extract discriminant functional connectivity. The imaging features (functional connectivity) were integrated with non-imaging features (such as age, gender) and the final feature set was presented to a Support Vector Machine (SVM) classifier for the final prediction. The method outperformed previous state-of-the-art results. However, this method was based on classical machine learning methods such as feature selection and classification.

In the second approach, as discussed in Chapter 6, we explored a Convolutional Neural Network (CNN) model, namely FCNet, for calculating functional connectivity of brain regions. In this work, a Siamese inspired network, namely feature extractor network, has been proposed in which a CNN is used to extract features from time-series signals. The feature extractor network maps the time-series signlas to abstract features. The second part of the FCNet is a similarity network. The abstracted features from the feature extractor network are passed to the similarity measure network which calculates the strength of similarity between the regions. This similarity measure serves as functional connectivity in this work. An Elastic Net is applied to extract discriminant features and finally, an SVM classifier is evaluated for the final prediction. The method was able to yield better classification accuracy as compared to other methods. However, the method relied on the classical machine learning methods for feature selection and classification.

Next, we proposed an end-to-end trainable deep network for classification of the ADHD, which incorporates functional connectivity in the proposed architecture. The proposed model, namely DeepFMRI, takes preprocessed fMRI time-series signals as input and produces the prediction label as output. The method is composed of three networks. First one is the feature extractor network that maps the preprocessed time-series signals to the abstracted features. The second network is a functional connectivity network which is comprised of multiple similarity measure networks and claculates functional connectivity of brain regions from their abstracted features. The last part is a classification network that takes the functional connectivity as input and yields the classification label as the final output of the network. The innovative network outperformed previous state-of-the-art methods.

Lastly, we evaluated the importance of functional connectivity in the proposed deep learning model. We observed that excluding the functional connectivity network from the deep network reduces the performance of the network. This finding shows that the functional connectivity plays a key role for classification of the ADHD. Our results suggest that the frontal lobe carries most discriminant power in classifying ADHD. The frontal lobe is known to be associated with cognitive functions like attention, memory, planning and mood. Our findings about the frontal lobe anomalies in ADHD support the earlier studies.

8.1 Outcomes

Our motivation of this work was to explore novel machine learning models to help medical experts for diagnosis of a brain disorder. We approached the solution by dividing the goal in multiple objectives in Section 1.2. Each of it can be addressed below.

- Functional connectivity as an important biomarker for diagnosis of a brain disorder: This objective has been addressed in Chapter 7. We proposed a deep learning network incorporating functional connectivity in the model. We have evaluated that functional connectivity serves as an important biomarker for predicting a brain disorder and improves the classification performance of the model

- Novel machine learning methods for evaluation of functional connectivity: This objective has been addressed in Chapters 5 and 6. In Chapter 5, we have proposed a clustering based novel method to calculate functional connectivity. The method achieved improved classification accuracy as compared to the existing work. In Chapter 6, a novel Convolutional Neural Network based model has been presented to calculate functional connectivity. This method achieved better classification result.

- Importance of non-imaging features for prediction of a brain disorder: Chapters 5 and 6 highlight the objective. The chapters show that the non-imaging features improve the classification accuracy of the proposed methods.

- Deep neural network for classification of a brain disorder: The objective has been highlighted in Chapter 7. We have presented an end-to-end deep network that takes fMRI preprocessed time-series signals as input and produces the prediction of label (healthy or ADHD) as output.

- Convoluional Neural Networks to map time-series functional MRI signals to features: Chapters 6 and 7 covers this objective. We have prosed a Convolutional Neural Network to map time-series signals to abstract features that are used for calculation of functional connectivity. Functional connectivity is used in next steps for the final prediction. The results show that the novel method of calculating functional connectivity improves the results.

In the next section, we present some limitations of the proposed work.

8.2 Limitations

One of the limitations of the proposed work is the small data size being evaluated. We have evaluated ADHD-200 dataset and different imaging sites contributed to the dataset. There is a small number of subjects in individual imaging sites. Also, the data is very heterogeneous across different sites requiring training the network separately for each institution. One main reason of data being heterogeneous is different protocols are followed by different imaging sites. The clinical protocols are constantly changing and there is a lack of agreement for a common protocol. There is a need to decide on a single protocol and a consistent way to acquire imaging data that can be followed by all data contributors. The heterogeneous nature of the data makes it difficult to train a single machine learning model that can yield better performance. Due to these facts, the classification accuracy achieved by the proposed work and other studies has room for improvement. One possible way to train a network on the heterogeneous data could be to explore multi-task learning as proposed by [161] for the ADHD-200 dataset. In this possible work, feature extractor networks could be trained for individual imaging site, while the rest of the network can be trained on all imaging sites. Using this network, the feature extractor network can be specific to individual imaging site while the functional connectivity and classification network can be learned from all datasets.

