Automating Computed Tomography Analysis for Early Diagnosis of Neurological Diseases

Nandakishor Desai

Submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Electrical and Electronic Engineering

The University of Melbourne

August 2020

Abstract

Neurological diseases are diseases of the nervous system that occur due to structural or biochemical abnormalities in the brain and nervous system. A diverse set of neurological diseases with varied symptoms makes it complicated to diagnose them with a standard protocol. Nevertheless, medical imaging can play a significant role in their early diagnosis by providing an accurate visualisation of internal body structures. However, analysis of the medical images mostly involves significant human intervention in complex disease cases. This process is not only time-intensive, but also laborious, and exhibits inter- and intra-observer variances. To this end, this study contributes to automating the early diagnosis of neurological diseases from computed tomography images.

The first contribution of the thesis involves early diagnosis of cerebral aneurysms from computed tomography angiograms. A large-scale computed tomography angiograms dataset is constructed to investigate the automated diagnosis of unruptured cerebral aneurysms. A novel convolutional neural network architecture is proposed and trained on the dataset to identify aneurysm voxels from the images and subsequently, diagnose its presence in the given image scan. The proposed approach achieves a sensitivity of 92% in diagnosing aneurysms and a dice score of 65.2% in their localisation, thus demonstrating the efficacy of the proposed work.

The second focus is on Parkinson's disease, a neurological disease affecting the control of body movements. It can cause significant speech impairment early its course. Therefore, analysing the abnormalities in vocal fold movements during phonation

can be a useful indicator for early signs. Computed tomography is an efficient imaging modality that captures dynamic vocal fold movements with a good spatial and temporal resolution. Therefore, it allows for a direct assessment of the movements of vocal folds and associated structures. A large-scale image dataset is constructed by capturing computed tomography scans of the neck during vocalisation period. First, a basic image processing-based approach is proposed that helps to explore and identify clinically useful feature points from arytenoid cartilages supporting the vocal fold movements. Further, convolutional neural network-based object detector is trained to fully localise the arytenoid cartilages. Inter arytenoid distance feature is then extracted to demonstrate its utility in differentiating Parkinson's patients from healthy controls.

In this final part of the contribution, novel machine learning interpretability techniques based on canonical correlation analysis, are proposed that assist in interpreting the representations learned by convolutional neural networks designed for the specific medical image analysis tasks. A set of novel two-dimensional multiset canonical correlation analysis algorithms are proposed that effectively capture the linear relationships between learned feature representations within and between neural networks. Results are presented by employing the proposed interpretability techniques to analyse the learned representations of neural networks trained to segment cerebral aneurysms from computed tomography angiograms.

In summary, the thesis contributes to automating the analysis of computed tomography images for early detection of neurological diseases.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Nandakishor Desai, August 2020

Acknowledgements

"The only true wisdom is in knowing you know nothing." - Socrates

The pursuit of PhD has been an invaluable experience enriching me with new perspectives, imparting me the courage to say "I can do anything", and also humbling me at every stage to say "I know nothing". It has been an adventurous journey that will significantly shape my personal and professional life.

I would like to thank my supervisor Prof Marimuthu Palaniswami. He has been a supervisor, father-figure, and a philosophical teacher guiding me to a PhD, sharing his vast knowledge, and imparting life lessons. The good-natured banter was the highlight of my PhD life. I would also like to thank Prema Palaniswami for her gracious luncheon invitations and making me feel at home through this long journey. The foundation of my research journey started at IIT Kharagpur under the valuable guidance of Prof Jayanta Mukhopadhyay, who is a collaborator in this PhD. I am grateful for his constant support and guidance especially during early days of my research life.

I would like to thank Prof Bernard Yan, who has been an incredible collaborator and an invaluable mentor, continually guiding me and putting in the hard yards with me in all the stages of the project from planning to execution.

I would like to thank Aravinda Rao, who has been a friend, a colleague, and a mentor, helping me navigate through difficult stages. He has been a constant support, and our discussions over coffee were enriching experiences.

I would also like to thank Prof Dominic Thyagarajan, who has been a constant presence through my PhD and has imparted his medical knowledge with utmost patience. I would also like to thank Paari Palaniswami, who has been a friend and valuable collaborator, continually providing me with new and exciting insights.

My extended greetings to Prof Kenneth Crozier, for chairing the PhD committee. He has been a generous and kind chair continuously supporting my candidature with constructive feedbacks and invaluable insights.

I am grateful to my labmates and friends: Punit, Nitisha, Shitanshu, Karishma, Shreyasi, Nandini, Bigi, Motin, Radha, and Emerson. Special thanks to my hangout buddies: Sanjay, Sanchari, Joydip, and Anu. You have all made my PhD journey pleasing and memorable. I should also thank my long distant rescue team from KGP days: Prantik, Sumeet, Ashwini, Anurag, and Mehul for always keeping the tempo high. Special mentions to Sumanth and Sunil, genius buddies from Mysore days and the best mates that one can have.

Finally, I must express my gratitude towards my parents without whom my existence would be neither possible nor meaningful. Heartfelt thanks to my little sister for bearing me throughout and supporting me. Special mentions to Mini, Shamant, RR and RP Desai, Dr BG and GP Desai for the continued encouragement. My shoutout to loving granny who epitomises constant learning and hunger for knowledge and has always inspired me with her scientific approach to life.

Preface

All the work in this thesis was conducted by the author of this thesis, including theoretical analysis, algorithm development, experiments, and manuscript writing. The thesis has not been submitted for other qualifications. All the work towards the thesis was carried out after the enrolment in the degree. No third party editorial assistance was provided in the preparation of the thesis.

The two datasets used in the thesis were acquired at (i) Melbourne Brain Centre, Royal Melbourne Hospital, Australia (approved by Melbourne Health Human Research Ethics Committee: HREC.2013.072) and (ii) Movement Disorders Clinic, Monash Medical Centre, Australia (approved by Research Ethics Committee of Monash Health: Application # 11230B).

The following sections list the publication during the candidature period. Section 'Thesis-related Publications' consists of publications that directly contribute to the thesis work. Section 'Other Publications' consists of publications from other collaborative research works with Melbourne Brain Centre, Royal Melbourne Hospital, Australia.

Thesis-related Publications

1. **Nandakishor Desai**, Abd-Krim Seghouane, and Marimuthu Palaniswami. "Multisubject fMRI data analysis via two dimensional multi-set canonical correlation analysis." In Biomedical Imaging (ISBI 2017), 2017 IEEE 14th Inter-

national Symposium on, pp. 468-471. IEEE, 2017.
This publication describes a novel two-dimensional multiset canonical correlation analysis approach for multisubject image analysis. It links to Chapter 6.

2. **Nandakishor Desai**, Aravinda S. Rao, Paari Palaniswami, Dominic Thyagarajan, and Marimuthu Palaniswami. "Arytenoid cartilage feature point detection using laryngeal 3D CT images in Parkinson's disease." In Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, pp. 1820-1823. IEEE, 2017.
This publication includes a novel exploratory approach, consisting of basic image processing techniques, to detect clinically relevant arytenoid cartilage feature points from neck computed tomography images. This work is connected to Chapter 5.

3. **Nandakishor Desai**, Abd-Krim Seghouane, and Marimuthu Palaniswami. "Algorithms for two dimensional multi set canonical correlation analysis." Pattern Recognition Letters 111 (2018): 101-108.
This publication discusses novel algorithms for two dimensional multiset canonical correlation analysis that can be applied to analyse the multiple sets of image data. It is connected to Chapter 6.

4. **Nandakishor Desai**, Aravinda S. Rao, Bernard Yan, and Marimuthu Palaniswami. "Automated detection and localisation of cerebral aneurysms from computed tomography angiogram images", to be submitted.
This publication constitutes a novel convolutional neural network approach to automatically detect and localise cerebral aneurysms from computed tomography angiogram images. It contributes to Chapter 4.

5. **Nandakishor Desai**, Aravinda S. Rao, Dominic Thyagarajan, and Marimuthu Palaniswami. "Analysis of movement of arytenoid cartilages from neck CT Images in Parkinson's patients", to be submitted.

This publication constitutes automating the analysis of movements of arytenoid cartilages from neck computed tomography images to detect vocal fold abnormalities in Parkinson's disease. It contributes to Chapter 5.

6. **Nandakishor Desai**, Aravinda S. Rao, Bernard Yan, and Marimuthu Palaniswami. "Interpretability of CNN representations using canonical correlation analysis methods" to be submitted.
   This publication consists of canonical correlation analysis-based machine learning interpretability techniques that assist in a better understanding of the inner workings of the convolution neural network models designed for medical image analysis task. It contributes to Chapter 4.

Other Publications

1. Weeden M, **Desai N***, Sriram S, Palaniswami M, Wang B, Talbot L, Deane A, Bellomo R, Yan B, "A pilot feasibility and exploratory study of accelerometry-based sedation monitoring in critically ill patients" submitted to critical care medicine.

2. Seghouane, Abd-Krim, Asif Iqbal, and **Nandakishor Desai**. "BSmCCA: A block sparse multiple-set canonical correlation analysis algorithm for multi-subject fMRI data sets." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

To my grandfather, who led a life full of hardships to ensure ours is a bed of roses.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Networks
AC	Anterior Commissure
ADHD	Attention Deficit Hyperactivity Disorder
ASPP	Atrous Spatial Pyramid Pooling
AUC	Area Under the Receiver Operating Characteristics
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Networks
CT	Computed Tomography
CTA	Computed Tomography Angiography
DICOM	Digital Imaging and Communications in Medicine
DSA	Digital Subtraction Angiography
DSC	Dice Coefficient Score
ECG	Electrocardiography
EEG	Electroencephalogram
EMG	Electromyography

EM	Expectation-Maximisation
IoU	Intersection Over Union
IAD	Inter Arytenoid Distance
MLP	Multilayer Perceptrons
NCG	Nerve Conduction Study
MRA	Magnetic Resonance Angiogram
MRF	Markov Random Fields
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative
PACS	Picture Archiving and Communication Systems
PD	Parkinson's Disease
ROI	Region of Interest
RCNN	Region CNN
PET	Positron Emission Tomography
SPET	Single Photon Emission Tomography
SAH	Subarachnoid Hemorrhage
SPP	Spatial Pyramid Pooling
SVM	Support Vector Machine
SVCCA	Singular Value CCA (SVCCA)
2DMCCA	Two-dimensional Multiset CCA

Chapter 1

Introduction

This chapter gives a brief overview of neurological diseases, their diagnosis using imaging modalities, and the need for automated methods to analyse the images towards diagnostic assistance. It summarises the overall objectives and outlines the structure of the proposed contribution.

1.1 Background

Neurological diseases refer to diseases of the nervous system that arise due to structural or biochemical abnormalities in the brain and associated nervous system. Neurological diseases put a significant burden on global health and economy with high mortality and morbidity rates. They are the second leading cause of death after cardiovascular diseases and the leading cause of disability, with a mortality rate of about 12% [1].

The neurological diseases can be classified by their primary cause, the primary location of origin, and primary dysfunction. There are several common forms of neurological diseases that disrupt the nervous system significantly. Neurodegenerative diseases are a group of diseases characterised by progressive degeneration of the structural and functional aspects of the nervous system. Alzheimer's disease, Parkinson's disease are some of the important neurodegenerative diseases. Cerebrovascular diseases are another vital category of neurological diseases that affect the vascular system. Common examples include brain aneurysms, stroke, vascular malformation, and others. Cerebrovascular diseases, and stroke, in particular, are some of the main

non-communicable diseases with a wide impact. Epileptic seizures, brain tumours, multiple sclerosis are some of the other usually occurring neurological diseases. Figure. 1.1 shows the mortality rates due to various neurological diseases.

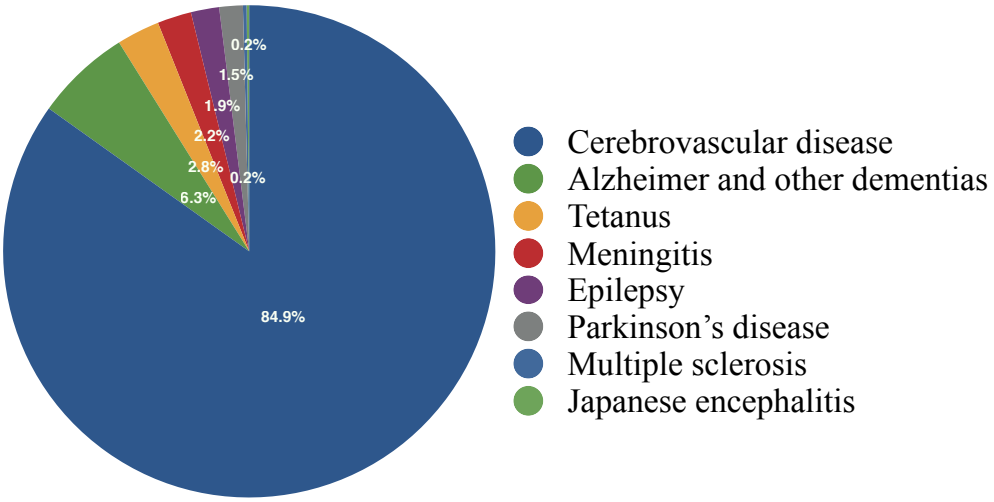


Figure 1.1: Mortality rates of neurological diseases. Data source: [1]

Neurological diseases are a diverse ensemble of diseases with varying causes and symptoms and some conditions with combinations of causes and symptoms, making it complicated to follow a standard diagnostic protocol. In addition to routine medical check-ups, some dedicated diagnostic tools include analysis of indirect electrical signals from the nervous system. Electroencephalogram (EEG) records the electrical activity of the brain, electromyography (EMG) and nerve conduction study (NCS) analyse the electrical activities of the muscles, and evoked potentials record brain's response to external stimuli among many other such assessment techniques. However, most of these methods are indirect and invasive, making them inefficient and reliant on indirect signals. In addition, invasive methods may also cause significant discomfort to the patient. Imaging, on the other hand, has added new dimensions to the field of medicine, greatly enhancing the capabilities of medical diagnostics.

Imaging has added an efficient diagnostic modality to medicine, with an ability to directly visualise the human body for further inferences. A visual representation of the relevant body region helps physicians to directly observe and assess rather

than rely on indirect measurement signals. A computed tomography (CT) scan using x-rays is generally modality of immediate choice in stroke-related bleedings and emergencies. Digital subtraction angiogram (DSA) is an x-ray based imaging enabling an accurate visualisation of the vascular system. CT angiography is an increasingly emerging, better and efficient alternative to DSA. Magnetic resonance imaging (MRI) is used to capture image representations with high-resolution. Ultrasound, positron emission tomography (PET) are some other frequently used imaging modalities.

The increased usage of medical imaging to visualise a diverse set of anatomical regions and to diagnose disease applications with complex characteristics have made it challenging to perform a robust manual analysis of the images. Further, increased availability and affordability of the imaging technology has been suitably augmented by the Internet's explosive growth to enable large-scale aggregation of the image data. A manual analysis of the image data would increase chances of a missed diagnosis with a potentially high cost, in addition to inducing inter- and intra- observer variances to the diagnostic accuracy. Therefore, efficient computerised approaches have increasingly become necessary to effectively analyse the image data and assist the clinicians in achieving high diagnostic accuracy. This study focuses on proposing and developing automated image analysis approaches for the early diagnosis of neurological diseases from CT images. CT imaging is commonly available in most hospitals (both in rural and urban areas). It is also cheaper, quick, and efficient, making it widely accessible. Further, it is applicable to a majority of disease conditions and anatomical regions. Therefore, this wide-availability, wide-applicability, and efficiency of CT imaging is the primary motivation for this work, focusing on two neurological diseases.

The first contribution investigates the development of automated approaches towards the diagnosis of cerebral aneurysms. A cerebral aneurysm is a form of the cerebrovascular disease characterised by weakness of the blood vessel wall. A severe complication with aneurysms is their rupture, leading to extravasation of blood

into the subarachnoid space, known as subarachnoid hemorrhage (SAH). Ruptured aneurysms are fatal in 40% of the cases, with around 46% of the survivors suffering from some form of long term cognitive impairment [11]. Therefore, the study focuses on an automated approach to diagnose unruptured aneurysms in the first contribution so that they can be diagnosed with greater accuracy and treated in time before the rupture.

The second contribution examines Parkinson's disease. It is a hypokinetic movement disorder that affects mainly the motor system and may result in the paucity of voluntary/involuntary movements. It is the second most common neurodegenerative disorder with an incidence rate of 2-3% amongst the population aged 65 or more [12] with an estimated 10 million patients in the world [13]. Early diagnosis of Parkinson's may help the clinicians to provide timely treatment and slow down its progress. The focus of the thesis revolves around investigating the relationship between speech abnormalities and the progress of Parkinson's disease. Around 70-89% of the Parkinson's patients suffer from abnormal vocal fold movements [14–16]. Therefore, movement of vocal folds/speech quality may be a useful clinical feature for the early detection of Parkinson's disease.

Large-scale image datasets are aggregated, and novel CNN-based automation approaches are proposed to analyse and understand the diseases from the images. Further, machine learning interpretability techniques are proposed that assist in acquiring a better understanding of the learning process to derive meaningful and easily understandable knowledge about the automation approach designed for a particular task.

1.2 Objectives

The objective of this work is to develop automated CT image analysis approaches to assist in the diagnosis of neurological diseases, and specifically:

- to detect and localise unruptured cerebral aneurysms from CTA examinations

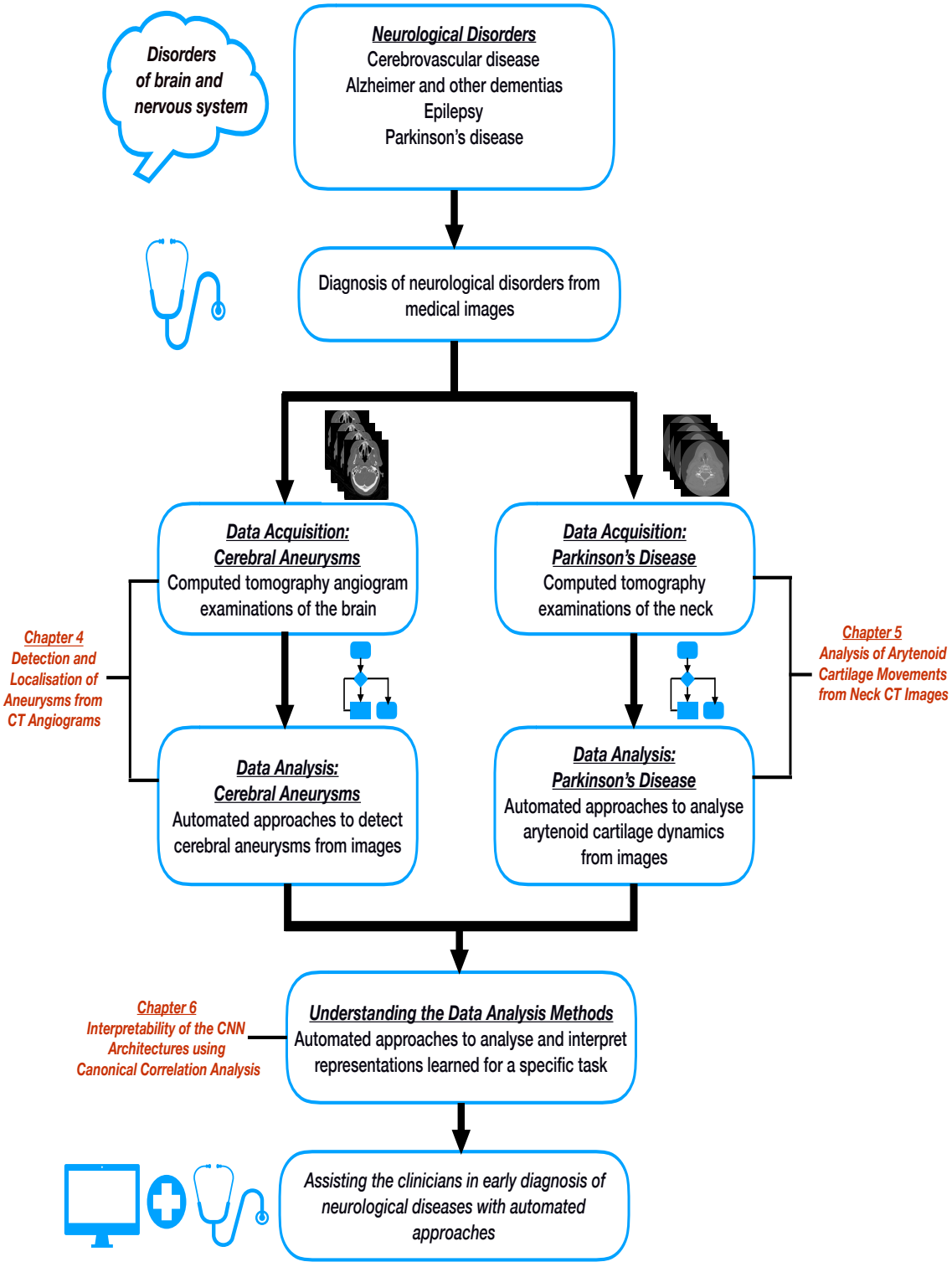


Figure 1.2: An infographic summary of the thesis objective and contributions.

of the brain. The aim is to detect the unruptured aneurysms and localise them through a region-of-interest.