We have used CNNs in Chapter 6 and Chapter 7 as the feature extractor network. CNNs require a fixed length of input data, therefore, our methods are designed to accept a fixed length of time-series signals. In order to evaluate any other dataset with different time-series length, the proposed models would need to be redesigned. One possible extension could be to design a separate feature extractor network for the individual dataset. The rest of the network would be the same as proposed in Chapter 7. Another option could be to map the original time-series to a standardised length. An autoencoder could be one of the best approaches to map the time-series length to a fixed length. The mapped dimensions could be used as input to the proposed method.

Another limitation is the lower accuracy of the proposed methods and existing work achieved on this dataset as compared to the accuracy achieved in other domains such as image classification. One of the main reason is the dataset itself, which is very challenging, as discussed earlier. In order to achieve reasonable classification results, a large amount of data with a common protocol needs to be acquired. Where the proposed methods achieved better classification results as compared to the state-of-the-art, there is room for improvement in results with more data.

In our approach, we have calculated functional connectivity of brain regions based on the assumption that functional connectivity of any two brain regions is consistent over the entire duration of scan. Such paradigm is called static functional connectivity. One interesting experiment could be based on dynamic functional connectivity. Dynamic functional connectivity is based on the assumption that the functional connectivity over the time of scan is not constant and it might show a different trend. Analysis based on dynamic functional connectivity might provide some better understanding of brain disorders.

8.3 Future Work

Brain imaging based methods show promise for solving the proposed problem as different independent studies reported ADHD detection accuracy higher than a chance factor. However, there are many areas to improve because none of the methods are good enough to replace the current manual diagnosis process. Further investigation needs to be performed regarding the data capturing protocols and the community needs to decide a standard method. As different protocols may lead to the variations of cognitive activities of brain which can reduce the performance of the diagnosis method. Some possible future extensions of this work are discussed below.

ADHD Sub-groups

In future work, we are interested in applying the proposed network to study ADHD sub-groups. The study may be based on treatment response, clinical scores, disorder outcomes etc. However, it will not require redesigning the proposed network. If the output variable is discrete, a classifier can be used as presented in this work. If the output variable is continuous, instead a regressor can be used. We are also interested in applying the proposed model to other disorders like epilepsy and Alzheimer's.

Synthetic Data Generation

Scarcity of data is a critical problem in a number of computer vision domains, particularly in the domain of medical imaging. Dataset imbalance is another related issue for any Artificial Intelligence based solution. We have addressed dataset imbalance in this dissertation through Synthetic Minority Oversampling Technique (SMOTE) which generates minority subjects from the available data. Recently, Goodfellow et al [162] proposed Generative Adversarial Networks (GANs) as a state-of-the-art generative network. A GAN is composed of two neural networks, a generator model and a discriminator model. The generator model maps random data to a data distribution of interest while the discriminator model estimates the probability that a sample came from the real training data rather than the generator model. The ultimate goal of the GAN is to generate data that is close to real data. GANs have been applied to a number of computer vision applications like natural images synthesis [163], style generation [164], conditional GANs [165], and many more. It would be interesting to design a GAN for fMRI data generation. With a reasonable amount of training data, the goal could be to design a GAN to generate synthetic subjects from the real training data. The synthesized data can be used to train any machine learning model.

Autoencoder-based Feature Extractor

In this dissertation, we have proposed an end-to-end deep network that includes a CNN feature extractor network. The feature extractor network used a set of weights pre-trained through FCNet. One possible extension could be to explore autoencoders for feature extractor network. The aim will be to train an autoencoder that can project the time-series data of brain regions to a lower dimension. The mapped lower-dimensional data can be used as input to the similarity network. With comparatively less number of parameters, this model might be expected to perform better.

Long Short Term Memory (LSTM)

In this dissertation, we have used CNN based feature extractor network. CNNs are well known for their representation learning capability, they learn the feature during the training. An interesting experiment could be to explore Long Short Term Memory (LSTM) for the feature extractor network. LSTMs have been shown to perform well in a number of time-series applications. An LSTM can be applied to extract features from the time-series fMRI signals. These extracted features can be used as input to the similarity measure network.

Fusion of Multiple Imaging Modalities

In this dissertation, we have explored fMRI imaging modality for classification of ADHD. One can speculate that multiple imaging modalities might add more discriminative capabilities to a machine learning model. For instance, along with functional data, structural data might also contain important information that can improve the performance of the model. One of the interesting future directions could be to integrate MRI with fMRI in a deep network and explore if it could yield better diagnosis results. One possible network design is presented in Figure 8.1. There are two network modules, one for each fMRI and MRI imaging modality. The fMRI network module is the same as the proposed in this work. For the MRI network module, a CNN can be used to extract the features. The extracted features from both the fMRI and MRI can be combined to yield final prediction results.

Integration of Non-Imaging Features in the End-to-end Network

In our end-to-end network, we have employed only imaging features for the prediction of ADHD. One interesting avenue could be to integrate non-imaging features to the network

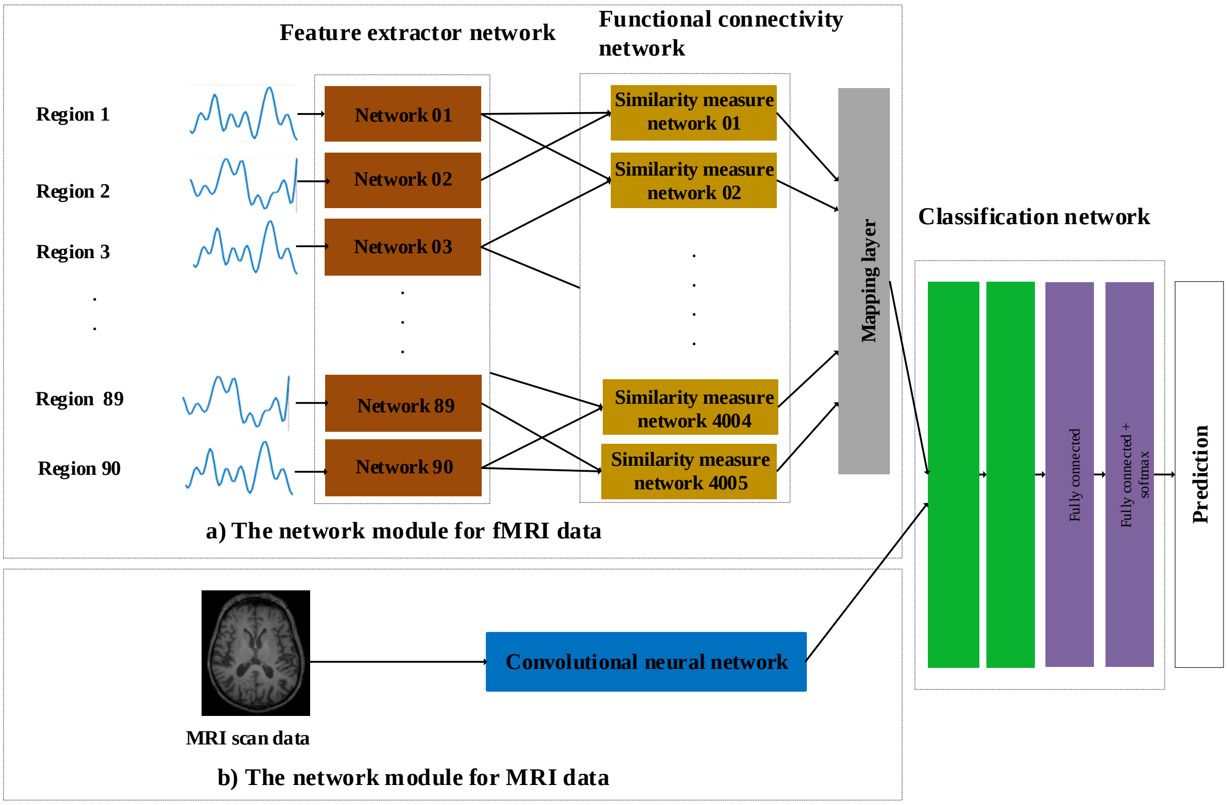

Feature extractor network
Functional connectivity network
Region 1
Region 2
Region 3
Region 89
Region 90
Network 01
Network 02
Network 03
Network 89
Network 90
Similarity measure network 01
Similarity measure network 02
Similarity measure network 4004
Similarity measure network 4005
Mapping layer
Classification network
Fully connected
Fully connected + softmax
Prediction
a) The network module for fMRI data
MRI scan data
Convolutional neural network
b) The network module for MRI data

Fig. 8.1 A possible extension for fusion of multiple imaging modalities. The network (a) is for fMRI data, which is the same as proposed in this dissertation and the network (b) is for MRI data.

and explore if it can improve the classification results or not. One possible implementation could be to add a layer prior to the classification network that will combine the imaging features and non-imaging features and present the combined features to the classification network. In this work, non-imaging features of age and IQ levels were not at the same scale as the imaging features, and this might affect the performance of the classifier. In future work we intend to rescale the non-imaging features to match the scale of the imaging features and evaluate performance.