- to assist in the analysis of vocal fold abnormalities from CT images of larynx. The idea is to analyse vocal fold movements by detecting and analysing the movements of supporting cartilages.

- to design and develop interpretability techniques that assist in analysing and a better understanding of the machine learning-based automated approaches. The aim is to complement the machine learning-based automated approaches with data driven interpretability tools for a better understanding of the system.

1.3 Challenges

The requirements for sophisticated automated analysis of the medical images has increased with the advancements in medical imaging technology. Non-invasive imaging and direct visualisation of the target region has greatly enhanced diagnostic accuracy; however, manual intervention and analysis of images is still a bottleneck in several applications. It is time-intensive, error-prone, and inefficient. There have been significant strides in image analysis techniques in general and in particular medical applications. However, there are several generic as well as domain-specific challenges as outlined below.

1.3.1 Generic Challenges

- It is difficult to acquire large amounts of medical data because of privacy issues, nature of certain diseases, availability, and affordability of the imaging facility.

- Ground truths are essential to validate automated approaches to assess whether they provide satisfactory performance. Obtaining large-scale ground truth is, however, not a straightforward task as it is time-intensive and requires relevant expertise.

- From the complex appearance of small brain aneurysms to capturing the dynamic movements of vocal folds from neck images, several other domain challenges exist, in developing automated approaches for specific applications.

- Machine learning algorithms for automated medical image analysis have varied time and space complexities. Further, algorithms processing large-scale complex datasets are harder to study because of their size, and the distributed data-parallel nature of the model development process.

1.3.2 Domain-specific Challenges

Cerebral Aneurysms

- Aneurysms anywhere within the internal carotid artery may easily be obscured by bone artefact, making it difficult to visualise.

- Arteries may loop and wind around each other. The looping can sometimes look like an aneurysm that makes it difficult to untangle the loop visually. It may require much time spent in a multi-planar examination that might still result in erroneous conclusions.

- DSA is usually considered to be a gold-standard in diagnosing cerebral aneurysms with high sensitivity. However, it is an invasive procedure requiring an expert operator to guide the catheter to significant blood vessels while trying to minimise injury risk. It is invasive and stressful for the patients.

Parkinson's Disease

- The standard motor symptoms associated with the progression of Parkinson's may not immediately manifest at the onset of the disease. However, perceptual voice changes (i.e., employing auditory perception to assess the voice characteristics) measured on the Unified Parkinson's Disease Rating Scale (UPDRS) were noted as far back as 9.8 years before diagnosis in [17].

- Laryngography, videolaryngostroboscopy, electromyography are some of the conventional techniques usually used to investigate the speech abnormalities. A major drawback of most of the works mentioned above is that they are invasive, causing significant patient discomfort.

- Additionally, they provide only surface visualisation of the vocal folds and fail to capture 3D movements. Further, most of them do not produce any objective measurement.

- Stroboscopic based techniques can be subjective as they rely heavily on operator skills, which can create significant inter- and intra-observer variance.

Interpretability

- Some of the significant interpretability techniques [18–20] focus on interpreting the CNN layers in isolation. They do not concentrate on analysing the population representations of the CNN layer and the relationships between layers.

- Methods, as mentioned above, are not well suited to compare representations across network architectures and to analyse multiple solutions simultaneously.

- The approaches need to be computationally inexpensive so that they can be used during the CNN training process.

1.4 Scope of the Thesis

This study attempts to address the challenges outlined in Section 1.3. It proposes automated approaches, using CNN architectures, to analyse neurological diseases from computed tomography images. Further, the 'blackbox' nature of CNNs is often considered a drawback in medical applications. The thesis proposes interpretability techniques that assist in analysing the automated approaches to get a better understanding of the internal dynamics of CNNs designed for a specific task.

1.5 Research Contributions

- The first contribution involves a novel automated assistance tool to diagnose cerebral aneurysms from CTA images.
  - A large scale image dataset is constructed consisting of DSA verified CTA examinations of brain aneurysms.
  - The aneurysm diagnosis objective is formulated as a dense voxel prediction problem.
  - Novel CNN-based segmentation architecture is proposed to identify and label aneurysm voxels.
  - An integrated prediction pipeline is developed that predicts aneurysm voxels, localises the aneurysm, and diagnoses the presence of aneurysm in a given CTA examination.
  - To the best of our knowledge, this is the first such approach for large scale automation of detecting aneurysms from CTA images, done concurrently with [21].
- The second contribution involves investigating the speech abnormalities in Parkinson's patients from the neck CT images.
  - A large scale neck CT dataset is constructed consisting of imaging examinations of neck acquired during phonation/breathing.
  - The vocal fold abnormality analysis is carried out by analysing the movements of arytenoid cartilages, which are attached posteriorly to the vocal folds.
  - An initial new exploratory approach is proposed to identify clinically relevant features from the arytenoid cartilages supporting the vocal folds.
  - Furthermore, a CNN-based architecture is employed to accurately localise

the arytenoid cartilages for further analysis.

  - To the best of our knowledge, this is the first work to focus on the automated analysis of neck CT images for investigating abnormal vocal fold movements.

- The third contribution revolves around designing novel interpretability techniques to analyse the learned representations from CNN architectures.

  - The interpretability problems are formulated as analysis of population representations of the CNN architectures.

  - Novel canonical correlation analysis (CCA)-based algorithms are developed to interpret and capture the relationships amongst the layer representations.

  - Layer representations are analysed by treating them as a collection of neurons, and filter feature maps are analysed by treating the layers as a collection of filter kernels.

1.6 Thesis Outline and Approach

Automated medical image analysis usually comprises two main stages: data acquisition and data analysis. However, with the emergence of machine learning-based automation approaches, a third stage needs to be considered - that helps the end-users to understand the decision-making process of a machine learning model. Therefore, the thesis focuses on the three stages of the medical image analysis: (i) data acquisition - retrospectively acquiring CT examinations to construct large-scale datasets for neurological diseases, (ii) data analysis - novel automated approaches to analyse the diseases from the images, and (iii) novel interpretability procedures to interpret the inner workings of the automated methods. Chapters 2 and 3 serve as introductory chapters that give a detailed overview of medical imaging modalities, existing automation approaches, the emergence of CNNs and their applications in

medical image analysis. Chapters 4,5, and 6 form the main contribution chapters outlining the neurological disease datasets, their analysis, and ways to interpret the CNN-based machine learning models. Chapter 7 concludes the thesis with a brief summary of the contributions and an outline of future research directions. Figure 1.2 shows an infographic summary of the thesis objective and contributions.

Chapter 2

This chapter gives a detailed introduction to medical imaging and its advantages over other diagnostic techniques. It outlines a multitude of imaging modalities and their advantages and disadvantages and applicability in different scenarios. It then details approaches to analyse the medical images and their necessity. Various image processing techniques, including histogram processing, image filtering, morphological operations, are described. Conventional image segmentation methods, along with their advantages and shortcomings, are then summarised. Chapter 2 provides a comprehensive overview of medical imaging and traditional processing techniques for image analysis.

Chapter 3

CNNs have greatly enhanced the field of image recognition and image analysis in general. It subsequently has impacted the analysis techniques in medical images. This chapter starts with an introduction to artificial neural networks (ANNs) and their construction. It then analyses CNNs and their state-of-the-art performances in several image analysis tasks, including image classification, object localisation, semantic segmentation, and others. The challenges in adapting CNNs to medical images analysis and some existing works that try to overcome these challenges, to increase the performance bounds, are outlined to conclude the chapter.

Chapter 4

This chapter describes the first contribution of the thesis towards the analysis of brain aneurysms from CTA images. Unruptured cerebral aneurysms pose a high mortality risk with a possible rupture consequence of stroke. This work first constructs a large-scale dataset of brain aneurysms from retrospectively acquired CTA images. It then proposes a novel CNN architecture to detect and localise aneurysms from the CTA images with high accuracy. The diagnostic, as well as localisation accuracies, are presented to demonstrate the effectiveness of the approach.

Chapter 5

This chapter summarises the second contribution of the thesis towards the analysis of vocal fold movements from neck CT images and investigating their abnormal movements in Parkinson's disease. An initial dataset consisting of Parkinson's and healthy controls are collected from a movement disorder clinic. An initial exploratory approach is first proposed to extract clinically relevant feature points from the arytenoid cartilages that support the movement of vocal folds. The dataset is then augmented with a large number of CT scans from healthy controls acquired during a breathing interval to construct a large-scale dataset of neck CT images. A convolutional neural network object detector is subsequently developed to localise the arytenoid cartilages for further analysis. Detailed localisation accuracies and simple feature information differentiating healthy control from Parkinson's are further presented to conclude the chapter.

Chapter 6

This chapter introduces contributions that complement the automated approaches presented in Chapters 4 and 5. The main objective of this work is to design and develop automated interpretability techniques that are useful to interpret the feature representations learned by a CNN architecture for a designated task. Novel interpretability tools employing CCA, a data-driven statistical tool, are proposed

to analyse the CNN representations. The chapter concludes with demonstrations of the techniques by applying them to explain the CNN representations learnt during a cerebral aneurysm detection task.

Chapter 7

This chapter provides a brief overview of the contributions of the thesis and an overall conclusion of the thesis. Future directions are also outlined in the process.

Chapter 2

Overview of Medical Imaging Modalities and Traditional Computing Methods

This chapter presents an overview of medical imaging technologies, different imaging modalities, and their applicability and limitations. Each imaging modality has its advantages, and using it to acquire data is usually the first step in medical image analysis. The next step involves the analysis of the image data to make further inferences. This chapter outlines some fundamental image processing methods that are useful to perform basic data analysis.

2.1 Introduction

Advancements in medical imaging technology have made it possible to capture the image representations of human organs with high accuracy. Digitisation of imaging technology and the rapid evolution of the computing software have contributed to a dramatic increase in the use of medical imaging in a clinical setting for diagnostic purposes. A visual representation of the relevant body region helps physicians to directly observe and assess rather than rely on indirect measurement signals (such as Electrocardiography, EMG). The computerised visual assessment of the human body has had a significant impact on the early diagnosis of several diseases leading to the prevention of fatal consequences through timely treatment. Over time, a multitude of imaging modalities has developed to capture numerous types of com-

plementary information about the human body. X-ray based imaging technologies such as planar radiography capture two-dimensional images efficiently with minimal exposure to radiation and are one of the oldest and widely used imaging modalities. X-ray-based CT imaging improves upon the radiography to be able to capture a three-dimensional image representation. X-ray-based imaging technologies are in-general fast, efficient, affordable, and widely available. However, exposure to radiation limits their frequent usage. MRI is another modality that does not use any radiation and captures image representation by extracting magnetic field information and the alignment of photons in the desired region. MRI produces high-resolution images and does not expose the body to any radiation. However, it is neither readily available nor easily affordable. Nuclear medicine is another form of imaging that is employed in positron emission tomography (PET) and single-photon emission tomography (SPET) to capture functional and anatomical image representations. Small amounts of radioactive materials (radioactive tracers) are introduced into the body that accumulate and decay over time. The emissions from radioactive tracers are then detected to construct image representations. A common disadvantage of the modalities mentioned above is that they capture static regions of interest with ease but, fail to image moving portions of the body. Ultrasonography is an imaging modality that produces images with high temporal resolution and assists in visualising moving regions of interest. Besides, it is affordable and does not expose to any harmful radiation; therefore, it can be used frequently without any hindrances. Ultrasonographic images, however, do not provide a comparable spatial resolution to that of other modalities [22–27].

It can be seen that there are certain advantages and limitations in each imaging technology. Their use depends primarily on the desired application and is more beneficial when used in conjunction to extract complementary information. A few widely used imaging modalities, their advantages and limitations, applicability, and other advancements are discussed in detail in the following section.

2.2 Medical Imaging Modalities

2.2.1 Computed Tomography Imaging

CT imaging uses X-rays to image human organs. A beam of X-rays is transmitted through the region of interest (ROI) in a CT machine, covering its field of view. The X-ray beam passes through the ROI and is received by the detectors on the other side. The amount of attenuated X-ray is a function of the composition of ROI and is, therefore, used to establish its cross-sectional view. Through transmitting the X-rays at various projection angles, multiple cross-sectional views are acquired simultaneously. These sectional views are then combined via geometry processing techniques to construct a three-dimensional volumetric representation of the ROI [28–30].

Depending on the structure of the X-ray beam, there are three main modes of data acquisition during a CT scan - a parallel beam, a fan beam, and a finely collimated cone-beam [31,32]. In general, a parallel beam and fan-beam CT produce images with relatively higher spatial resolution compared to the images produced using cone-beam of X-rays. Nevertheless, they use more X-rays inefficiently and eventually increase the exposure to radiation. A cone-beam CT, on the other hand, is a 3D array of fan beams and effectively captures a 3D image in a quicker time with significantly reduced exposure to radiation. The use of chemical substances known as radiocontrast is another widely used supplementary technique to improve the visibility of the ROI to produce images with enhanced contrast. A bolus of radiocontrast is injected intravenously into the body before the CT representations are acquired. CT imaging generally uses iodine-based chemical agents and in such cases, is called contrast-enhanced CT [33].

CT imaging technology has drastically progressed over the years and is generally affordable with wide availability and accessibility. It is typically the first choice of modality in several cases, despite limited exposure to radiation [34]. CT scan of the brain is usually used to determine hemorrhage, infarction (amount of tissue

that is dead after a stroke), calcification [35]. Cardiac CT scan is another critical application to visualise the cardiac anatomy in the diagnosis of coronary artery disease [36]. Contrast CT scan is usually employed to image the artery and veins to understand vascular diseases better. This is known as CT angiogram [37]. A neck CT scan is usually employed towards understanding thyroid cancer and other neck-related diseases [38].

2.2.2 Magnetic Resonance Imaging

MRI uses magnetisation and nuclear spin properties of an object to encode its information, and construct image representation. A strong static magnetic field is used that initially aligns the protons in the body molecules with the field. A minor radiofrequency pulse magnetic field is subsequently applied that displaces the protons out of equilibrium. The time to return to the equilibrium state and the energy released during the process are a function of the chemical nature of the molecules. This information is used to reconstruct and produce a volumetric image representation of high resolution [39–42].

The critical difference with CT imaging is that the MRI imaging does not employ any form of radiation and instead uses the magnetic properties of body molecules to acquire image representations. Furthermore, to construct a volumetric representation, CT imaging needs to obtain cross-section views at multiple projection angles. MRI, however, is inherently a 3D imaging procedure that can be used to acquire the data in any plane by employing appropriate radiofrequency pulses. MRI is also suitable for functional imaging and is widely used to image a functional brain map to assess the active brain regions during various cognitive tasks. MRI scanning, though, is a long and uncomfortable procedure. In addition, the technology is expensive and may not be widely available.

Over the years, MRI imaging has evolved to use stronger magnetic fields. Compared to CT imaging, muscles, soft tissues, and ligaments are more evident in an MR image. MR angiogram is used to image the vascular system [43]. MRI is generally the first

choice of diagnostic for neurological diseases and conditions that affect the central nervous system as it provides the best contrast between grey matter and white matter [44]. Functional MRI is commonly used to design numerous studies and task paradigms to understand brain activations and functionalities better [45].

2.2.3 Ultrasonography

Ultrasonography or medical ultrasound is a form of imaging technology that generates object representations using high-frequency sound waves. An electrical transducer (probe) is used to convert the electrical signals to high-frequency sound waves. The sound waves transmitted into the body echo off of the bodily structures and are reflected back to the transducer. These reflected waves are converted back to electrical signals by the transducer, which are then fed to a computer to construct an image representation of the structure. Different structures have different reflective properties and act as differentiable information in encoding image representation [46, 47].

A significant difference between ultrasound imaging and CT, MRI imaging is the ability to construct real-time representations, which is useful to image moving structures with high temporal resolution. Further, it does not use any ionising radiation and is portable to use at the bedside. Ultrasound scanning can, therefore, be done efficiently, repeatedly, and without any harmful effects of radiation. However, it is constrained by the inability of sound waves to penetrate gas and bone in body structures and also does not create high spatial resolution images, making differentiating between structures at times challenging.

Cardiology is an important application of ultrasound imaging to assess the abnormalities in heart rate and other structural abnormalities. Visualisation and assessment of the growth of a fetus during pregnancy is another critical application [48]. Ultrasound imaging is often used to view abdominal structures [49] (liver, kidney, pancreas, spleen, and others) and is generally also the first diagnostic option to image appendix and related inflammation [50].

2.2.4 Miscellaneous

Other frequently used imaging modalities include nuclear medicine-based PET [51] and SPET [52] capture functional and anatomical representations of objects. Small amounts of radioactive materials (radioactive tracers) are introduced into the body that accumulate and decay over time. Radioactive tracer emissions are then detected to create image representations. PET imaging is frequently used in combination with CT imaging (PET-CT) to combine their functional and anatomical image representations, respectively.