Different Brain Disorders

In this work, we have explored ADHD. We are interested in studying other brain disorders with our proposed methods. It could be an interesting future direction to extend the proposed models to other brain disorders like schizophrenia, epilepsy, dementia etc. Our proposed model would be easily scalable to other disorders also with very minimum changes.

References

[1] Atif Riaz, Kashif Rajpoot, and Nasir Rajpoot. A connectivity difference measure for identification of functional neuroimaging markers for epilepsy. In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pages 1517–1520. IEEE, 2013.

[2] Raymond Salvador, John Suckling, Martin R Coleman, John D Pickard, David Menon, and ED Bullmore. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral cortex*, 15(9):1332–1342, 2005.

[3] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3286, 2015.

[4] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[5] Soumyabrata Dey, A Ravishankar Rao, and Mubarak Shah. Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects. *Frontiers in Neural Circuits*, 8, 2014.

[6] Marta Nuñez-Garcia, Sonja Simpraga, Maria Angeles Jurado, Maite Garolera, Roser Pueyo, and Laura Igual. Fadr: Functional-anatomical discriminative regions for rest fmri characterization. In *International Workshop on Machine Learning in Medical Imaging*, pages 61–68. Springer, 2015.

[7] Atif Riaz, Eduardo Alonso, and Greg Slabaugh. Phenotypic integrated framework for classification of adhd using fmri. In *International Conference Image Analysis and Recognition*, pages 217–225. Springer, 2016.

[8] Atif Riaz, Muhammad Asad, SM Masudur Rahman Al-Arif, Eduardo Alonso, Danai Dima, Philip Corr, and Greg Slabaugh. FCNet: A Convolutional Neural Network for Calculating Functional Connectivity from functional MRI. In *International Workshop on Connectomics in Neuroimaging*, pages 70–78. Springer, 2017.

[9] Ronald C Kessler, Sergio Aguilar-Gaxiola, Jordi Alonso, Somnath Chatterji, Sing Lee, and T Bedirhan Üstün. The WHO world mental health (WMH) surveys. *Psychiatrie (Stuttgart, Germany)*, 6(1):5, 2009.

[10] Colin Mathers, Doris Ma Fat, and Jan Ties Boerma. The global burden of disease: 2004 update. 2008.

[11] Pamela Y Collins, Vikram Patel, Sarah S Joestl, Dana March, Thomas R Insel, Abdallah S Daar, Isabel A Bordin, E Jane Costello, Maureen Durkin, Christopher Fairburn, et al. Grand challenges in global mental health. *Nature*, 475(7354):27–30, 2011.

[12] Naomi A Fineberg, Peter M Haddad, Lewis Carpenter, Brenda Gannon, Rachel Sharpe, Allan H Young, Eileen Joyce, James Rowe, David Wellsted, David Nutt, et al. The size, burden and cost of disorders of the brain in the uk. *Journal of Psychopharmacology*, pages 761–770, 2013.

[13] Valerie A Harpin. The effect of adhd on the life of an individual, their family, and community from preschool to adult life. *Archives of disease in childhood*, 90(suppl 1):i2–i7, 2005.

[14] Joel T Nigg. Attention-deficit/hyperactivity disorder and adverse health outcomes. *Clinical psychology review*, 33(2):215–228, 2013.

[15] Ariadna Albajara Sáenz, Thomas Villemonteix, and Isabelle Massat. Structural and functional neuroimaging in attention-deficit/hyperactivity disorder. *Developmental Medicine & Child Neurology*, 61(4):399–405, 2019.

[16] Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541, 2009.

[17] Peter A Bandettini. Twenty years of functional mri: the science and the stories. *Neuroimage*, 62(2):575–588, 2012.

[18] Alistair M Howseman and Richard W Bowtel. Functional magnetic resonance imaging: imaging techniques and contrast mechanisms. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 354(1387):1179–1194, 1999.

[19] Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[20] Seong-Gi Kim and Seiji Ogawa. Biophysical and physiological origins of blood oxygenation level-dependent fmri signals. *Journal of Cerebral Blood Flow & Metabolism*, 32(7):1188–1206, 2012.

[21] Edson Amaro Jr and Gareth J Barker. Study design in fmri: basic principles. *Brain and cognition*, 60(3):220–232, 2006.

[22] Nikos K Logothetis and Josef Pfeuffer. On the nature of the bold fmri contrast mechanism. *Magnetic resonance imaging*, 22(10):1517–1531, 2004.

[23] Marcus E Raichle and Mark A Mintun. Brain work and brain imaging. *Annu. Rev. Neurosci.*, 29:449–476, 2006.

[24] Douglas N Greve, Gregory G Brown, Bryon A Mueller, Gary Glover, Thomas T Liu, et al. A survey of the sources of noise in fmri. *Psychometrika*, 78(3):396–416, 2013.

[25] Kevin Murphy, Rasmus M Birn, and Peter A Bandettini. Resting-state fmri confounds and cleanup. *Neuroimage*, 80:349–359, 2013.

[26] Thomas T Liu. Noise contributions to the fmri signal: An overview. *NeuroImage*, 143:141–151, 2016.

[27] Douglas N Greve, Bryon A Mueller, Thomas Liu, Jessica A Turner, James Voyvodic, Elizabeth Yetter, Michele Diaz, Gregory McCarthy, Stuart Wallace, Brian J Roach, et al. A novel method for quantifying scanner instability in fmri. *Magnetic resonance in medicine*, 65(4):1053–1061, 2011.

[28] Arno Klein, Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, D Louis Collins, James Gee, Pierre Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, 46(3):786–802, 2009.

[29] Alexandre Guimond, Jean Meunier, and Jean-Philippe Thirion. Average brain models: A convergence study. *Computer vision and image understanding*, 77(2):192–210, 2000.

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

[31] Jingyuan E Chen and Gary H Glover. Functional magnetic resonance imaging methods. *Neuropsychology review*, 25(3):289–313, 2015.

[32] Stephen LaConte, Jon Anderson, Suraj Muley, James Ashe, Sally Frutiger, Kelly Rehm, Lars Kai Hansen, Essa Yacoub, Xiaoping Hu, David Rottenberg, et al. The evaluation of preprocessing choices in single-subject bold fmri using npairs performance metrics. *NeuroImage*, 18(1):10–27, 2003.

[33] Statictical Parameter Mapping. www.fil.ion.ucl.ac.uk/spm.

[34] FMRIB Software Library (FSL). www.fmrib.ox.ac.uk/fsl.

[35] Yan Chao-Gan and Zang Yu-Feng. Dparsf: a matlab toolbox for “pipeline” data analysis of resting-state fmri. *Frontiers in Systems Neuroscience*, 4, 2010.

[36] Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[37] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[38] Bharat B Biswal, Joel Van Kylen, and James S Hyde. Simultaneous assessment of flow and bold signals in resting-state functional connectivity maps. *NMR in Biomedicine*, 10(45):165–170, 1997.

[39] Kashif Rajpoot, Atif Riaz, Waqas Majeed, and Nasir Rajpoot. Functional connectivity alterations in epilepsy from resting-state functional mri. *PloS one*, 10(8), 2015.

[40] Atif Riaz, Muhammad Asad, Eduardo Alonso, and Greg Slabaugh. Fusion of fmri and non-imaging data for adhd classification. *Computerized Medical Imaging and Graphics*, 65:115–128, 2018.

[41] Andrew Ng. Cs229 lecture notes. *CS229 Lecture notes*, 1(1):1–3, 2000.

[42] Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. CRC Press, 2007.

[43] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[44] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[45] Jianping Hua, Waibhav D Tembe, and Edward R Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424, 2009.

[46] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: algorithms and applications*, page 37, 2014.

[47] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[48] Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media, 2012.

[49] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[50] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[51] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[54] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. Joint learning of words and meaning representations for open-text semantic parsing. In *Artificial Intelligence and Statistics*, pages 127–135, 2012.

[55] Dan C Cireşan, Ueli Meier, and Jürgen Schmidhuber. Transfer learning for latin and chinese characters with deep neural networks. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–6. IEEE, 2012.

[56] Jimmy SJ Ren and Li Xu. On vectorization of deep convolutional neural networks for vision tasks. In *AAAI*, pages 1840–1846, 2015.

[57] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[58] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[59] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[60] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[65] Dai Dai, Jieqiong Wang, Jing Hua, and Huiguang He. Classification of adhd children through multimodal magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6:63, 2012.

[66] Alejandro Tabas, Emili Balaguer-Ballester, and Laura Igual. Spatial discriminant ica for rs-fmri characterisation. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

[67] Yuhui Du, Zening Fu, and Vince D Calhoun. Classification and prediction of brain disorders using functional connectivity: Promising but challenging. *Frontiers in Neuroscience*, 12:525, 2018.

[68] Kaiming Li, Lei Guo, Jingxin Nie, Gang Li, and Tianming Liu. Review of methods for functional brain connectivity detection using fmri. *Computerized Medical Imaging and Graphics*, 33(2):131–139, 2009.