2.3 Medical Image Computing

The two primary stages of medical image analysis are image acquisition and image computing. The image computing methods enable a computerised assessment of the medical images towards extracting clinically relevant knowledge from the images for subsequent clinical assessment and diagnostic inferences. Imaging technology has evolved rapidly over the past few decades producing better images with an improved spatial and temporal resolution. The increased usage of medical imaging to visualise various anatomical regions and to diagnose diseases with complex characteristics have made it challenging to carry out manual analysis of the images. Further, increased availability and affordability of the imaging technology has been suitably augmented by the era of connectivity to enable large-scale aggregation of the image data spanning several pathologies and anatomical regions. These developments have accelerated further the need for efficient image computing techniques that are capable of analysing a diverse set of image modalities with varied resolution capabilities and have the ability to learn from large quantities of data effectively. The image computing techniques have a notable impact on two broad areas: (i) assisting the clinicians in a diagnostic setting, (ii) assisting the medical researchers in gaining new knowledge. Medical image computation methods cover a wide range of objectives from the ability to visualise image data, to identifying specific regions of interest within the image, to the extraction of structured quantitative information from the

data. Diverse techniques exist that span preprocessing methods improving image quality, as well as computing methods extracting new information. The computing methods can be grouped into different categories depending on their overall objective. Some relevant categories of techniques include: (i) interactive visualisation of 2D/3D image data for initial data exploration; (ii) image enhancement techniques to remove noise and improve image contrast; (iii) image segmentation to accurately identify regions of interest in the image with minimal human interaction; (iv) statistical analysis to derive quantifiable and structured knowledge from the data; (v) image registration to align images from different acquisitions. In addition to the aforementioned conventional image computing methods, pattern recognition and machine learning advances in recent years have added new dimensions to analysing the medical images, and have also significantly pushed performance boundaries of the conventional methods. Some of the basic medical image computing methods and their evolution are described in the following sections.

2.3.1 Image Enhancement

The primary objective of image enhancement methods is to improve the quality of the given image so that it is easier for subsequent methods to process the desired feature information from the images. Usually, image enhancement involves some form of mapping function that maps the pixel intensities to a domain that distinguishes the desired features effectively from non-relevant background information. Some of the simple image enhancement techniques include pixel intensity transformations in which basic mathematical operations such as logarithmic transformation, power-law transformations are used to transform pixel intensities. Intensity scaling is often used to enhance the focus on desired pixel intensity ranges. Piecewise linear transformations such as contrast stretching, grey-level slicing improve low contrast images and enhance intensity regions in specific windows.

Histograms are another compact way of representing the distribution of pixel intensities to understand the content of the image and its properties. Histograms represent

the relative frequency of occurrence of intensity levels in the image. Histogram equalisation is a technique that equalises the histogram of an image to produce a uniform histogram, in which most intensity levels are distributed approximately evenly. It usually improves the global contrast of the images and enhances the low contrast regions in the image to attain high contrasts. Histogram matching is another technique in which the image histogram is transformed to match a specific histogram specification. Local histograms can be useful to determine localised intensity distribution specific to regions of interest in images. Histograms are usually used to determine an appropriate level of threshold intensity to threshold images. Histograms are powerful and computationally inexpensive methods that are widely used to explore unknown image data, enhance its contrast, and derive several useful quantitative image measurements [53, 54].

Image filtering is another widely employed technique, whether as a preprocessing step or as an image enhancement step, to enhance image contrast and extract useful knowledge. Usually, this is done either in the spatial domain by directly operating on the pixel intensities or in the frequency domain by first translating the pixel intensities into the frequency domain through Fourier transform, Wavelet transform, and other related transformations. Low-pass filters and its variations in the spatial domain are used to eliminate noise or small details and construct a more uniform and smooth image with less abrupt intensity changes. Gradient filters are helpful to determine orders of intensity changes and extract useful information such as curves, edges, and other such image shapes related information. Median filters and variations are used to remove specific noise like salt and pepper noise. Fourier transforms are usually used to extract frequency domain information to carry out filtering in the frequency domain. It includes low-pass filters that filter out abrupt changes and high-frequency content; high-frequency filters that enhance edges and boundaries; unsharp masking and high-boost filtering to maximise the contribution of the original image to the filtered image; and several others. Wavelets are another widely used transformation to carry out filtering as they provide a better spatial resolution as well as a frequency resolution. Wavelets are useful in image denoising

and constructing image pyramids at multiple resolutions [53, 54].

Morphological operations are another group of non-linear operations related to the processing of morphological characteristics in an object and are frequently used as a preprocessing or to extract basic image features. Each pixel in the image is modified based on specific neighbourhood properties, and a small template called the structuring element defines the size and shape of the neighbourhood. The neighbourhood property used to modify pixel intensities depends on the type of operation performed. Some of the fundamental morphological operations include erosion (eliminates irrelevant details in the image), dilation (bridges gaps in images), opening (generally smoothes contours of an object and eliminates narrow breaks), and closing (generally smoothes contours of an object and fuses narrow breaks). Other advanced operations that are a combination of these basic ones include hit-or-miss transform, morphological gradient, watershed algorithm, and others. Although normally applicable to binary images, morphological operations have been extended to the greyscale images recently.

Morphological operations are useful as operations similar to image filters to remove noisy details, enhance specific image details, close boundaries, and many others. They usually are used either in a preprocessing step or in a sequence of operations to extract shape-based feature information for further analysis, such as image segmentation and others [55].

2.3.2 Image Segmentation

Image segmentation is an important class of algorithms in medical image computing methods that is crucial towards accurately localising RoI from the images. It is often a critical step to extract feature information from disease locations or anatomical target structures, either to better visualise it or to extract detailed, quantifiable information for further processing. The overall objective of an image segmentation algorithm requires an accurate delineation of the desired target region/structure from its surroundings in the image.

Multitudes of approaches exist to perform image segmentation that may or may not entirely involve the end-user. It may be a manual process in which the user locates the target region by manually delineating its area with visualisation assistance or a semi-automated approach in which the user provides an initial input, and the computer algorithm builds upon it to segment the object. It may comprise an interactive, iterative algorithm involving continuous interaction between user and computer till the segmentation is satisfactory; or an automated computer algorithm segmenting the target structure in a fully automated manner. Whatever the means, the end-goal for an image segmentation algorithm is to delineate the target structure as accurately as possible. However, fully automated algorithms are generally the most desired ones for a variety of reasons such as the ability to minimise the inter- and intra-user variance, generating a highly accurate segmentation in a shorter time, segmenting large quantities of data without any fatigue generally associated with manual observers.

Segmentation algorithms generally operate in two broad modes: segment the target structure by determining its boundary outline or segment the target structure by identifying each pixel (2D)/voxel (3D) in the image belonging to the target object. There are algorithms, however, that use a combination of the two objectives and improve the overall segmentation accuracy by using both complementary information. The former generally use a gradient, edge, shape-based information, whereas the latter mainly use the image intensity-based features [56–59]. The image intensity-based approaches have especially taken significant strides with the evolution of pattern recognition and machine learning algorithms recently. Image segmentation is usually formulated as a labelling problem in the pattern recognition paradigm that involves assigning target structure labels to the pixels/voxels. Further, the explosion of big data and other internet technologies coupled with advancements in general image processing and machine learning has further boosted the state-of-the-art in medical image segmentation. The following sections outline some of the important classes of medical image segmentation algorithms from simple intensity-based thresholding to model-based approaches.

Threshold Segmentation Methods

Threshold-based segmentation converts the medical image into a binary segmentation map by using a threshold based on the domain knowledge, desired target structure, and other image properties. Global thresholding involves individually marking each pixel as a target or background using a single threshold for the entire image. Local adaptive thresholding involves dividing the image into smaller spatial sub-images and then determining the optimal threshold for the local neighbourhood before combing the results to obtain the final segmentation map. Multiple threshold intensity windows can be used to segment multiple target points in the image [60,61]. Thresholding is sometimes followed by connected component analysis to determine the connected pixels and determine the final target structure. Threshold selection is often a critical step that determines the final accuracy segmentation. In addition to using predefined domain knowledge about the target, image histogram and its related properties are often used to determine a suitable threshold, using threshold selection methods such as Otsu's [62]. Thresholding usually does not produce highly accurate results and is often used as a starting point for more sophisticated segmentation algorithms [63].

Region-growing segmentation methods are another group of segmentation algorithms that build on the thresholding methods. The central premise is that pixels in a neighbourhood have similar values and properties. A region membership property can, therefore, be defined using which the local neighbourhoods can be progressively combined to construct a segmentation map. Starting seed points need to be defined in the form of pixels along with a membership property and convergence criteria. The seed points are progressively grown to aggregate neighbourhoods based on the region membership property until the convergence criteria are satisfied. Seed points may be determined using either an automated algorithm or a user input in the form of manual intervention. The region membership property (based on pixel intensity, colour, derived feature, texture, and others) is a crucial aspect of the region-growing methods that determine the process of aggregating the regions and usually affects

the accuracy of the final segmentation map. Region growing methods are simple but effective segmentation methods that work well for images with well-separated entities in which membership property can be clearly defined. However, their computationally intensive process that might even require user intervention and prioritising local information over the global image context often limits their success in complex segmentation tasks [64, 65].

Model based image segmentation

Segmentation of target anatomical structure by a human observer may be time-consuming, inconsistent, and inefficient. However, the central premise used by the human observer is very useful. Observers are almost always experienced practitioners with previous domain knowledge of imaging modality and 'a visual model' about the location, size, and shape of the target structure. Clinicians use this visual model, and a priori information to segment the target structure from its surroundings. Model-based segmentation methods translate this domain knowledge and incorporate the prior visual model information to develop accurate and efficient computer algorithms for image segmentation. Model-based segmentation methods work mostly in two stages: (i) incorporate a prior domain knowledge about the target structure through a user intervention or prior automated algorithms, and (ii) build upon the initial information in (i) to optimise iteratively until an accurate segmentation is obtained agreeing with a predefined objective. The prior knowledge is generally in the form of a contour drawn by the user with a computer mouse, and the second step involves optimisation of the contour until a predefined objective is satisfied. The prior information might include image reconstruction filter used, contrast agent administered, the position of the patient during the scanning. Such information usually are obtained using the DICOM header associated with the imaging data [66, 67].

Active contour models: Active contour models often called as snakes, are a class of deformable model-based segmentation methods that iteratively deform an initial

contour, supplied generally by the user or through an automated algorithm until an energy functional is minimised. The final contour, a closed curve, usually is at least a local minima that best represents the boundary outline of the target object. Energy functional is a combination of internal energy that smoothes the deformation curve and external image energy that resists the internal energy and captures the image features. Active contour methods can be extended to volumetric images by applying active contour models on individual slices and then connecting the contours to get the 3D shape. It can also be applied by using the final contours in a given slice as initialisation for subsequent slices. Active contour methods are in general sensitive to the initial user contour and do not work well when there is no enclosing boundary outline to the target object. Further, they are computationally intensive and in some cases require frequent user intervention [68, 69].

Active shape models: The active contour models start with an initial user-defined contour. The final segmented result, however, is not dependent on any other prior knowledge of the shape of the target structure. Active shape models build on active contour models and explicitly incorporate the target shape information. It may be done by manually locating distinct boundary coordinate points, known as landmark points, along the object outline in a set of training examples. Point distribution models are then employed to build a mean geometric shape and capture main modes of shape variations. It may be done by using a statistical tool like a principal component analysis to capture the most active eigenvectors. The approach generally works similar to active contour models by iteratively optimising an initial guess to the final target structure. A significant difference, however, lies in the explicit shape constraint term that forces the model to look for only the trained shape with limited variations. It is an important advantage when segmenting well defined anatomical structures with clear prior knowledge. It, however, is a limitation if the prior shape information is not clearly defined and object shapes are flexible [70–72].

Active appearance models: Active appearance models extend the active shape models by explicitly incorporating target shape information and intensity variation

within the target structure. Besides the landmarks along the boundary, an intensity vector is constructed from the target structure using the training set of examples. Principal component analysis projections are then carried out to capture main modes and variations in shape information and intensity vector. The subsequent steps and the iterative process are similar to those in active shape models [73].

Level set methods: The previously mentioned active contour models and model-based image segmentation approaches view the curves as a collection of one-dimensional coordinate points and iteratively develop the curve until it is optimised. One-dimensional curve representation, however, makes segmentation of split objects difficult, makes the process vulnerable to noisy outliers, amongst several other limitations. Level set methods are used to overcome some of these limitations by representing the curves using level sets. A level set is defined as a set of points that have the same function value. Rather than representing the curve as one-dimensional points that evolve over a period of time, the curves are embedded in a high-dimensional space, as a zero level set, and the level sets are modified overtime at fixed coordinates. Also, if in case of any split objects and other scenarios, the curve changes its topology, level sets remain a valid function making the segmentation robust [74, 75].

Segmentation using Graphs

Graph-based image segmentation algorithms are an important class of methods that model the image properties using graphs. Graph-based methods model the image as a graph $G = (V, E)$. The image pixels/voxels are represented by the vertices of the graph V, and the edges E represent the cost of connecting the two pixels. Edges with high cost indicate the corresponding pixels belong to different regions of interest, and low-cost edges indicate the pixels belong to a homogeneous region of interest. The objective, therefore, is formulated as a graph minimisation problem and is generally effectively solved by employing a graph cut method. The graph cut method involves finding partition/cuts of a graph such that the overall cost is minimised. Discrete

optimisation algorithms are usually employed to solve this minimisation problem. Some of the recent works incorporate prior information about target structures and integrate model-based image segmentation methods with graph techniques. Graph search techniques are another class of graph image segmentation techniques that efficiently model and segment two-dimensional surfaces from volumetric image data. Graph search techniques are a sophisticated extension to graph cut methods that formulate surface detection problem as determining a minimum closed set in a weighted graph [76, 77].

Random walker image segmentation method is another popular graph approach that first needs initialisation in the form of starting seeds/pixels for different regions of interest. A random walker is assumed to release from the unlabelled pixels, and the probabilities of random walker reaching the initialised pixels are computed. This probability is generally computed by solving a sparse system of a linear equation that is both positive definite and symmetric [78]. Some of the other methods include employing Markov Random Fields (MRF) [79], unsupervised approaches using a combination of MRF and Expectation-Maximisation (EM) algorithm [80], watershed image segmentation [53].

Machine learning methods

The previously described conventional medical image analysis methods are a combination of techniques, some of which use a priori information, and some learn to optimise indirect energy functional. Machine learning extended their performance boundaries by building methods that learn effectively from data and get benefited from increases in data size to better generalise the ability of the computation algorithm to diversified data. A machine learning algorithm, in general, involves learning from patterns present in the data, for a particular task, to optimise predefined performance criteria. The algorithm might learn from both the data as well as the expected output, a process known as a supervised learning paradigm. A standard pipeline in a supervised machine learning usually involves two stages:

feature extraction - extracting manually engineered distinctive features from the data and classification - training a classifier from the features to construct decision boundaries. Unsupervised learning, on the other hand, learns entirely from the data without any reference outputs. It usually involves optimisation of predefined criteria that reflects the overall learning objective. Other learning approaches include reinforcement learning that aims to optimise cumulative rewards, feature learning that attempts to directly learn features from the data, instead of using manually engineered features.

Machine learning approaches significantly improved upon the conventional computing methods to produce better performing, more generalisable image analysis techniques efficiently applicable to large amounts of diversified data from different medical image modalities. The work in [81] investigates several machine learning algorithms for automated classification of clustered microcalcifications towards achieving an accurate diagnosis of breast cancer on mammograms. Support vector machine (SVM), kernel Fisher discriminant (KFD), relevance vector machine (RVM), ensemble classifiers were some of the machine learning classifiers considered. A marginal space learning framework is proposed in [82] for localising anatomical structures from 2D/3D medical images in a series of marginal spaces with increasing dimensionality, by sequentially performing object position estimation, position-orientation estimation, and position-orientation-scale estimation. Marginal space learning is extended for automated segmentation of 3D cardiac CT volumes in [83], and automated liver segmentation in [84]. Random forest classifiers, their training process, and their ability to quantify the importance of discriminative features in classification are analysed in detail in [85] to improve the performances of medical image segmentation. Some of the other noteworthy classifiers include logistic regression, k-nearest neighbour classifier, neural network classifiers. A significant limitation of the machine learning methods is that their performances are often constrained by the manually engineered features that may not be optimised for the particular data.

2.4 Summary

This chapter gave a brief outline of the importance of medical imaging in a diagnostic setting and the necessity of efficient computing methods to perform their computerised assessment. Different imaging modalities have evolved over the years that capture different kinds of information from the human body, with focus ranging from a high spatial resolution to high temporal resolution. Accordingly, multiple computing methods have been developed that operate on different image properties to perform different tasks. From image enhancement techniques for improving image quality to image segmentation methods in an accurate target localisation, medical image computing methods have drastically progressed over the years assisted ably by the accelerated growth of data aggregation.

Chapter 3

Convolutional Neural Networks and their Applications in Medical Image Analysis

This chapter presents an overview of convolutional neural networks and their advances in image analysis tasks such as image recognition, image segmentation, and others. It then summarises the challenges in applying neural networks to medical image analysis tasks and documents some of the works that overcome a few of these limitations.

3.1 Introduction

Artificial neural networks (ANNs) are a class of computing algorithms, loosely inspired by the biological neural networks in the brain [86], that can be modelled to recognise patterns in data. They can be built to perform specific tasks on different types of data from speech to natural languages to images. ANNs represent a network of nodes, called artificial neurons, interconnected by trainable weights that can be trained to achieve optimum performance for the desired task. The overall objective of an ANN is to compute a function f, consisting of a set of trainable weights W, that is an optimum mapping of input x to output y for a given task T. Several approaches exist to train the weights depending on the availability of the prior ground truths, from the supervised learning process to unsupervised learning.

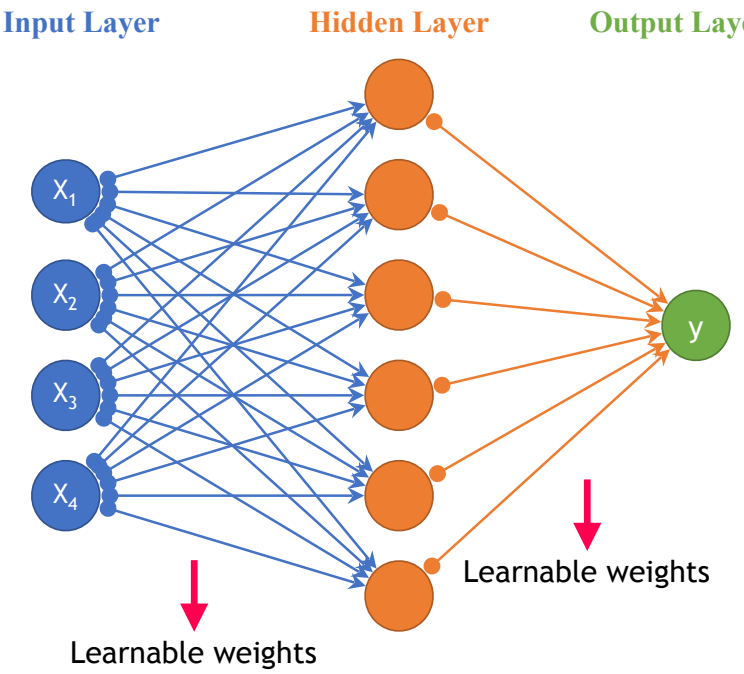


Figure 3.1: A multilayer perceptron consisting of an input layer with four nodes, a hidden layer with six nodes, and an output node. It can be noted that all the nodes of a layer are connected to all the nodes of the subsequent layer.

ANNs with only feedforward connections between network nodes, called feedforward neural networks, are one of the primary and commonly used neural network architectures in visual recognition tasks. Multilayer perceptrons (MLP), shown in Figure 3.1, form an important class of feedforward neural networks. In general, MLPs consist of three layers of neurons: the input layer, a hidden layer, and an output layer. The input layer connects the input nodes (input data) to hidden layers using learnable weight connections, which then connect to output nodes. MLPs are typically shallow networks consisting of one or very few hidden layers. Further, they connect each node of a layer to every other node of the next layer. Therefore, they fail to scale with increases in the data dimensions, increased depth of the networks, and, more significantly, fail to capture local data patterns. Convolutional neural networks improve upon the MLPs by introducing local neuron connectivity. Further, the weights of locally connected neurons are shared across the entire spatial extent of data. CNNs mostly operate on image data, and weight sharing of the

neurons essentially acts as convolution operation between a filter kernel of neuron weights and the input image, leading to the name convolutional layer. Weight sharing allows replication of learned features throughout the data extent, in addition to regularising the network by reducing learnable parameters.