[69] Ali-Mohammad Golestani and Bradley G Goodyear. Regions of interest for resting-state fmri analysis determined by inter-voxel cross-correlation. *Neuroimage*, 56(1):246–251, 2011.

[70] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, 2010.

[71] Debra A Gusnard and Marcus E Raichle. Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience*, 2(10):685, 2001.

[72] Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005.

[73] Dietmar Cordes, Victor M Haughton, Konstantinos Arfanakis, Gary J Wendt, Patrick A Turski, Chad H Moritz, Michelle A Quigley, and M Elizabeth Meyerand. Mapping functionally related regions of brain with functional connectivity mr imaging. *American Journal of Neuroradiology*, 21(9):1636–1644, 2000.

[74] Keri S Taylor, David A Seminowicz, and Karen D Davis. Two systems of resting state connectivity between the insula and cingulate cortex. *Human brain mapping*, 30(9):2731–2745, 2009.

[75] Adriana Di Martino, Anouk Scheres, Daniel S Margulies, AMC Kelly, Lucina Q Uddin, Zarrar Shehzad, B Biswal, Judith R Walters, F Xavier Castellanos, and Michael P Milham. Functional connectivity of human striatum: a resting state fmri study. *Cerebral cortex*, 18(12):2735–2747, 2008.

[76] Jessica R Andrews-Hanna, Abraham Z Snyder, Justin L Vincent, Cindy Lustig, Denise Head, Marcus E Raichle, and Randy L Buckner. Disruption of large-scale brain systems in advanced aging. *Neuron*, 56(5):924–935, 2007.

[77] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012.

[78] Jason W Bohland, Sara Saperstein, Francisco Pereira, Jérémy Rapin, and Leo Grady. Network, anatomical, and non-imaging measures for the prediction of adhd diagnosis in individual subjects. *Frontiers in Systems Neuroscience*, 6:78, 2012.

[79] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[80] Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, Eric Reiman, Alzheimer's Disease NeuroImaging Initiative, et al. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[81] Rajan S Patel, F DuBois Bowman, and James K Rilling. A bayesian approach to determining connectivity of the human brain. *Human brain mapping*, 27(3):267–276, 2006.

[82] A. Eloyan et al. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6:61, 2012.

[83] Wei Cheng, Xiaoxi Ji, Jie Zhang, and Jianfeng Feng. Individual classification of adhd patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. *Frontiers in Systems Neuroscience*, 6:58, 2012.

[84] Jie Zhang, Wei Cheng, ZhengGe Wang, ZhiQiang Zhang, WenLian Lu, GuangMing Lu, and Jianfeng Feng. Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy. *PloS one*, 7(5):e36733, 2012.

[85] Liangsuo Ma, Binquan Wang, Xiying Chen, and Jinhu Xiong. Detecting functional connectivity in the resting brain: a comparison between ica and cca. *Magnetic resonance imaging*, 25(1):47–56, 2007.

[86] Martin J McKeown, Scott Makeig, Greg G Brown, Tzyy-Ping Jung, Sandra S Kindermann, Anthony J Bell, and Terrence J Sejnowski. Analysis of fmri data by blind separation into independent spatial components. *Human brain mapping*, 6(3):160–188, 1998.

[87] Kaiming Li, Lei Guo, Jingxin Nie, Gang Li, and Tianming Liu. Review of methods for functional brain connectivity detection using fmri. *Computerized Medical Imaging and Graphics*, 33(2):131–139, 2009.

[88] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[89] Alex Fornito, Andrew Zalesky, and Michael Breakspear. Graph analysis of the human connectome: promise, progress, and pitfalls. *Neuroimage*, 80:426–444, 2013.

[90] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330–342, 2008.

[91] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[92] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207, 2005.

[93] A Paul Alivisatos, Miyoung Chun, George M Church, Ralph J Greenspan, Michael L Roukes, and Rafael Yuste. The brain activity map project and the challenge of functional connectomics. *Neuron*, 74(6):970–974, 2012.

[94] Lichtman J.W. Kasthuri, N. Neurocartogtaphy. *Neuropsychopharmacol*, pages 342–343, 2010.

[95] Stephen M Smith. The future of fmri connectivity. *Neuroimage*, 62(2):1257–1266, 2012.

[96] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

[97] Anderson dos Santos Siqueira, Claudinei Eduardo Biazoli Junior, William Edgar Comfort, Luis Augusto Rohde, and João Ricardo Sato. Abnormal functional resting-state networks in adhd: graph theory and pattern recognition analysis of fmri data. *BioMed Research International*, 2014, 2014.