The fundamental ideas behind the CNNs were developed as early as the 1980s; nonetheless, the traction was not gained until the early 2010s [87–91]. CNNs have made rapid progress in recent years in the field of image recognition and computer vision by significantly pushing the performance boundaries. The advancements were primarily due to two crucial reasons: (i) exponential data growth and an era of connectivity leading to aggregation of large-scale datasets, and (ii) advancements in computing hardware to efficiently train deeper and complex CNNs. The initial advances in CNN are primarily for RGB colour images (RGB - Red, Green, Blue colour space), which can be easily acquired and distributed on a large scale. It is, however, not straightforward to similarly acquire extensive amounts of medical imaging data. Privacy, affordability, ease of access, and harmful radiation effects of some imaging modalities form significant hurdles in a fast and smooth data collection process. Despite these challenges, there have been several significant advances in medical image processing, thanks to engineering strategies that include data augmentation to increase the data size, employing transfer learning to transfer representations learned on a different task to the desired task, combining data of multiple imaging modalities, and several others.

The following sections discuss underlying CNN concepts, prominent CNN architectures in image analysis tasks, the applicability of CNNs to medical images and associated challenges.

3.2 Convolutional Neural Networks (CNN)

3.2.1 Basics

CNNs are a class of neural network architectures similar to MLPs that are designed to operate on image data. The assumption that the input data is image constitutes a primary constraint in the design of CNN architectures and is used to limit the connection of a node in a higher layer to a small set of nodes in a local neighbourhood in the lower layer. The learnable weights connecting the nodes are optimised to extract basic low-level image features from the neighbourhood. These low-level features are subsequently combined in the following layers to obtain high-level, complex, non-linear features. It leads to hierarchical features from the image that progressively learn high-level semantic concepts at the deeper layers.

A generic CNN architecture consists of an input layer, multiple hidden layers, and an output layer. The primary components in constructing a CNN architecture include a convolutional layer, a non-linear activation function, a pooling layer, a fully connected layer, a loss function reflecting the task to be performed, and an optimisation process to determine the weights optimised for the defined loss function.

Convolutional layer. The convolutional layer is the essential component of a CNN architecture, which extracts the local image features from raw image data. It typically consists of multiple filter kernels, which have fixed spatial dimensions and operate on a local neighbourhood in the image. The filter-size, known as the receptive field, defines the spatial extent of the neighbourhood on which the filter operates. The learnable weight parameters of the filter kernel are replicated across the entire spatial extent of the image, based on the assumption that similar feature information exists throughout. Besides reducing learnable parameters, this parameter sharing enables translation invariance by extracting the same set of features independent of their location. The outputs by convolutional layer filters are usually called activation maps/feature maps. Different filters in a convolutional layer extract different types of feature maps from the image, and the number of filter-kernels in

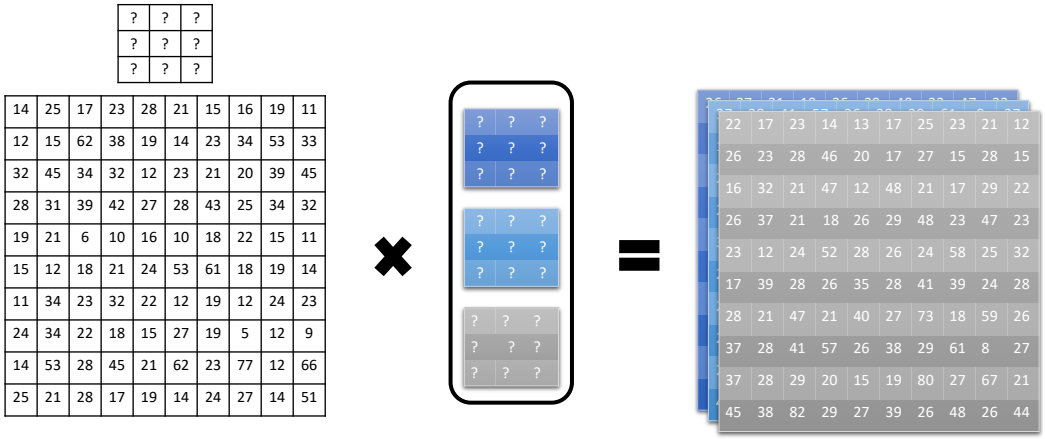


Figure 3.2: A convolutional layer consisting of 3×3 filter kernels. The layer consists of three filters kernels, leading to three output feature maps. Therefore, the layer is said to have three output channels.

the layer defines the depth of the feature map. Some of the essential hyperparameters in constructing a convolutional layer include (i) receptive field (K), (ii) filter stride (S), (iii) number of filters (D), and (iv) padding (P). Padding appends input, usually with zeros, to prevent losing the boundary pixels during convolution. The size of an output feature map is given by $(W - F + 2P)/S + 1$, where W represents the size of the input feature map/image. Figure 3.2 shows an example convolutional layer consisting of 3×3 filter kernels. The depth of the layer is 3, leading to 3 output feature maps.

Activation function. convolution is essentially a matrix product that extracts linear features from the images. If the feature maps directly connect to subsequent higher layers, the entire learning process of CNN architecture would be linear, and the layers may not be able to learn higher hierarchies of features to construct decision boundaries in high-dimensional spaces to solve pattern recognition problems. Therefore, non-linear activation functions are incorporated into the CNN architectures that apply elementwise non-linearity to the output feature maps of convolutional filters, before connecting them to higher layers. The non-linear operations, sandwiched between convolutional layers, allow CNN to gradually learn high-level, complex non-linear feature representations from the data and construct effective

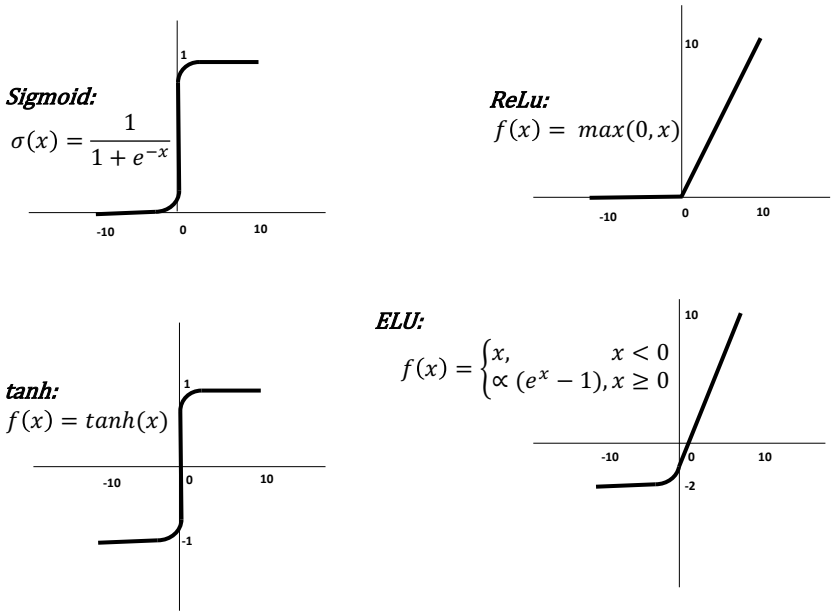


Figure 3.3: Commonly used non-linear activation functions. The functions operate elementwise on output feature maps of convolutional layers to produce non-linear feature maps.

decision boundaries. Gradient-based optimisation techniques are usually employed to train the weights of a CNN; therefore, the activation functions need to be (i) monotonic (monotonically increasing/decreasing), and (ii) differentiable. Some of the commonly used non-linear activation functions are shown in Figure 3.3.

Pooling layer. pooling is another widely used operation in standard CNN architectures, between any two sequential convolutional layers. Pooling, alternatively referred to as downsampling, progressively reduces the spatial dimensions of a feature map, thereby reducing the number of learnable parameters. Besides reducing the computational complexity and controlling the overfitting through the reduction of learnable parameters, pooling introduces an element of spatial invariance since it discards the information about exact spatial locations of features through downsampling. Pooling is a familiar presence in many high-performing CNN architectures and operates in a predefined spatial window in a feature map. Different forms of pooling include (i) max-pooling that selects a feature with maximum intensity in the window, (ii) min-pooling that selects a feature with minimum intensity in the

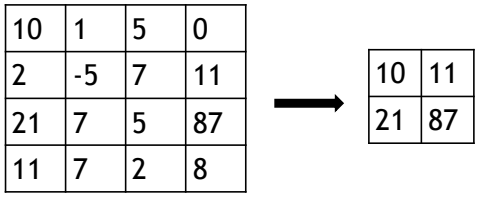


Figure 3.4: Max pooling operation. It operates in a 2×2 spatial window of the feature map and selects the feature with maximum intensity.

window, (iii) average-pooling that averages the feature values in the window and several others. Max-pooling is generally the preferred choice as it has demonstrated a significant performance boost in many practical applications. Some of the essential hyperparameters in constructing a pooling layer include (i) pooling window w, and (ii) pooling stride s. A 2×2 max pooling operation is shown in Figure 3.4.

Fully connected layer. CNN architectures mainly consist of interconnected, convolutional layers followed by non-linear activation functions to progressively obtain high-level feature information. Pooling layers are used in-between select convolutional layers to control overfitting and incorporate spatial invariance. These sequence of layers, however, extract feature information specific to spatial neighbourhoods, and image-level feature information is essential to ultimately determine the image content and perform the desired visual recognition task. Therefore, fully connected layers are used usually towards the final stages of a CNN architecture to combine all the lower level feature information extracted by the sequence of convolutional layers. Fully connected layers are a variation of MLPs that connect all the nodes in a layer to all the other nodes in the previous layer. They do not incorporate any non-linear transforms and act as a matrix dot-product that linearly combines all the lower-layer features to obtain global, image-level feature information.

The above-mentioned layer components are integral to building a standard CNN architecture for a visual recognition task. The next important step involves training CNN to determine the learnable weights optimised for the desired task, for the given data distribution. In a supervised learning framework, the essential components for the training process include (i) loss function that reflects the task, and (ii)

optimisation algorithm.

Loss function. Training a CNN requires an objective function, which represents the target task, and that needs to be optimised to achieve the desired task performance. The objective is usually formulated as a loss function that reflects some form of cost between the actual CNN output and expected CNN output. The loss function is a crucial CNN component, which decides the final task performance. Many facets of the task need to be condensed into the loss function so that optimal task performance is eventually achieved through its minimisation. Gradient-based optimisation algorithms are used to minimise the loss functions; therefore, they need to be differentiable. Varied loss functions exist designed depending on whether the final objective is a classification formulation, regression formulation, and many other criteria [92].

Cross entropy loss is a commonly used loss function in classification problems computed as

$$CE(Y,p) = -(y\log p + (1-y)\log(1-p)) \tag{3.1}$$

where Y represents the actual output, p represents the predicted probability for the two-class classification problem. It measures the probability output of CNN. The loss reduces when the predicted probability of the model is closer to the ground truth classification label. Hinge loss is another loss function used in classification problems, computed as:

$$L(y) = max(0, 1-\hat{y}y) \tag{3.2}$$

where y and $\hat{y}$ represent the actual and predicted output, respectively. Some of the commonly used loss functions in regression formulations include the L_1, L_2 losses that compute the L_1, L_2 norm of the errors between the actual and predicted outputs. Several other losses exist, and their use primarily depends on the desired task and the performance that needs to be optimised.

Optimisation algorithm A standard way to optimise the CNN weights proceeds with a combination of backpropagation framework and gradient-based learning algo-

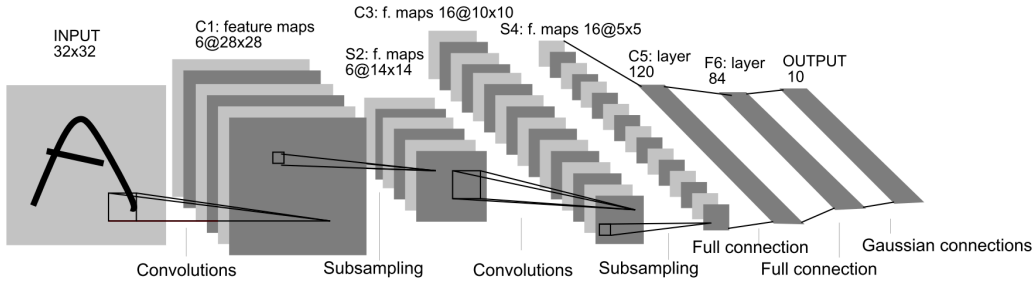


Figure 3.5: A five layer CNN architecture trained for handwritten digit recognition. It consists of three convolutional layers and two fully connected layers, in addition to two pooling layers. Figure source: [2]

rithms that involve computation of the gradient of the loss function with respect to the weights of the CNN and backpropagating the error through preceding layers [92]. The efficient method of backpropagation enables usage of the gradient-based method to iteratively minimise the loss function. Stochastic gradient descent [93], Momentum [94], Adam [95] are some of the commonly used gradient-based optimisation algorithms. There have been several foundational works in the conception and design of basic principles behind CNNs [87–89] by effective usage of the above-mentioned components. The work in [2], however, made a significant impact, which documented the importance of learning features from data, elucidated the advantages of gradient-based learning techniques, and comprehensively demonstrated the applicability and superior performance of neural network architectures in visual recognition tasks (e.g. handwritten recognition). The CNN architecture used in [2] is shown in Figure 3.5. Highlights of the architecture include:

- The primary objective in [2] is to develop a pattern recognition model to recognise and classify the handwritten digit present in the input image to one of the ten classes from $0-9$. CNN architecture is one of the approaches employed, and the data set used is MNIST [2]. The input images are of size 28×28 and the output is 10 dimensional vector indicating a probability for the presence of each digit.

- The architecture consists of eight layers in total: one input layer, three convo-

lutional layers, two pooling layers, and two fully connected layers.

- The kernel sizes and depth of the convolutional layers, and the number of nodes in a fully connected layer are indicated above the respective layers.
- It performs sub-sampling in a non-overlapping 2×2 window by adding the four values, multiplying them by a trainable bias, and adding a trainable bias to get the sub-sampled value for the window.
- Two fully connected layers are included as final layers in the architecture. The first FC layer has 84 nodes combining the lower-layer feature information. The second FC layer has 10 nodes representing the 10 classes in the image dataset.
- The minimum squared error (MSE) is used as a loss function to train and learn the learnable parameters of the CNN architecture in a supervised paradigm using backpropagation and gradient-based optimisation.

The recent advances in convolutional neural networks and their rapid growth have been possible mainly due to the exponential growth of data. Extensive collections of RGB images for visual recognition tasks have enabled the evolution of complex CNN architectures with sophisticated learning techniques to advance the performance boundaries significantly. Image classification is one image analysis task, where CNN architectures have made considerable progress and culminated in state-of-the-art results. It has subsequently influenced the other image analysis tasks by formulating them in an image classification framework. Therefore, the following section first discusses some of the significant developments in CNN architectures for image classification to outline the significant developments in CNN architectures.

3.3 Image classification Using CNN

The following CNN architectures have been developed over several years to address the image classification challenge on ImageNet LSVRC database, which consists of about 1.2 million high-resolution images, having 1000 different classes. The input

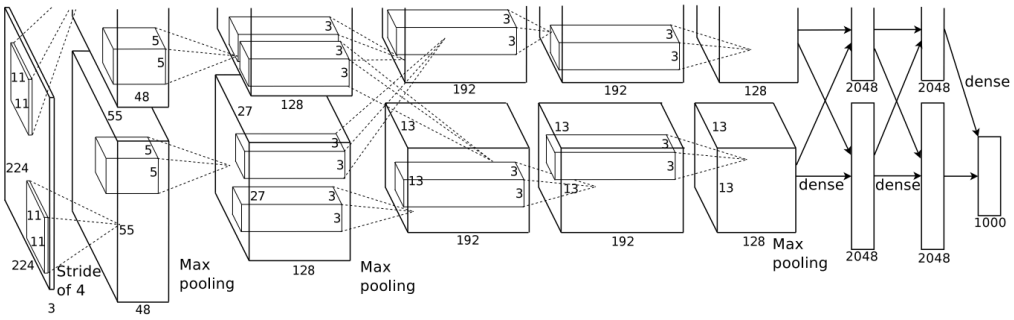


Figure 3.6: An eight-layer CNN architecture trained for large-scale image classification. It consists of five convolutional layers and three fully connected layers, in addition to three pooling layers. Figure source: [3]

images are of size $224 \times 224 \times 3$, and the outputs usually are 1000 dimensional vectors indicating a probability for each class [96].

AlexNet

The work in [3] is a significant milestone in the reinvention of CNNs that has led to the development of deep CNN architectures and their ubiquitous presence in a multitude of machine learning applications. Some of the highlights of the works include:

- The objective is to perform image classification on the ILSVRC database, and the architecture consists of eight layers in total: five convolutional layers, and three fully connected layers (60 million parameters and 650,000 neurons). The kernel sizes and depth of the convolutional layers, and the number of nodes in a fully connected layer are indicated above the respective layers in Figure 3.6.

- Three pooling layers are used, and max-pooling is employed on overlapping windows of 3×3 at a stride of 2.

- ReLU non-linear activation function is used.

- Data augmentation is used to artificially boost the number of images and dropout[] is used as a regularisation technique.

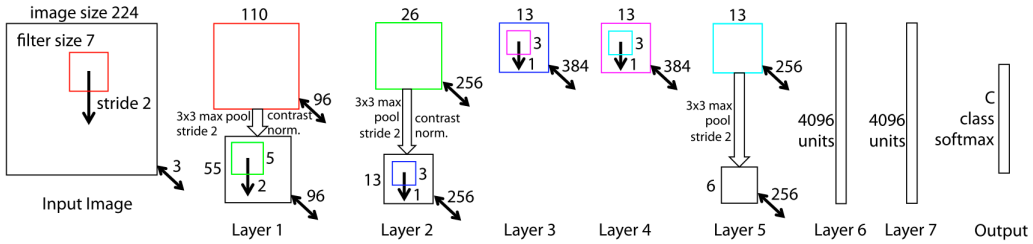


Figure 3.7: An eight layer CNN architecture trained for large-scale image classification. It consists of five convolutional layers and three fully connected layers. A major contribution of this work is employing novel techniques to visualise the features maps of intermediate CNN layers to understand their learning process better and improve the hyperparameters. Figure source: [4]

- Binary cross-entropy [92] is used as a loss function to train the network parameters. The network is trained using stochastic gradient descent [97] with a batch size of 128 examples, the momentum of 0.9, and a weight decay of 0.0005.

ZFNet

- A major contribution of this work is the development of novel visualisation techniques to visualise the feature maps of intermediate and final layers in a CNN architecture to understand their learning process, advantages, and limitations. Deconvolution, unpooling, and rectified non-linearity are used to transform the feature maps for visualisation purposes.

- The visualisation techniques are employed to analyse the limitations in [3] and modify the architecture accordingly. It consists of 8 layers in total: 5 convolutional layers, and 3 fully-connected layers. The kernel sizes, filter strides, and depth of the convolutional layers, and the number of nodes in a fully connected layer are indicated besides the respective layers in Figure 3.7.

- Generalisability of CNN architecture trained on Imagenet database to other imaging datasets are demonstrated. Ablation studies are conducted to analyse the contributions of different components of CNN.

- Pooling operations, non-linear activation function, data augmentation techniques, regularisation techniques, and training methods are the same as in [3].

Very Deep networks

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 3.8: Very deep convolutional neural networks. An important contribution of this work is to demonstrate the significant performance enhancements with increased network depths. Figure source: [5]

- A notable contribution of this work [5] is to demonstrate the correlation between depth of CNN architecture(number of layers) and image classification performance. It develops network with increased depth(deeper networks), using convolutional layers with small receptive fields, to significantly enhance the classification performance.

- The two important variations of the architectures proposed in this work include a 16 layer network (VGG16) and a 19 layer network (VGG19). The number of layers in the network and other layer related hyperparameters are shown in Figure 3.8.

- A multinomial logistic regression is used as the objective function of the network and is optimised using mini-batch gradient descent with momentum. The training was regularised using weight decay and dropout regularisation for the first two fully-connected layers.

Inception Architectures

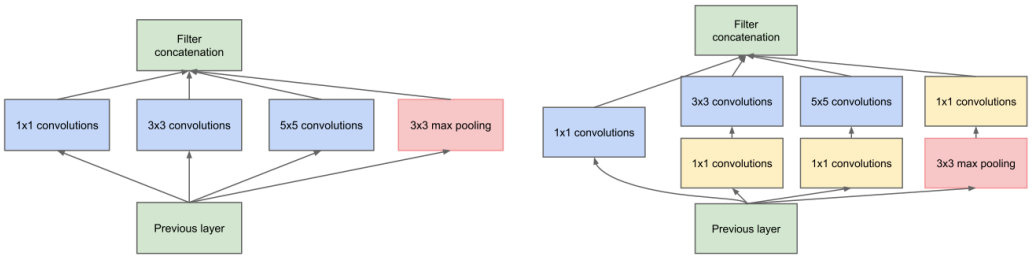


Figure 3.9: A basic inception module. It employs convolution filters with different kernel sizes to extract feature maps at different resolutions. An important inference of this work is that the width of a CNN architecture contributes significantly to an enhanced final performance. Figure source: [6]

- The work in [6] was carried concurrently with the VGG [5] architecture to demonstrate the ability of deeper networks to enhance the image classification performance. A deeper, wider, and more complex CNN architecture was developed; however, without increasing the number of parameters compared to previous works.

- Building on the works in [98], "inception modules" are developed, shown in Figure 3.9, that are the primary building blocks of the CNN architecture. It is based on approximating an optimal local sparse structure in a CNN architecture using available dense components. Convolutions with very small receptive fields(1×1) cover concentrated features, and with increasing receptive fields extract spatially spread-out features. Dimension reductions and projections are used to embed the features into lower-dimensional space and reduce computational complexity.

- The inception modules usually are stacked to build inception CNN architec-

tures. They allow the network width and depth to be increased without increasing the computation complexity, allow the features to be extracted at multiple scales, and therefore, significantly boost the performance of CNN architectures in image classification and object localisation tasks.

The work in [99] modifies the inception architecture to develop new design principles useful to scale the CNN architecture to deeper and complex networks efficiently. Some ideas include: balancing width and depth in a network can contribute to optimum performance; lower-dimensional spaces are useful to spatially combine features without losing valuable information; higher-dimensional representations are easier to process locally within a network.

Residual Learning Framework

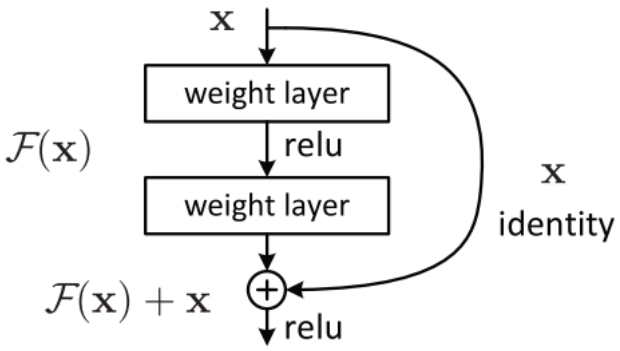


Figure 3.10: A generic residual learning block. The residual blocks learn referenced mappings and improve the gradient flow between layers. They can be effectively deployed to build significantly deeper CNN architectures that enhance, instead of degrading, the final performance. Figure source: [7]

- As demonstrated in previous works, deeper and wider CNN architectures have significantly improved image classification performance. The increase in depth, however, makes the network training difficult. The addition of extra layers to an already deep model saturates the accuracy, increases the training error, and eventually leads to performance degradation. The work in [7] attempts to address the degradation problem in deeper networks by developing a residual learning framework.

- It is not straightforward to determine the mapping between input and output of a convolutional block in a deep CNN architecture, as the mappings are unreferenced. The residual learning framework in [7] addressed this issue by developing residual blocks and explicitly constraining them to learn a residual mapping between input and output. The residual blocks are realised by using convolution blocks with 'shortcut connections', as shown in Figure 3.10. The identity shortcut connections, as shown, do not add any extra parameters and, therefore, do not make any changes to the computational complexity.

- The residual learning model was developed to create significantly deeper networks of up to 152 layers that improve the performance of image classification and object detection and do not suffer from performance degradation issues associated with regular CNNs. Besides, the residual network could still be trained in a backpropagation framework similar to standard CNNs using gradient-based optimisation procedures.

- An extension of the work in [100] further delves on the residual learning framework and conducts a detailed study and series of ablation experiments to document the importance of identity shortcut skip connections and identity shortcut additions in the propagation of the signal through the network. A 1000 layer CNN architecture is developed to demonstrate the significance of residual learning blocks in developing very deep architectures.

Inception architectures are combined with the residual learning framework in [101] to develop deeper networks efficiently, and it also helps accelerate the network training process.

3.4 Object Localisation and Segmentation

3.4.1 Object Localisation

The initial deep neural net architectures were developed primarily for image classification - to classify an image based on its content into one of the predefined categories. Within a single image, there can be multiple instances of a single type of object, and localisations of object positions can be challenging and computationally intensive. Object localisation using deep neural networks is explored in [102] in a regression-based formulation. A standard CNN architecture, similar to [3], is designed to predict rectangle coordinates bounding an object instance. The classification layer in the CNN is replaced with a regression layer, and the network is trained using a L_2 error for predicting the ground-truth mask. In combination with refinement techniques, multi-scale box inferences are used to produce precise bounding box locations of object instances. In [103], an integrated multi-scale, sliding-window system is proposed to perform image classification, object localisation, and object detection. Localisation is considered an extension to image classification involving bounding box localisation of a primary object in the image. Whereas, detection is a harder problem involving bounding box localisations of all the object instances within the image. By first training the architecture for classification and then replacing the classification layer with the regression layer to produce bounding box coordinates, the three tasks share the same feature extraction architecture. Classification and regression outputs are subsequently combined to generate bounding boxes with confidence scores. Detection scores are improved by combining localisations at multiple scales.

The CNN-based object localisation approach in [8] made a departure from the regression-based formulations discussed previously and adopted a recognition using regions approach [104], known as region CNN (RCNN), to build an object detector using bottom-up region proposals, shown in Figure 3.11. Region proposal algorithms like selective search [105] are used at test time to compute region proposals, which

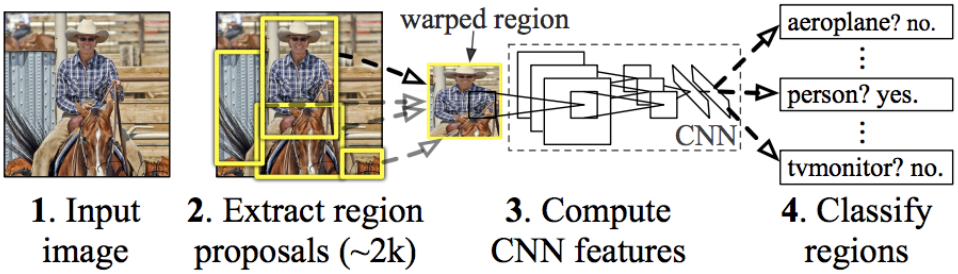


Figure 3.11: CNN operating on regions for object localisation. Region proposal algorithms are employed during test time to generate object proposals, which are passed through trained CNN to extract proposal-specific features to identify the object content of the proposal. Figure source: [8]

are then passed through trained neural nets to extract proposal-specific features. The features subsequently are classified by class-specific linear SVM classifiers to identify the class of the object within the proposal. Pooling in a CNN operates on fixed-size windows leading to varying output sizes depending on the image size. Therefore, all the input images to the network need to be of the same size. Spatial pyramid pooling (SPP) is incorporated into standard CNN architectures in [106] that outputs fixed size pooled outputs and therefore, allows input images of varying sizes. A key difference with RCNN is that features are computed on the entire image only once, instead of passing each region proposal through CNN reducing the computational complexity.

Fast RCNN was proposed [107] by building a CNN architecture that trains in a single end-to-end pipeline framework. The region proposals are input into the network with the image. The CNN processes the whole image before extracting features for the proposals. The features are then fed into sibling fully connected layers performing classification and regression. Faster RCNN was proposed in [108] that removed the usage of region proposal algorithms by developing region proposal networks (RPN). RPN is a fully convolutional network that generates object proposals and objectness scores to be further used by fast RCNN. It is incorporated into the fast RCNN framework to share the layers with fast RCNN, thereby making producing computation-free region proposals. Fast and faster RCNN significantly improved the object detection accuracies and inference times.

3.4.2 Image Segmentation

The machine learning framework formulates image segmentation as a pixel-classification problem that involves classifying image pixels to one of the object classes. CNNs substantially improved image recognition performances, in general, by extracting powerful hierarchical features through a deeper and complex set of convolutional layers. An obvious extension of the applicability of CNNs is to image pixel segmentation using the architectures experimented against image classification and object detection. A multi-scale CNN is explored in [109] to perform scene parsing by labelling all image pixels. After transforming the input image through a scale pyramid, images at multiple scales are processed by a CNN architecture to capture features at different resolutions. The multi-scale features are subsequently post-processed by methods such as conditional random fields to obtain final pixel labels. A feature learning framework is developed in [110] for RGBD images that propose a CNN based detection and segmentation architecture using depth images encoded using a new geocentric embedding. Using the detection output from CNN architecture, a decision forest is trained to perform instance segmentation, and superpixel classification framework is used to perform semantic segmentation. The RCNN framework is exploited in [8] to also perform semantic segmentation by using CPMC for the region proposal and using linear SVMs for final regression.

The work in [111] substantially improved segmentation performance and led to an evolution of a new class of CNN architectures suited for image segmentation. Simple changes were introduced into standard CNN architectures by removing the fully connected layers and directly using the final convolution feature maps to learn semantic segmentation directly from images and ground-truth images, leading to a new class of segmentation architectures known as fully convolutional architectures. The architecture significantly evolved over the previous CNN based segmentation approaches by not using multiple post-processing techniques and multi-stage training pipelines. The standard classification CNN architectures [3, 5, 6] are modified to suit the fully convolution class of segmentation architectures by removing the final

fully connected layers. Fine-tuning the feature representations, learned for image classification yields substantially improved segmentation performance.

Skip connections are further proposed and incorporated into fully convolutional architectures to refine the spatial precisions of the segmented outputs [112]. Skip connections combine low-level coarse information with fine high-level information, significantly refining the segmented output. However, it cannot handle multi-scale image information and can produce segmented output based only on single-scale information, making it difficult in case of objects much larger or smaller than the receptive field. In addition, a simple deconvolution/interpolation is used to construct the segmentation map from a high-level, low-resolution feature map. The work in [113] proposes to overcome these limitations by proposing a multi-layer deconvolutional network through a sequence of deconvolution, unpooling, and non-linear layers. The fully convolutional-deconvolutional architecture significantly improves the semantic/instance segmentation performances. Individual object proposals are segmented through the network and combined to get the instance segmentation output, alleviating the multi-scale issues.

Incorporating multi-scale contextual information in obtaining high-level feature maps, and using them to construct a full-resolution segmentation map to obtain pixel-level predictions are two critical sub-tasks in a dense prediction task. The regular fully convolutional architectures [111] aggregate contextual information using the convolutional and pooling layers. However, in the process, feature resolution is reduced. The work in [114] proposes a multi-scale context aggregator module that uses dilated convolutions to aggregate contextual information without affecting the feature resolution. The dilated convolution can cover multiple exponentially increasing receptive fields, by using different dilation factors, without affecting the feature resolution. It does not change either the number of feature channels in the output or the feature resolution. It, therefore, can be directly plugged into any existing CNN architectures to modify them to be suitable for the segmentation task. A vital inference from the work is that the structure of a segmentation architecture is primarily different

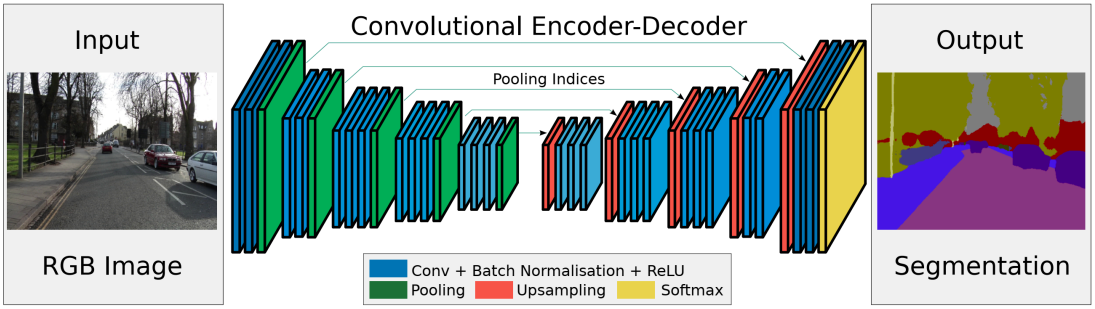


Figure 3.12: A convolutional encoder-decoder framework for image segmentation. The encoder stage extracts semantically rich low-resolution feature maps, which are subsequently processed by the decoder to generate high resolution pixelwise segmentation maps. Figure source: [9]

from that of image classification, and they have to be suitably modified to boost the segmentation performance.

Building on the fully convolutional segmentation architectures, the work [9] proposes an encoder-decoder segmentation framework for the CNNs. Figure 3.12 shows a generic encoder-decoder framework for semantic segmentation. The main contribution involves a novel upsampling scheme to construct the full-resolution segmentation map from the low-resolution feature map extracted using the encoder. The corresponding pooling indices from the max-pooling layers of the encoder are stored and then used to upsample the low-resolution encoder feature maps non-linearly. The upsampled non-linear feature maps are sparse and are transformed through convolutional layers to obtain the dense segmentation maps.

A fully convolutional architecture in conjunction with conditional random fields was proposed in [115] to improve the segmentation accuracies. It employs the fully convolutional versions of the standard CNN architectures [3, 5] by modifying them to reduce the spatial information loss due to pooling layers. It incorporates atrous convolution similar to [114] in the final stages of convolutional architectures to preserve spatial resolution, and include contextual information. Spatial invariance is another essential property of CNNs that enhances the image classification accuracy; however, it affects a precise object localisation. Fully connected conditional ran-

dom fields are integrated into the segmentation architecture to improve the precise boundaries of the segmented objects and enhance segmentation accuracy. In [116], this work is extended by directly incorporating the atrous convolutions into existing classification CNN architectures to extract dense feature maps without increasing the overall network parameters. The atrous convolution is useful to arbitrarily control the receptive field of the convolutional kernel by varying dilation rates. It is further incorporated to develop an atrous spatial pyramid pooling (ASPP) layer that aggregates feature information at multiple contexts. Parallel convolutional filters with multiple dilation rates are used to extract multi-scale features, which are then used to generate a final dense segmentation map.

These works are improved upon in [117] by incorporating image-level features into the atrous spatial pyramid module to capture global image contexts. It also discusses in depth the choice of dilation rates and their efficiency in the extraction of dense and multi-scale features. Besides, the post-processing by CRFs is discarded; however, the segmentation performance is still significantly enhanced. An encoder-decoder architecture is proposed in [118] as an extension to the previous work, that employs dense CNN with atrous convolution as an encoder to extract powerful features and incorporates an atrous spatial pyramid pooling module for aggregating multi-scale feature information. A decoder network is further developed to obtain a full-resolution precise segmentation map from the low-resolution feature maps. Also, depthwise separable convolution is employed in ASPP module and decoder network to reduce the computational complexity without compromising the segmentation accuracy significantly.

3.5 Applying CNNs to Medical Images

3.5.1 Domain Specific Challenges and Adaptations

The convolution neural networks have significantly pushed the performance boundaries in computer vision and image processing ranging from image classification to scene parsing and object tracking and many related tasks. It was, therefore, a natural extension to employ the deep neural nets in medical image analysis. CNNs have been previously used in medical image analysis [119–122] before the recent advent of deep neural nets; however, it was computationally intensive, and performance was not satisfactory. A primary contributing factor in advancing the performances of deep neural nets has been the availability of extensive amounts of data. An exponential data growth in the era of connectivity has enabled the development of complex deep neural nets by training them with diversified data to help model different image properties. It is, however, not straightforward to acquire large amounts of medical image data quickly because of: (a) privacy issues, (b) nature of certain diseases, and (c) difficulty in obtaining expert annotations from qualified radiologists in large volumes. Nevertheless, affordability, accessibility, and advancements in medical imaging technology have steadily contributed to the growth of medical imaging for diagnostic purposes. Sophisticated approaches to address data shortage issues have complemented the steady growth in medical imaging significantly to enhance the performance of several medical imaging tasks.

The transferability of learned CNN representations is a vital aspect explored in the previous works. In the case of limited data applications, deep neural nets are trained on auxiliary tasks, and the trained knowledge is then transferred to be used for the desired task. There have been many successful approaches that exploit the limited medical image data efficiently by (i) directly using models trained on natural images; (ii) by finetuning models for the desired domain; (iii) by using a simpler architecture that would require less amount of data; and (iv) by employing simple and effective data augmentation techniques. For instance, select convolutional layers or final fully

connected layers of a network are replaced in a neural net trained on an auxiliary task, and the modified network is trained using the medical data for the desired task. It has yielded promising results on several medical imaging tasks, in which it is challenging to aggregate a large collection of data. CNNs learn hierarchical feature representations by extracting low-level generic image features in lower-layers and high-level semantic features specific to the application in higher layers. This property is incorporated to finetune a pretrained model in a layerwise manner to enhance the performance of the medical imaging task.

The CNN applications in medical image analysis have spanned multiple modalities, which include CT, MRI, ultrasonography, and others. Each of these modalities has its constraints. MRI produces high-resolution images; however, it is not suitable for moving structures and requires many preprocessing steps. CT is widely available, but it has radiation effects limiting its frequent usage. Ultrasonography can be used without any radiation effects; however, low-spatial resolution and high temporal resolution make the computation a challenge. CNNs have contributed immensely to several medical image analysis tasks in recent years, such as disease classification, disease localisation, anatomy segmentation, medical image registration, group analysis of medical image and radiology reports, and others [123–126]. Some of the important technical advancements and applications are detailed below.

A five-layer CNN architecture is developed in [127] to perform lung pattern classification for interstitial lung diseases (ILDs) from CT images. The work in [128] developed a computer-aided detection system to detect pulmonary nodules from lung CT. Multi-stream 2D CNNs are trained on patches centred on the nodules extracted from different orthogonal views, which are then combined to produce final nodule detection. In [129], a two-stage CNN-based object detection architecture is developed and experimented on the detection of sclerotic metastases, detection of lymph nodes, and detection of colonic polyps from CT images. The first stage achieves a high-sensitivity object detection, and the outputs from the first stage are transformed through random translation and scale transformations and aggregated

to train the second stage network, which significantly reduces the false positives while retaining the sensitivity.