[98] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[99] Gopikrishna Deshpande, Peng Wang, D Rangaprakash, and Bogdan Wilamowski. Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE Transactions on Cybernetics*, 45(12):2668–2679, 2015.

[100] Heung-Il Suk, Chong-Yaw Wee, Seong-Whan Lee, and Dinggang Shen. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage*, 129:292–307, 2016.

[101] Wei Liao, Daniele Marinazzo, Zhengyong Pan, Qiyong Gong, and Huafu Chen. Kernel granger causality mapping effective connectivity on fmri data. *IEEE transactions on medical imaging*, 28(11):1825–1835, 2009.

[102] Junghoe Kim, Vince D Calhoun, Eunsoo Shim, and Jong-Hwan Lee. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*, 124:127–146, 2016.

[103] Saman Sarraf and Ghassem Tofighi. Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*, 2016.

[104] Xiaoxiao Li, Nicha C Dvornek, Xenophon Papademetris, Juntang Zhuang, Lawrence H Staib, Pamela Ventola, and James S Duncan. 2-channel convolutional 3d deep neural network (2cc3d) for fmri analysis: Asd classification and feature learning. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1252–1255. IEEE, 2018.

[105] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[106] ADHD-200 Online. http://fcon_1000.projects.nitrc.org/indi/adhd200/, 2011.

[107] Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

[108] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.

[109] Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

[110] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009.

[111] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

[112] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012.

[113] Walter J Trybula. Data mining and knowledge discovery. *Annual review of information science and technology (ARIST)*, 32:197–229, 1997.

[114] Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.

[115] Patricia Riddle, Richard Segal, and Oren Etzioni. Representation design and brute-force induction in a boeing manufacturing domain. *Applied Artificial Intelligence an International Journal*, 8(1):125–147, 1994.

[116] Wael Khreich, Eric Granger, Ali Miri, and Robert Sabourin. Iterative boolean combination of classifiers in the roc space: An application to anomaly detection with hmms. *Pattern Recognition*, 43(8):2732–2752, 2010.

[117] Mahbod Tavallaee, Natalia Stakhanova, and Ali Akbar Ghorbani. Toward credible evaluation of anomaly-based intrusion-detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(5):516–524, 2010.

[118] Z Yang, WH Tang, Almas Shintemirov, and QH Wu. Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(6):597–610, 2009.

[119] Zhi-Bo Zhu and Zhi-Huan Song. Fault diagnosis based on imbalance modified kernel fisher discriminant analysis. *Chemical Engineering Research and Design*, 88(8):936–951, 2010.

[120] Pablo Bermejo, Jose A Gámez, and Jose M Puerta. Improving the performance of naive bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. *Expert Systems with Applications*, 38(3):2072–2080, 2011.

[121] Yi-Hung Liu and Yen-Ting Chen. Total margin based adaptive fuzzy support vector machines for multiview face recognition. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1704–1711. IEEE, 2005.

[122] Claire Cardie and Nicholas Howe. Improving minority class prediction using case-specific feature weights. In *ICML*, pages 57–65, 1997.

[123] Catherine L Blake. Uci repository of machine learning databases, irvine, university of california. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

[124] P Domingos. A general method for making classifiers cost-sensitive. *Artificial Inelligence Group, Instituto Superior Técnico, Lisboa*, pages 1049–001.

[125] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[126] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[127] Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[128] C Cortes and V Vapnik. Support vector machine [j]. *Machine learning*, 20(3):273–297, 1995.

[129] Matthew D Sacchet, Gautam Prasad, Lara C Foland-Ross, Paul M Thompson, and Ian H Gotlib. Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Frontiers in psychiatry*, 6:21, 2015.

[130] Feng Liu, Wenbin Guo, Jean-Paul Fouche, Yifeng Wang, Wenqin Wang, Jurong Ding, Ling Zeng, Changjian Qiu, Qiyong Gong, Wei Zhang, et al. Multivariate classification of social anxiety disorder using whole brain functional connectivity. *Brain Structure and Function*, 220(1):101–115, 2015.

[131] Ali Khazaee, Ata Ebrahimzadeh, and Abbas Babajani-Feremi. Identifying patients with alzheimer's disease using resting-state fmri and graph theory. *Clinical Neurophysiology*, 126(11):2132–2141, 2015.

[132] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[133] Edgar Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. 1997.

[134] Cèline Chayer and Morris Freedman. Frontal lobe functions. *Current neurology and neuroscience reports*, 1(6):547–552, 2001.

[135] Michel Desmurget and Angela Sirigu. A parietal-premotor network for movement intention and motor awareness. *Trends in cognitive sciences*, 13(10):411–419, 2009.