Multi-instance learning is integrated into deep neural nets in [130] to develop a weakly supervised anatomy localisation framework that learns discriminative and non-discriminative features of an image contributing to its classification using weak image-level labels. A single-stage CNN is trained to detect the presence of desired anatomical structure in the three orthogonal views of a medical image volume (axial, sagittal, coronal) in [131]. The detection results from 2D views are then combined to localise the target structure in the volumetric data through a 3D bounding box. Additionally, a spatial pyramid pooling layer is incorporated into this architecture to analyse medical images of different dimensions. A Marginal space deep learning framework is introduced in [132] that uses CNNs to perform object localisation from medical image data in hierarchical marginal spaces. A sparse adaptive data sampling is used to capture the structure of the data, and a deep learning-based active shape model is developed that segments the target of interest given the object localisation. A two-step approach is proposed in [133] that consists of a shallow neural network at the first stage to test all the voxels in the 3D image data, followed by a deep neural network that further performs classification on the voxels obtained after the first stage. This two-stage approach employs 3D convolutions that operate directly on the volumetric data rather than operating on the three planes separately. Separable filter decomposition and network sparsification are used to account for the increased complexity due to the 3D operations and speed up the computation. A pair of CNNs are used in [134] to automatically quantify the amount of coronary artery calcification (CAC) from coronary CT angiography. CAC is a strong predictor of cardiovascular events. The first CNN identifies CAC-like voxels, thereby discarding the majority of non-CAC like voxels. The second CNN further classifies the voxels after the first stage by differentiating CAC and CAC-like negative voxels. Both the CNNs share the same architecture and the input data structure is either 2.5D or 3D. Similarly, a pair of CNNs are used in [135] to perform automated detection of the coronary artery, thoracic aorta, and cardiac valve calcifications from low

dose chest CT in cardiovascular risk assessment. It uses dilated convolutions to capture larger receptive fields in the images and incorporate contextual information. Some other works include [136] for pulmonary nodule detection, [137] for breast cancer histopathology images, [138] for detections from CT colonography and several others.

Precise localisation of the target structure of interest in medical images is often a critical step to compute target volume, precisely understand the disease location, and other such objectives. Segmentation of anatomical structures using CNNs is formulated as a pixel-classification problem that labels each image pixel/ voxel belonging to the target structure. For image segmentation, as discussed previously, many variants of CNN architectures have been developed. The fully convolutional networks and encoder-decoder architectures are widely used frameworks that greatly enhance the segmentation performance. The work in [139] builds a similar architecture, called "U-Net", consisting of a contracting encoder path to extract and encode the high-level features from the image and an expanding decoder path to derive precise localisations from the low-resolution feature map. The contracting path consists of standard convolutional and pooling layers without the fully connected layers, and the decoder path consists of up-convolution and convolution paths to recover the full-scale segmentation map. Different techniques of data augmentation are used to mitigate the issue of data-scarcity and to increase the reliability of segmentation. The U-Net can process only 2D images directly and is extended in [140] to develop "V-Net" that can directly handle 3D volumes. In addition, a novel loss function based on the dice coefficient is developed that helps in optimising the deep neural nets in highly imbalanced datasets.

Segmentation of brain MR images is performed in [141] using multi-stream convolutional architectures that operate independently on image patches of multiple scales centred at image voxels. The multi-stream outputs are combined at the softmax layer to get the final segmentation map for different tissue classes. A 19 layer fully convolutional encoder-decoder architecture is proposed in [142], similar to U-Net, to

segment skin lesions from dermoscopic images. Skin lesions usually occupy a tiny portion of the image resulting in a high imbalance between foreground-background voxel. A novel loss function based on Jaccard distance is proposed to train the network that accounts for the data imbalance. In [143], a novel deep weak supervision, is formulated under multiple instance learning framework to develop weakly supervised, fully convolutional architecture to segment cancerous regions in histopathology images. Area constraints are incorporated into the objective function to constrain the expansion of positive instances during training. Graph cut-based GrabCut segmentation approach is extended in [144] to develop DeepCut, a deep neural net architecture to segment medical objects from bounding box annotations. A fully connected conditional random field is used to regularise the segmentation. An iterative optimisation procedure is formulated using conditional random fields, which is used to update the parameters of CNN to obtain pixel-wise labels. A dual pathway, 3D CNN architecture, followed by post-processing using a fully conditional random field, is proposed in [145] to segment brain lesions from MR images.

A 3D fully convolutional architecture is developed in [146], which extends the concept of dense feedforward connections to multi-modal image segmentation. The proposed HyperDenseNet processes each modality in a separate path, with dense connections across paths, and is employed to segment brain tissues from multi-modal MR images. A novel attention gating module is developed in [147] that learns to highlight salient features in the target structure while learning to suppress irrelevant image regions. The attention gating module is then incorporated into U-Net to develop an attention U-Net model and is experimented against segmenting the pancreas from CT images. U-Net is extended to develop U-Net++ in [148] by using dense and nested skip connections in encoder and decoder sub-networks of the architecture. It helps to minimise the semantic gap between encoder and decoder networks and is experimented against multiple CT datasets. A dense hybrid U-Net is developed in [149] to segment liver tumours liver CT images. It consists of a 2D dense net that can effectively extract the features from CT slices and a 3D net that can efficiently combine the volumetric context along the depth dimension to aggregate the

final tumour segmentation map. Some other interesting works include hierarchical CNNs for segmenting breast tumours in MRI [150], novel focal Tversky loss function for lesion segmentation in [151], a dense dilated network for vessel detection in retinal fundus images [152], squeeze-and-excitation blocks towards medical image segmentation [153].

3.5.2 Neuroimaging

Functional MRI is another imaging modality widely employed to study neurological disorders. It images brain activity by determining the associated changes in the blood flow to the brain region under consideration. Deep belief networks and restricted Boltzmann machine are examined in [154] to investigate their ability to learn physiologically relevant feature representations and also determine latent relations from the imaging data. The deep networks are shown to perform similar to the independent component analysis for identifying the functional networks. Review in [155] documents the studies during the initial days of deep learning, which investigate its applications to analysing neuroimaging data of psychiatric and neurological disorders. The review article [156] summarises the applications of deep learning in neuroradiology by focusing on the several classes of applications that include image acquisition and improvement, image transformation, end-to-end diagnostic pipeline, and other fMRI applications. Deep neural network-based classification of ADHD from resting-state fMRI is the focus of another study in [157]. In [158], data-driven deep network optimisation approach that modifies the fully connected layers of a CNN architecture to determine an overall CNN architecture optimised for classifying ADHD from resting-state fMRI. A deep transformation method to extract discriminant latent feature space from resting-state fMRI is proposed in [158]. Spatio-temporal correlations of the functional activities are extracted as features, which are subsequently processed by convolutional layers to construct high-level features to classify ADHD from the resting-state fMRI. Deep neural nets are used in [159] to extract predictive markers from neuromelanin sensitive magnetic resonance imaging (NMS-MRI) images for diagnosing Parkinson's disease. NMS-MRI is

an effective modality to image the abnormalities in substantia nigra associated with Parkinson's disease.

3.5.3 Radiologist-level Performances

Despite the constraints in collecting large-scale datasets, technologies such as Picture archiving and communication systems (PACS) have enabled steady and efficient aggregation of medical image databases. CNN architectures are rapidly leveraging the steady growth of medical image databases, embracing domain-dependent needs to develop deep neural net-based analytical techniques that perform on-par with expert radiologists.

Application of CNNs to detect diabetic retinopathy from retinal fundus images is explored in [160]. A dataset of 128175 retinal images, graded for diabetic retinopathy, diabetic macular edema, and image gradability, by a panel of 54 ophthalmologists and ophthalmology senior residents was used to train an Inception-v3 [99]-based architecture. An ensemble of 10 networks was trained, and an average of the ensemble was used to compute the final output. CNNs were employed in [161] to develop an approach to extract cardiovascular risk factors from retinal fundus images, such as age, gender, systolic blood pressure. The models were developed and validated using two large-scale retinal fundus datasets from: UK Biobank (`http://www.ukbio-bank.ac.uk/about-biobank-uk`), and EyePACS (`http://www.eyepacs.org`). A 121-layer dense CNN architecture, called CheXNet, is developed in [162] to detect pneumonia from frontal- view chest X-ray images. The performance of CNN exceeds that of four practising radiologists. It is trained on a chest X-ray dataset, containing over 100,000 frontal- view X-ray images of 30,805 unique patients, with 14 disease classes. In addition to the diagnosis of pneumonia, it also produces a heatmap localising the most affected regions of the image. A single-stage GoogleNet Inception v3 CNN architecture [99] was employed in [10] to automate the classification of skin lesions from dermoscopy images, as shown in Figure 3.13. The CNN was initially pretrained on the ILSVRC image classification database, before using trans-

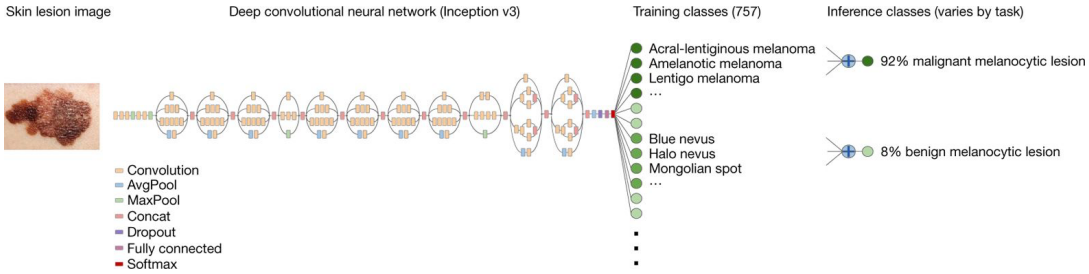


Figure 3.13: CNN architecture is trained on 129,450 dermoscopy images to perform binary skin cancer classification, achieving state-of-the-art performance on par with 21 board-certified radiologists. Figure source: [10]

fer learning to train on 129,450 clinical images consisting of 2,032 disease classes. Two binary classification experiments were designed to recognise the most common cancer and deadliest skin cancer from the images, and the results were on par with when compared against performances of 21 board-certified radiologists. Deep neural networks were demonstrated in the detection of tuberculosis (TB) on chest radiographs in [163]. AlexNet [3] and GoogLeNet [99] were employed to perform binary classification of an image as pulmonary TB or healthy. The networks used both the combination of pretraining on ILSVRC and transfer learning, and training the networks from scratch. The ensemble of the best performing algorithms was used to obtain final predictions, which were augmented by cardiothoracic radiologist. A patch-level fully convolutional neural network-based multi-task architecture was developed in [164] to perform a joint task of image segmentation and classification for the identification of acute intracranial hemorrhage from head CT images. A training subset of 4,396 head CT images, composed of 1,131 examinations positive for intracranial hemorrhage and 3,265 negative examinations, was used to train the FCN architecture based on Dilated ResNet [165]. The performance of the algorithm was on par with that of 4 American Board of Radiology (ABR) certified radiologists on an independent test set of 200 head CT images.

Application of deep neural nets for detecting Lymph Node Metastases from whole-slide pathology images in women with breast cancer is reported in [166]. Whole-slide images(399) and their reference annotations generated under the supervision of

expert pathologists were collected to organise a coding competition (CAMELYON16 - Cancer Metastases in Lymph Nodes Challenge 2016) for aggregating solutions from research groups around the world. The algorithms were validated against two tasks: Metastasis Identification, and Whole-Slide Image Classification, and the performances of the automated algorithms was compared with that of a panel of 11 pathologists. Some of the deep neural nets performed at par or exceeding the performance of pathologists. Retrospectively collected 313318 non-contrast head CT images, from multiple medical centres, were analysed in [167] to develop and validate a CNN architecture for automated detection of intracranial haemorrhage, calvarial fractures, midline shift, and mass effect, achieving a high AUC that was validated independently by three radiologists. A deep neural net is trained in [168] to determine the amount of to predict the amount of salvageable tissue from MRI images in acute ischemic stroke. Other notable mentions include pathology detection in heat CT scans [169], hemorrhage detection in ct scans [170], and others.

3.5.4 Gaps and Limitations

Despite these advancements, several challenges and limitations exist. In the case of cerebral aneurysms, most of the existing works concentrate mostly on DSA and MRA images. DSA is usually considered to be the gold standard. Some of the works include decision tree-based approach in [171], a new aneurysm feature in [172], a novel surface descriptor to quantify how closely a given region approximates a tubular structure in [173], and surface Voronoi diagrams-based approach in [174]. Most of these approaches rely on DSA images, which are acquired following an invasive and operator-reliant procedure, making it an uncomfortable process for the patients. Further, they experiment on very limited datasets. They have been explained in greater detail in Chapter 4.

Functional imaging is generally used to image the brain activities in Parkinson's patients. The review article [175] outlines the importance of structural and functional imaging procedures in diagnosing Parkinson's disease and documents some signifi-

cant relevant works. Dynamic causal modelling (DCM) is used to compare different competing models of brain function within a group or between groups and to analyse the abnormal voluntary actions [176, 177]. In [178], advances in structural imaging for the diagnosis of Parkinson's disease are studied. Notably, structural imaging of substantia nigra, extra-nigral structural imaging, and advances for related structural imaging biomarkers for diagnosing PD are discussed. The review also documents several other works focusing on the prognosis of cognitive dysfunction in PD patients, MRI markers for differential diagnosis, and others. Most of the works in literature concentrate on analysing PD from MRI images. However, MRI is an expensive imaging modality, and also, it may not be applicable in patients with certain existing health conditions. Further, these works do not concentrate on analysing the progression of speech abnormalities, which is present in a significant portion of PD patients [14–16]. Non-imaging modalities to analyse speech abnormalities and their limitations and other relevant works are discussed in Chapter 5.

3.6 Summary

This chapter presented an overview of advancements in the field of image recognition using convolutional neural nets. The concepts behind CNNs have existed for some time; however, the growth in computing hardware and aggregation of large-scale datasets have dramatically pushed their performance limits. Different kinds of CNN architectures have been investigated that underline the importance of depth, width, and other complexities in a network and their contributions to performance enhancements. Most of the initial developments in CNN architectures were made on an image classification challenge; however, they have been subsequently extended to object detection and image segmentation, among other tasks. Frameworks such as region-based CNNs for object detection and an encoder-decoder class of architectures for image segmentation have substantially improved the performances in their respective objectives. The straightforward extension of CNNs to medical images is constrained by several challenges, including a lack of adequate data and expert

reference standards. There are, however, several works that have tried to somewhat address these challenges through a reduction in the complexities of architectures, transfer learning, and other such measures. The final section in this chapter briefly summarised some of the works that have successfully addressed challenges for some pathologies, which have subsequently resulted in state-of-art algorithms that perform at par or even exceed the performances of expert radiologists.

Chapter 4

Detecting and Localising Cerebral Aneurysms from Computed Tomography Angiograms

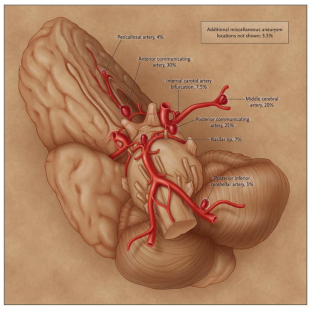
A cerebral aneurysm is a form of the cerebrovascular disease characterised by weakness of the blood vessel wall, which may rupture, leading to hemorrhagic stroke. This chapter focuses on the automated detection of unruptured cerebral aneurysms from CT angiograms (CTA). A large-scale CTA image dataset is constructed and is used to propose and develop a novel CNN-based machine learning approach to detect and localise the unruptured cerebral aneurysms.

4.1 Introduction and Background

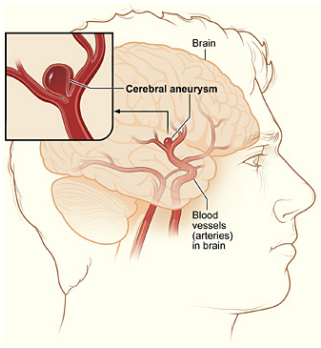
An aneurysm is a form of the cerebrovascular disease characterised by weakness of the blood vessel wall. It may occur anywhere in the circulatory system but is usually formed in the blood vessels of the brain. Aneurysms in the brain are called cerebral aneurysms (Figure 4.1). Cerebral aneurysms are primarily classified by their shape. A most common type is a saccular aneurysm that accounts for about 80-90% of all the cerebral aneurysms. It has a neck and a stem and is also known as berry aneurysm because of its shape. The fusiform aneurysm is the other type of brain aneurysm, which is less commonly occurring and does not have a stem. Mycotic aneurysms, which are typically very rare, occur when an infection spreads through the bloodstream from other areas of the body. The scope of this study is limited to

saccular aneurysms.

Cerebral aneurysms have an incidence rate of 3-5% amongst the general population [179]. A severe complication of cerebral aneurysms is that they can rupture, leading to extravasation of blood into the subarachnoid space, known as subarachnoid hemorrhage (SAH). Ruptured aneurysms are fatal in 40% of the cases, with around 46% of the survivors suffering from some form of long term cognitive impairment [11]. Subarachnoid hemorrhage accounts for 5% of deaths from stroke [180] and about 27% of all stroke-related years of potential life lost before the age of 65. Several long-term experiments study the unruptured aneurysms, their formation and causes, rate of growth, rupture risks, rupture probability, and rupture consequences [181–184].



(a) Intracranial vasculature, showing the most frequent locations of intracranial aneurysms. Figure source: [179]



(b) Cerebral aneurysms. Figure source: [185]

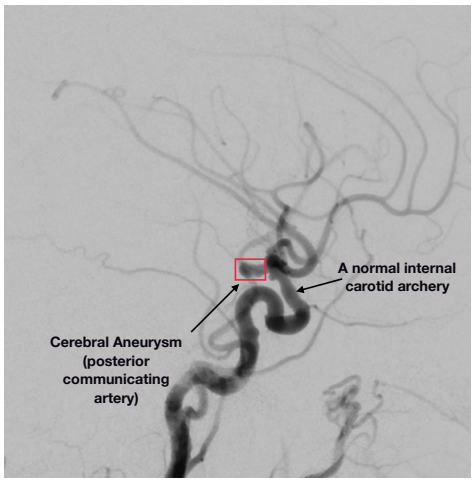
(c) Aneurysm rupture. Figure source: [186]

Figure 4.1: Cerebral aneurysm is a cerebrovascular disease characterised by weakness of blood vessels of brain. Rupture of aneurysms may lead to hemorrhagic stroke.

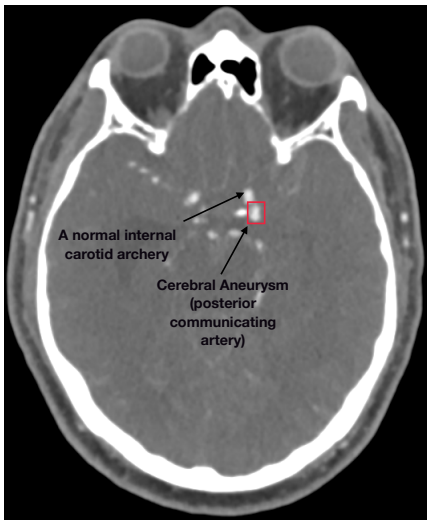
Cerebral aneurysms are generally discovered after rupture or accidentally during diagnostic examinations. Unruptured aneurysms must not be missed as this would provide opportunities for preemptive treatment, preventing the fatal consequences. Neuroimaging is a widely preferred mode of diagnostic to identify cerebral aneurysms. Digital subtraction angiogram (DSA) is generally considered to be the gold-standard in identifying brain aneurysms. DSA is a fluoroscopy technique that uses X-rays and

catheter to visualise the blood vessels. The catheter is guided to the blood vessels, and a contrast agent is injected to acquire multiple 2D views that are used to obtain a 3D representation of the vessels. It produces images with a high spatial resolution and is generally considered to be the reference standard. However, it requires contrast agents, uses ionising X-ray radiations, and more importantly, is an invasive procedure that requires an expert operator to guide the catheter to major blood vessels while minimising the risk of injury. It is, therefore, invasive and stressful to the patients [187]. Recent advancements in CT imaging have enabled a more efficient and non-invasive imaging mechanism to capture the blood vessels. CTA uses multi-detector CT scanners to visualise blood vessels, with the help of a contrast agent. It has been reported that the state of the art multi-detector scanners can help diagnose aneurysms greater than 4 mm with almost 100% sensitivity [188]. This means that CTA is a reliable imaging modality that does not miss many aneurysms when compared to the DSA, which is usually considered to be the gold-standard. Advances in spatial resolution capabilities of the CT scanners have further increased the sensitivity in detecting even smaller aneurysms. Figure 4.2 shows the identification of cerebral aneurysm on DSA and CTA slices.

The main advantage of CT technology is that it is affordable and efficient to use. This makes it a preferred modality for initial diagnostics, although, the radiation effects can be significant beyond a specified limit. Magnetic Resonance Angiography (MRA) uses a strong static magnetic field and radio-frequency pulses to produce 3D vessel representations rather than ionising X-ray radiations used by the CT scanners. MRA uses time-of-flight sequences to visualise the blood vessels that eliminates the need to use contrast agents. The vessels are better visualised on a 3 Tesla MRA over a CTA. A 7 Tesla MRI scanner can result in images with a high spatial resolution that helps to detect even smaller aneurysms and their variations with high precision. Despite these advantages, MRA is generally not the initial diagnostic modality used and is reserved for follow-up, as (a) the technology is highly expensive, (b) it is difficult and less efficient to obtain MRA images compared to CTA, (c) it cannot be used with people having certain health conditions.



(a) Cerebral aneurysm on a DSA slice.



(b) Cerebral aneurysm on a CTA slice.

Figure 4.2: DSA is usually considered the gold-standard in diagnosing cerebral aneurysms. The recent advances in CT imaging have enabled a high sensitivity diagnosis of aneurysms from CTA images.

CTA is generally the initial diagnostic choice as it is convenient to use, non-invasive, and affordable. It produces images with a spatial resolution that is good enough to detect aneurysms with high precision. The scope of this study is restricted to data acquired using CTA. A common approach to diagnose aneurysms is to acquire the CTA, which is then observed by expert radiologists to identify possible aneurysms by characterising their shape, size, and location. Clinicians may use this information to plan their subsequent treatment procedures. However, manual identification of aneurysms can be challenging and laborious, presenting several challenges:

- Aneurysms anywhere within the internal carotid artery can easily be obscured by bone artefact. For instance, in the region of the temporal bone, the internal carotid artery traverses from the neck, enters into the skull (bone), directs upwards into cavernous sinus (next to bone), before settling underneath the belly of the brain.

- Arteries may loop and wind around each other. The looping can sometimes

look like an aneurysm that makes it difficult to visually untangle the loop. It may require much time spent in a multi-planar examination that might still result in erroneous conclusions.

- Inter- and intra- observer variance can result in inconsistent conclusions, making it difficult to come to a unanimous diagnostic decision.

There are a limited number of works in literature that concentrate on automated detection of aneurysms from the angiographic datasets. A framework for detection and quantification of the morphology of aneurysms from DSA images is proposed in [171] that uses a decision tree based detector employing responses of blobness and vesselness filters. Aneurysm location is then used to seed growcut, followed by neck extraction based on intravascular ray-casting. The proposed approach achieved a sensitivity of 100%; however, it was evaluated on only ten cerebral 3D-DSA images. Many of these aforementioned works follow the general flow of using some form of a priori information about the dataset to guide the feature extraction process. These features are then used to model the aneurysms. A new feature to characterise the aneurysms and an algorithm based on sphere enhancing filter were proposed in [172] to detect aneurysms from multimodal angiographic datasets automatically. The approach was tested on CTA and MRA datasets, with a sensitivity of higher than 93%. A modified k-means clustering-based approach was proposed in [189] to identify cerebral aneurysms from 3DRA, 3DMRA, and 3DCTA images, achieving high sensitivity. However, a primary limitation with these approaches is the very limited datasets of about 20 images per modality. It makes it challenging to assess the generalisability of the algorithm. Further, [189] reports high false positive of about 37 per datasets. In addition, these approaches report only binary identification metrics and do not include the localisation accuracies.

The work in [173] proposed an approach applicable to the segmented cerebral vasculature. This approach detects aneurysms as suspect regions on the vascular tree by inspecting small regions along the tree and developing a novel surface descriptor to quantify how closely a given region approximates a tubular structure. The approach

was tested against 3D rotational angiographic (RA) and 3D CTA datasets, achieving a sensitivity of 100%. However, the datasets used contain only ten RA and ten CTA volumes, making it difficult to ascertain the generalisability of the algorithm. Further, it requires a preprocessed and mainly, segmented cerebral vasculature before it can proceed with aneurysm detections. The accuracy of the entire process depends on the segmentation accuracy of the structure.

An automated aneurysm detection approach using surface Voronoi diagrams was developed in [174] by incorporating a new definition for the aneurysm neck based on the minimum cost path around the aneurysm sac. This approach uses the fast marching method to obtain a scalar field on the surface and then computes minimal path along this scalar field to reliably identify the aneurysm neck. A semi-automated approach based on the Geodesic Active Counters (GAC) is proposed in [190] to segment cerebral aneurysms from the 3D CTA data. Image intensity, gradient magnitude, and intensity variance are used in the level set evolution, and the parameters for the level set function are derived from the image statistics in the local neighbourhood of a user-defined point.

To the best of our knowledge, there was no existing work in the literature documenting a large scale CTA dataset of brain aneurysms and developing/validating automated machine learning-based aneurysm identification method at the time of beginning the work. However, there has been a similar work done concurrently in [21] that develops a CNN based architecture to identify aneurysm from CTA images. The dataset consists of about 328 aneurysms examinations in addition to the controls with no aneurysms. A semantic segmentation architecture is developed employing 3D convolutions operating directly on the CTA voxels. Performance of the radiologists with and without the model augmentation are reported. The work was carried out concurrently; therefore, we are unable to explicitly address the limitations of the work and document performance comparisons. However, it does have a few limitations that are complemented in our work. The approach consists of a 3D CNN architecture to segment the cerebral aneurysms from CTA images. The 3D

architecture, however, is inherently computationally intensive, resulting in a significantly increased training complexity and necessitating a larger dataset to achieve satisfactory performance. Further, even though their work addresses the localisation of the aneurysms through their segmentation from CTA images, the performance metrics document only binary diagnostic accuracies without any form of localisation accuracies.

4.2 Overview of the Proposed Contribution

CTA examinations of the brain were acquired, retrospectively, at the Melbourne Brain Centre, Royal Melbourne Hospital, Australia. The images were acquired in a compressed DICOM file format, which was converted to NIfTI for subsequent processing. The images were analysed by expert neuroradiologists, with access to radiology reports and corresponding DSA examinations. DSA is considered a gold-standard in finalising the diagnosis of aneurysms. Open-source software ImageJ [191] was used to analyse the CTA images and generate ground-truth annotations on the three orthogonal views (axial, sagittal, and coronal) of the CTA volume. Aneurysms were first diagnosed and then delineated by drawing a tight bounding rectangle around them such that the walls of the bounding boxes are tangential to the boundaries of the aneurysm. The bounding boxes are parameterised by the left-lower coordinates, width, and height of the rectangle.

Further, this work proposes and implements a CNN architecture to detect and localise the cerebral aneurysms from the CTA volumes. A novel multiview architecture is proposed that incorporates the information from all the orthogonal views of the 3D data and can localise the aneurysm voxels from the images.

4.3 Problem Formulation

The overall goal of this work is to detect and localise unruptured cerebral aneurysms from the CTA images. Specifically, it involves two objectives: (i) a binary task of de-

tecting whether a CTA image contains an aneurysm, and (ii) an accurate localisation of the aneurysm by identifying the region of interest within the CTA image containing the aneurym. To achieve both the objectives, we formulate a dense prediction task that involves categorising all the CTA voxels as aneurysm or non-aneurysm voxels. The labelled voxels are subsequently used to detect the presence of the aneurysm in the CTA examination and also, localise the region of interest.

4.4 Method

4.4.1 Dense Prediction

CNNs have made significant strides, as explained previously in Chapter 3, in several image analysis tasks from image recognition to image segmentation. In particular, encoder-decoder class of CNN architectures [9] have gained a significant prominence due to their ability to recover the full resolution semantic feature map of an image from its high-level, low-resolution feature map, with high accuracy. U-Net [139] is one such encoder-decoder architecture widely employed in medical image analysis that produces state-of-art results in image segmentation.

We propose a novel, modified U-Net architecture, as shown in the Figure 4.3 to localise and identify the aneurysms from CT angiogram volumes. It addresses two critical challenges in detecting an aneurysm-like small object from CTA volumes: (i) loss of discriminative feature information due to pooling layers, and (ii) incorporating contextual information to enhance the accuracies of dense prediction results.

Dense prediction in image recognition tasks involves predicting a label for each image pixel. Therefore, the dense prediction from a CTA volume is the process of assigning a label to each CTA voxel depending upon whether it belongs to an aneurysm.

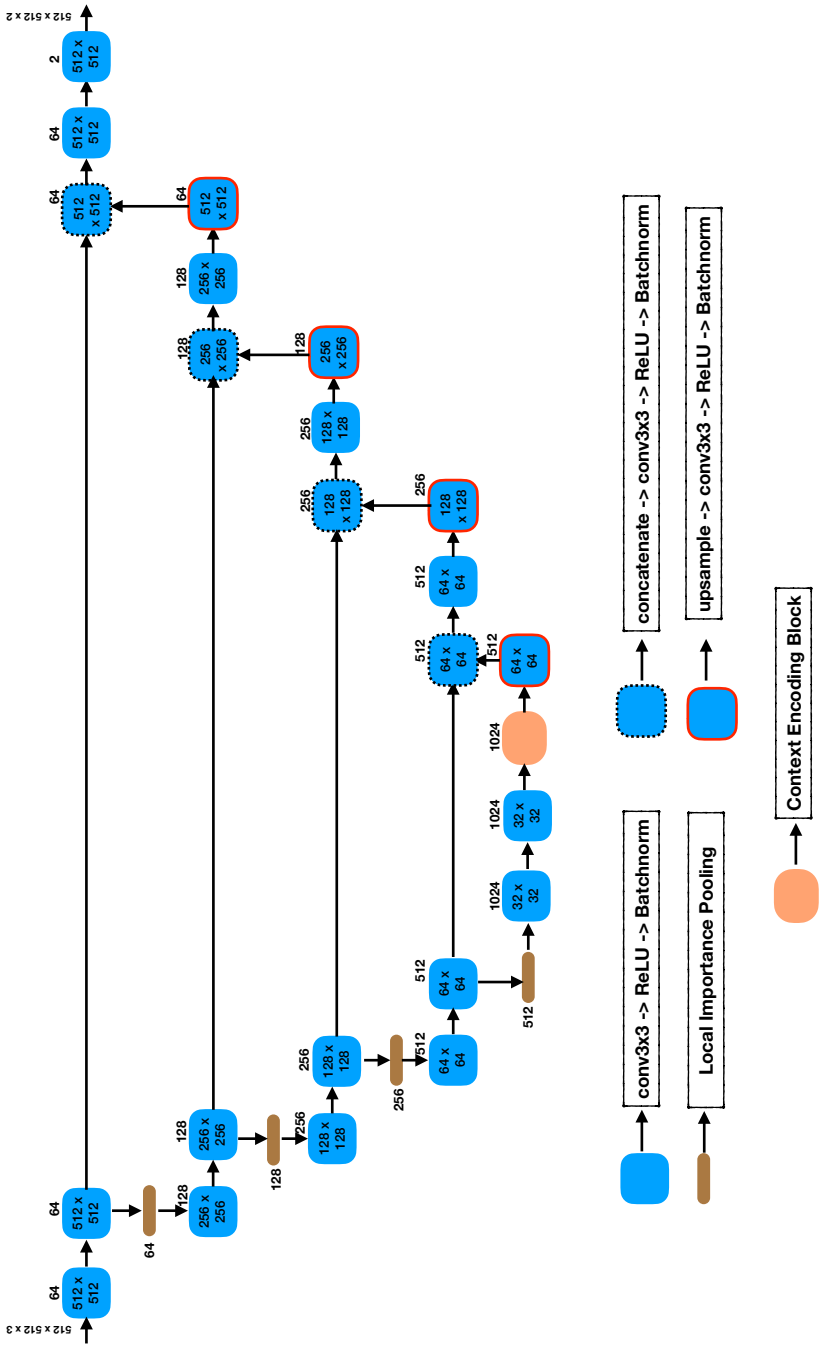


Figure 4.3: Context U-Net with local importance pooling. The local importance pooling blocks efficiently model the discriminative voxels in a neighbourhood and capture their distribution in the downsampled output. The context encoding block captures the feature map at multiple resolutions using atrous convolutions.

4.4.2 Pooling

Pooling is a critical feature of CNN architectures in the dense prediction tasks. It progressively reduces the spatial dimensions of a feature map, thereby reducing the number of learnable parameters. Besides reducing the computational complexity and controlling the overfitting, pooling also introduces spatial invariance by discarding the information about exact spatial locations of features. The different forms of pooling include max-pooling, min-pooling, average-pooling. However, a significant limitation of pooling is the loss of information. The downsampling operation leads to not only loss of feature map information, but also loss of discriminative feature information. Pooling works typically by downsampling the voxels in a predefined neighbourhood of the feature map without considering the discriminatory ability of a voxel towards the final performance. The loss of discriminative information may prevent the subsequent layers from learning class-specific hierarchical feature representations and may ultimately affect the predicted label for the CTA voxel. It is especially crucial in dense voxel prediction tasks involving tiny aneurysm-like objects that constitute a small part of the image, making it necessary to minimise loss of any discriminative information.

Pooling operation in a CNN architecture, as discussed in [192], can be formulated as:

$$O_{x',y'} = \frac{\Sigma_{(\Delta x,\Delta y)\in\Omega} F(I)_{x+\Delta x,y+\Delta y} I_{x+\Delta x,y+\Delta y}}{\Sigma_{(\Delta x,\Delta y)\in\Omega} F(I)_{x+\Delta x,y+\Delta y}} \tag{4.1}$$

where, I is the input feature map, Ω is the kernel indices set consisting of relative sampling locations $(\Delta x, \Delta y)$ in a sliding window, and (x,y) represents the left-top location of the sliding window in the input feature map, in reference to the output location (x',y). $F(I)$ represents the importance map for the image I, with the same size as that of I and $F(I) \geq 0$ over space. Pooling in this formulation can be interpreted as weighted sum over each sliding window with locally normalised weights:

$$\frac{F(I)_{x+\Delta x,y+\Delta y}}{\Sigma_{(\Delta x,\Delta y)\in\Omega} F(I)_{x+\Delta x,y+\Delta y}} \tag{4.2}$$

Therefore, $F(I)$ represents the importance map of a feature map image I and gives importance to voxels in a neighbourhood during pooling operations. An average pooling, for instance, gives equal importance to all the voxels in the neighbourhood. Max-pooling gives importance to the maximally active feature in the neighbourhood. However, none of these operations explicitly address the discriminatory features and assign them the maximum priority. The equal importance in average pooling may cancel out the discriminating and non-discriminating features, whereas max-pooling works on a predefined assumption that the maximally active feature is always the most discriminative. An ideal pooling operation would instead adaptively prioritise the voxels in the neighbourhood that are the most discriminative and contribute the most for the particular task, based on the final objective. We incorporate locally important pooling [192] by constructing an importance map F using a small, fully convolutional network that learns the importance weights for the voxels during the training of the CNN architecture based on the final objective. A small fully convolutional network G is followed by an exponential operator to obtain non-negative importance weights for the voxels, formulated as:

$$F(I) = \exp\left(G(I)\right) \tag{4.3}$$

where I is the feature map and G is the learnable logit module that learns during the end-to-end CNN training. Therefore, combining both the equations, learnable locally important pooling operation can be formulated as:

$$O_{x',y'} = \frac{\Sigma_{(\Delta x,\Delta y)\in\Omega} \exp\left(G(I)\right)_{x+\Delta x,y+\Delta y} I_{x+\Delta x,y+\Delta y}}{\Sigma_{(\Delta x,\Delta y)\in\Omega} \exp\left(G(I)\right)_{x+\Delta x,y+\Delta y}} \tag{4.4}$$

In this work, the logit block $G(I)$ is constructed through a sequence of convolutional, instance normalisation, and non-linear sigmoid layers, as shown in Figure 4.4. It learns during the CNN training process to construct the importance map that helps to identify the discriminative voxels in a neighbourhood window that contribute significantly towards the final voxel prediction performance. Adaptive importance weights in the importance map $F(I)$ for feature map I enable the network to effi-

ciently model the discriminative voxels in a neighbourhood and capture their distribution in the downsampled output.

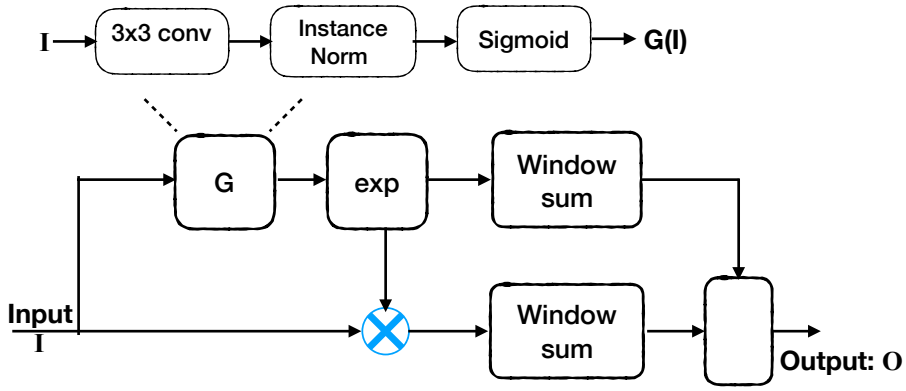


Figure 4.4: Logit block, in local importance pooling, consisting of a fully convolutional network followed by an exponential block to ensure the adaptive importance weights are non-negative.

4.4.3 Context Encoding

Contextual information is another crucial aspect that needs to be incorporated into dense voxel prediction tasks. Global contextual details of the image and multi-scale contexts of a small aneurysm-like object are essential to compensate for the loss of resolution associated with pooling layers and capture fine-grained image details. Atrous/dilated convolution is an efficient and widely used approach to capture feature information at multiple scales. Atrous convolution allows the network to control the resolution of the feature map without the associated increase in overall trainable parameters. It can be seen from Figure 4.5 that the increase in atrous convolution rate allows an increase in field-of-view and enables analysis of the feature map at increased resolutions without adding to the overall trainable parameters.