[136] Guang Yang, Xiahai Zhuang, Habib Khan, Shouvik Haldar, Eva Nyktari, Xujiong Ye, Greg Slabaugh, Tom Wong, Raad Mohiaddin, Jennifer Keegan, et al. A fully automatic deep learning method for atrial scarring segmentation from late gadolinium-enhanced mri images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 844–848. IEEE, 2017.

[137] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature Verification using a “Siamese" Time Delay Neural Network. In *Advances in Neural Information Processing Systems*, pages 737–744, 1994.

[138] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[139] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[140] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3286, 2015.

[141] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015.

[142] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5007–5015, 2015.

[143] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015.

[144] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1420–1429, 2016.

[145] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, 2013.

[146] Guillaume Chanel, Swann Jean Antoine Pichon, Laurence Conty, Sylvie Berthoz, Coralie Chevalier, and Julie Grèzes. Classification of autistic individuals by merging information from multiple fmri experiments. 2016.

[147] Xiaoyan Tang, Weiming Zeng, Yuhu Shi, and Le Zhao. Brain activation detection by modified neighborhood one-class svm on fmri data. *Biomedical Signal Processing and Control*, 39:448–458, 2018.

[148] Guillaume Chanel, Swann Pichon, Laurence Conty, Sylvie Berthoz, Coralie Chevallier, and Julie Grèzes. Classification of autistic individuals and controls using cross-task characterization of fmri activity. *NeuroImage: Clinical*, 10:78–88, 2016.

[149] Fabio De Sousa Ribeiro, Liyun Gong, Francesco Calivá, Mark Swainson, Kjartan Gudmundsson, Miao Yu, Georgios Leontidis, Xujiong Ye, and Stefanos Kollias. An end-to-end deep neural architecture for optical character verification and recognition in retail food packaging. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2376–2380. IEEE, 2018.

[150] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[151] Veena Kumari, Emmanuelle R Peters, Dominic Fannon, Elena Antonova, Preethi Premkumar, Anantha P Anilkumar, Steven CR Williams, and Elizabeth Kuipers. Dorsolateral prefrontal cortex activity predicts responsiveness to cognitive–behavioral therapy in schizophrenia. *Biological psychiatry*, 66(6):594–602, 2009.

[152] Stefan P Koch, Claudia Hägele, John-Dylan Haynes, Andreas Heinz, Florian Schlagenhauf, and Philipp Sterzer. Diagnostic classification of schizophrenia patients on the basis of regional reward-related fmri signal patterns. *PloS one*, 10(3):e0119089, 2015.

[153] Xunheng Wang, Yun Jiao, Tianyu Tang, Hui Wang, and Zuhong Lu. Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder. *European journal of radiology*, 82(9):1552–1557, 2013.

[154] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[155] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

[156] Siri DS Noordermeer, Marjolein Luman, Corina U Greven, Kim Veroude, Stephen V Faraone, Catharina A Hartman, Pieter J Hoekstra, Barbara Franke, Jan K Buitelaar, Dirk J Heslenfeld, et al. Structural brain abnormalities of attention-deficit/hyperactivity disorder with oppositional defiant disorder. *Biological psychiatry*, 82(9):642–650, 2017.

[157] Luca Cocchi, Ivanei E Bramati, Andrew Zalesky, Emi Furukawa, Leonardo F Fontenelle, Jorge Moll, Gail Tripp, and Paulo Mattos. Altered functional brain connectivity in a non-clinical sample of young adults with attention-deficit/hyperactivity disorder. *Journal of Neuroscience*, 32(49):17753–17761, 2012.

[158] Liang Wang, Chaozhe Zhu, Yong He, Yufeng Zang, QingJiu Cao, Han Zhang, Qiuhai Zhong, and Yufeng Wang. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Human brain mapping*, 30(2):638–649, 2009.

[159] Ming-guo Qiu, Zhang Ye, Qi-yu Li, Guang-jiu Liu, Bing Xie, and Jian Wang. Changes of brain structure and function in adhd children. *Brain topography*, 24(3-4):243–252, 2011.

[160] Mariya V Cherkasova and Lily Hechtman. Neuroimaging in attention-deficit hyperactivity disorder: beyond the frontostriatal circuitry. *The Canadian Journal of Psychiatry*, 54(10):651–664, 2009.

[161] Tillman Weyde, Gregory Slabaugh, Gauthier Fontaine, and Christoph Bederna. Predicting aquaplaning performance from tyre profile images with machine learning. In *International Conference Image Analysis and Recognition*, pages 133–142. Springer, 2013.

[162] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[163] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[164] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[165] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.