Given an input image I, a filter kernel w, atrous convolution at a rate r is formulated as:

$$y[i] = \sum_{K} I[i + r.k]w[k] \tag{4.5}$$

where, $y[i]$ is the convolved output at location i. The atrous rate r represents the

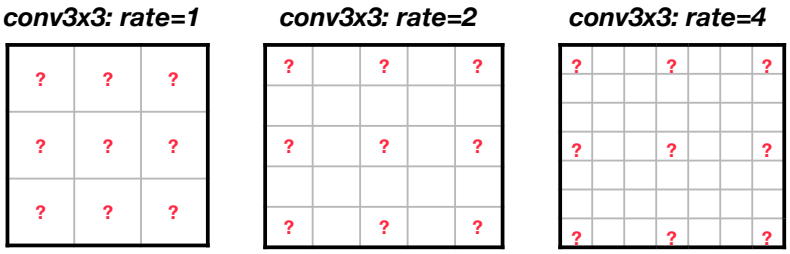


Figure 4.5: Atrous convolution filters. The rate parameter controls the receptive fields of an atrous convolution filter. Atrous convolution at a rate of 1 is the same as standard convolution and an increase in the rate increases the receptive field.

upsampling stride used to upsample the filter before the convolution operation. A standard convolutional operator has $r = 1$. As can be seen from Figure 4.5, it allows the network to explicitly control the receptive field-of-view and modify the density of features in a given spatial window, accordingly. We incorporate a context encoding block based on the atrous spatial pyramid pooling [118] to probe the high-level, low-resolution features at multiple scales and field-of-view. It is implemented by varying the atrous convolution rates and constructing an encoding block that performs the multiple-rate atrous convolutions in parallel. The atrous convolution blocks at different rates in the context encoding block, as shown in Figure 4.6, probe the incoming high-level, low-resolution feature map at multiple scales and extract features with varying feature densities. The global average pooling block incorporates the global image context and adds the image image-level information to the multi-scale feature pyramids. This contextual information is combined via a convolution operator in a 1×1 to fuse the information channels and capture multi-resolution feature information with varying scales.

4.4.4 Loss Function

A critical component in training a CNN architecture in a supervised learning paradigm is the loss function. The loss function needs to be defined apriori to the training and is based on the desired objective of the task. Generally, it fully reflects some variation of the performance metric for the task, which needs to be optimised for overall high performance.

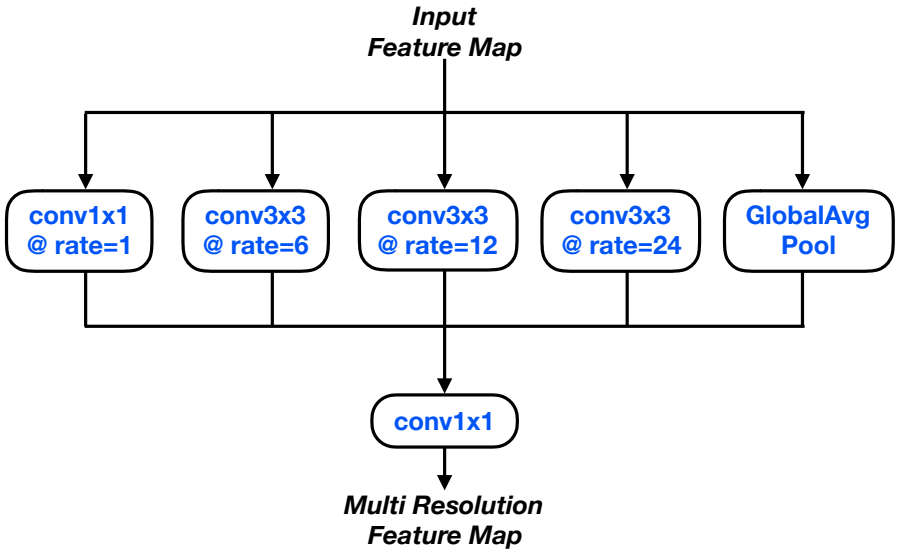


Figure 4.6: Context encoding block. The convolution blocks with different atrous rates probe the incoming low-resolution feature maps at multiple receptive fields to encode the context at multiple scales.

Binary cross-entropy is a usually employed classification loss in several CNN architectures for supervised training. A binary cross-entropy loss for a voxel prediction in this task involving two classes can be formulated as:

$$BCE(y,\hat{y}) = -(y\log(p) + (1-y)\log(1-p)) \tag{4.6}$$

where, y represents the true class label (aneurysm or background) for the voxel, and p represents the predicted probability for the aneurysm class. Binary cross entropy is a convex loss function, smooth, and is easy to optimise. However, aneurysms are usually tiny objects, and in a CTA volume of dimensions $(512 \times 512 \times 600)$, they may occupy as small a portion as $10 \times 10 \times 5$ (500 voxels), leading to an overall class-balance of much higher 100,000. The localisation accuracies usually are measured using the dice coefficient score formulated as:

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4.7}$$

where A represents the ground-truth voxels, and B represents the voxel predic-

tions.We incorporate a soft dice loss based on the DSC as: $SDL = 1 - DSC(A,B)$. It can be seen that the dice loss is based on the amount of intersection between the predicted and ground-truth voxel classes and therefore, automatically accounts for the heavy class imbalance in the data. A problem, however, with the dice loss is that the loss is limited to $[0,1]$ and does not have a smooth gradient, making it difficult to optimise using a gradient-based optimisation algorithm. Therefore, we propose to use a weighted combination of binary cross-entropy and soft dice loss to construct the final loss function L, formulated as:

$$L(A,B) = \alpha.BCE(A,B) + \beta.DSC(A,B) \tag{4.8}$$

where the weights α and β are determined empirically through experimentations.

4.5 Dataset and Annotations

The aneurysm dataset consists of 215 brain CT angiograms of 215 individuals, acquired retrospectively, at the Melbourne Brain Centre, Royal Melbourne Hospital, Australia. The data acquisition was limited to individuals admitted in the time-period 2010 to 2018 at a single centre. The subjects mostly had a single aneurysm. Tables 4.1 and 4.2 list the preliminary demographic characteristics of the CTA images and cerebral aneurysm present in the images, respectively. Figure 4.7 shows the distribution of the aneurysm based on their location of occurrence.

Table 4.1: Preliminary characteristics of the CTA image dataset.

Age	54.1 ± 15.15
Gender	67% Female
Slice spacing	0.44 ± 0.07 mm
Slice thickness	0.75 mm
Number of slices	744 ± 136
Slice spatial dimensions	512×512

Table 4.2: Preliminary characteristics of the cerebral aneurysms present in the CTA image dataset.

Minimum aneurysm size	$(1.76, 1.32)$ mm
Maximum aneurysm size	$(29.48, 32.56)$ mm
Mean aneurysm size	$(6.95 \pm 4.63, 6.73 \pm 4.64)$ mm
Minimum aneurysm thickness	0.75 mm
Maximum aneurysm thickness	47.25 mm
Mean aneurysm thickness	8.98 ± 6.71 mm

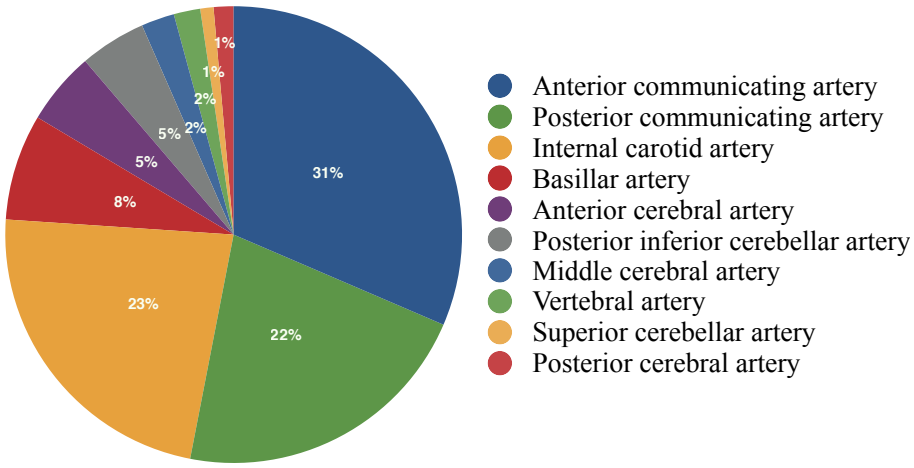


Figure 4.7: The locations of incidence for the cerebral aneurysms present in the current CTA dataset.

Inter-rater agreement

The diagnosis by a single observer, however, may not be reliable, and multiple observers need to be consistent and agree on a diagnosis to ensure its validity. Therefore, 30 CTA volumes are randomly sub-sampled from the full dataset, and three experienced neuroradiologists independently diagnose and annotate the aneurysms from the images. The inter-observer agreement measures are then computed on the three sets of annotations in two forms: (i) diagnosis agreement, and (ii) localisation agreement.

Diagnosis agreement describes the agreement in reliably diagnosing a slice containing an aneurysm. We assign a score of 1 if the aneurysm is identified in a slice and 0

otherwise. Fleiss's kappa coefficient [193], which measures inter-rater agreement for categorical items between multiple observers, is used to determine the agreement among the three neuroradiologists in identifying a slice with an aneurysm. The scores are computed over every subject and then averaged over all the subjects to obtain the final averaged agreement measure. Fleiss' kappa is computed as:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e} \tag{4.9}$$

where $\bar{P}$ is the average observed probability of agreement and P_e is the average chance agreement. The mean Fleiss's kappa for the three sets of observers across all the patients is 0.869, with a standard deviation of 0.005. A high value of Fleiss's kappa indicates that there is a high level of agreement amongst the three observers confirming a reliable diagnosis of aneurysms.

However, the main objective of this work is not only to identify the aneurysms but to localise them in the CTA slices. The radiologists localise the aneurysms by drawing enclosing bounding rectangles around them. The localisation measure describes the agreement between the three observers in localising aneurysms through bounding boxes. The amount of agreement between two bounding rectangles generated by any two observers can be computed using the intersection-over-union (IoU) measure. IoU measures the intersecting area between the two bounding rectangles relative to the combined area of the rectangles, thus measuring the amount of agreement between the two rectangles. IoU for any two bounding rectangles A and B is computed as:

$$IoU = \frac{intersection(A,B)}{union(A,B)} \tag{4.10}$$

.

IoU measure for a CTA volume is computed by aggregating the IoUs calculated over the set of slices for which there exist bounding rectangle measurements by all the observers. These slice level IoUs are then combined to obtain a per-patient measure, which is then averaged to obtain a final agreement score amongst the observers. We

compute two agreement scores by aggregating the slices level scores in two different approaches. We first take a simple average of all the slice level IoUs to determine the per-patient IoUs, which are then averaged to obtain a type-1 score. A limitation of the type-1 score is that in the case of small aneurysms, a small offset between rectangles reflects in a poorer slice IoU. Further, it is sometimes difficult to accurately identify the edge slices of aneurysms. Therefore, we then compute type-2 IoU scores for a patient by aggregating the slice level IoUs over five slices with the best slice level IoUs. It is done to verify that the independent observers agree on at least five slices in a CTA volume. An average type-1 IoU score in aneurysm localisation amongst three radiologists is 0.57 with a variance of 0.009. Average type-2 IoU score amongst the three observers is 0.78 with a variance of 0.01. The higher IoU scores greater than 0.5 indicate that there is at least 50% overlap in the aneurysm locations localised by a pair of observers. It indicates a reliable and consistent localisation between observers, especially when accounting for the existence of aneurysms as small as 4 mm in the dataset.

4.6 Implementation

The following subsections describe the implementation, training, and testing of the CNN architectures, in addition to documenting the performance metrics of several methods.

Training

The positive CTA examinations with aneurysms were split into percentages of 70/15/15 subsets for training, validating, and testing the implementations. CTA examinations without aneurysm were excluded from the training set. To train and develop the neural network methods, we used a Ubuntu 18.04 computing machine with the Nvidia GeForce RTX 2080Ti graphics card, which has 24 Gigabyte global memory. The implementation was done using PyTorch library [194] in Python, and only the slices containing aneurysms were used for training. We employ 2D convolutions in the

proposed architecture, which may not fully incorporate the information along the depth axis of the CTA volume. In order to partially account for this, we construct a 3-channel image input by combining a slice at position i with slices at locations $i-1$ and $i+1$. This pseudocolour, 3-channel image enables the network to account for the loss of depthwise information and act as a 2.5D computation. The images were subsequently normalised and cropped as a part of data preprocessing pipeline. Horizontal flipping was employed as a data augmentation technique to increase the diversity of the training set. Stochastic gradient descent optimisation algorithm was used to train the neural networks with a mini-batch size of two and a momentum of 0.9 for about 125K iterations. A polynomial learning rate scheduler with an initial learning rate of 0.007 was used to adjust the learning rate throughout the training process. Weight decay of 0.0005 was applied to all learnable parameters, including those of batch normalisation. The networks were trained for 50 epochs, with training times of about 18-20 hours.

Testing

In addition to the testing subset constructed by splitting the original CTA dataset with aneurysms, negative CTA examinations without aneurysms were included in the test set to be able to test the diagnostic accuracy of the proposed method. A multiview architecture was employed during the testing phase by testing the trained model on all the three orthogonal views of the image volume. The multiview aneurysm predictions were subsequently combined to get the final prediction for an aneurysm voxel. The implemented CNN architectures, described in detail below, take an average of about 45-50 seconds to perform predictions on a CTA view.

4.7 Results and Discussion

4.7.1 Binary Classification of the CTA Examination

A 3D connected component analysis is carried out on the final dense predictions of the CTA volume to determine a 3D bounding rectangle tightly enclosing the voxels belonging to an aneurysm, if any. Prior clinical information-based rules are then formulated, based on the 3D bounding boxes, to assign a binary label to the CTA examination reporting the presence/absence of an aneurysm. The performances of the CNN architectures for the binary classification task are documented using the performance metrics in (4.11)- (4.14).

$$Sensitivity/Recall = \frac{TP}{TP+FN} \tag{4.11}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4.12}$$

$$Precision = \frac{TP}{TP+FP} \tag{4.13}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{4.14}$$

where TP, FP, TN, FN represent the true positive, false positive, true negative, and false negative predictions in classifying the CTA examination. The binary classification performances accuracies of the proposed approach and some of the widely used CNN architectures are reported in Table 4.3. U-Net11 and U-Net16 adapt standard classification architectures VGG11 and VGG16 to construct encoders in the encoder-decoder type segmentation architectures [195]. AlbUNet uses ResidualNet encoder to construct segmentation architecture. An attention gate based U-Net architecture is proposed in attention U-Net [147] that automatically learns to fo-

cus on target structures by highlighting salient features useful for a particular task. Tiramisu [196] extends the dense CNNs to construct a 100 layer fully connected DenseNet segmentation architecture.

It can be observed from Table 4.3 that the proposed architecture achieves a better overall performance. The local importance pooling prevents the loss of discriminative information, while the context encoding block operates at multiple receptive fields and captures multiscale feature information. AlbUNet achieves a similarity sensitivity score; however, it has a very high false-positive rate. The U-Net11, U-Net16, attention U-Net architectures, on the other hand, have a low false-positive rate; however, their sensitivity is poor in detecting aneurysms. Aneurysms are complex structures with varying size, shape, and intensity distributions. It is challenging for clinicians to diagnose aneurysms without assistance and generally requires expert neuroradiologists. Lack of timely diagnosis may lead to their rupture resulting in subarachnoid haemorrhage, with a high mortality/ morbidity rate. The cost of missed unruptured aneurysm detections is very high; therefore, sensitivity is an important performance criteria. It can be seen from Table 4.3 that the proposed approach achieves high sensitivity while providing a slightly reduce false-positive rate performance.

Table 4.3: Performances of the CNN architectures in the binary classification task of whether a CTA examination contains an aneurysm.

Method	**Sensitivity/Recall**	**Specificity**	**Precision**	F_1 **Score**
U-Net11 [195]	0.84	0.938	0.72	0.792
U-Net16 [195]	0.8	0.988	0.12	0.597
AlbuNet [195]	0.92	0.08	0.5	0.647
Attention U-Net [147]	0.68	0.8	0.926	0.723
U-Net [139]	0.64	0.88	0.84	0.727
Tiramisu [196]	0.4	0.92	0.83	0.540
Ladder [197]	0.88	0.32	0.56	0.687
Proposed Work	**0.92**	0.6	0.69	**0.793**

4.7.2 Localisation of the Cerebral Aneurysms from the CTA Examination

Another important objective of the work is to localise aneurysms in addition to the binary classification of whether a CTA examination. We report the findings of the localisation accuracies through dice score, defined as:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{4.15}$$

In addition to the binary classification of the CTA examinations, another essential objective is to localise the identified aneurysms. The final dense prediction of the CTA volumes includes the predictions for each CTA voxel. The accuracies of voxelwise predictions are reported in this subsection. In addition to the voxelwise sensitivity and specificities, the dice coefficient score (DSC) is used to evaluate the accuracy of the voxel predictions by computing the overlaps between CTA predictions and ground truth. Table 4.4 tabulates the voxelwise sensitivity and specificity performances of the CNN architectures, along with the DSC score. It can be observed that the specificity is saturated at 0.999 for all the methods. It is because of the substantial data imbalances in the CTA volume and the tiny portions that the aneurysms occupy. The voxel sensitivity is a hard measure that accounts for the classification of each voxel belonging to an aneurysm; therefore, it is difficult to achieve a high value as that of binary classification sensitivity. The DSC shows the accuracy of the delineated aneurysm structure, about the ground truths. The aneurysms are tiny structures that have an average size of about 5-10 mm, making it difficult to achieve highly accurate localisations. Further, 3D connected component labelling is used to increase the certainty of aneurysm localisation and identify its enclosing region-of-interest, as shown in Figure 4.8. It can also be seen that the networks with similar sensitivity/ specificity score achieve different dice scores. The sensitivity and specificities in Table 4.3 document the diagnostic ability of the automated algorithms. However, the DSC represents their ability to localise an area of a given aneurysm. For example, an algorithm maybe able to detect difficult aneurysms

with better precision but may localise them poorly resulting in lower DSC scores.

Table 4.4: Performances of the CNN architectures in localising the aneurysm voxels.

Method	**Sensitivity/Recall**	**Specificity**	**DSC**
U-Net11 [195]	0.398	0.999	0.645
U-Net16 [195]	0.42	0.999	0.613
ABUNet [195]	0.45	0.999	0.612
Attention U-Net [147]	0.36	0.999	0.649
U-Net [139]	0.314	0.999	0.646
Tiramisu [196]	0.18	0.999	0.16
Ladder [197]	0.45	0.999	0.58
Proposed Work	**0.45**	0.999	**0.652**

4.7.3 Ablation Study

Ablation study refers to analysing the neural network architectures through a gradual analysing the performance changes with new added modules/blocks. Our proposed method is built on a core U-Net architecture with the encode-decoder framework. We first incorporate the contextual encoding block into the U-Net architecture, and the patient-level sensitivity improves from 0.64 to 0.68 with an associated dice score of 0.352 to 0.39. Similarly, with the addition of locally important pooling in place of standard max-pooling operators, patient-level sensitivity improves from 0.68 to 0.92 with an associated dice score of 0.39 to 0.423. The increases in patient-level classification and localisation metric with the incorporation of the context module and locally important pooling show their contribution towards the final performance in the diagnosis of aneurysms from the CTA volumes.

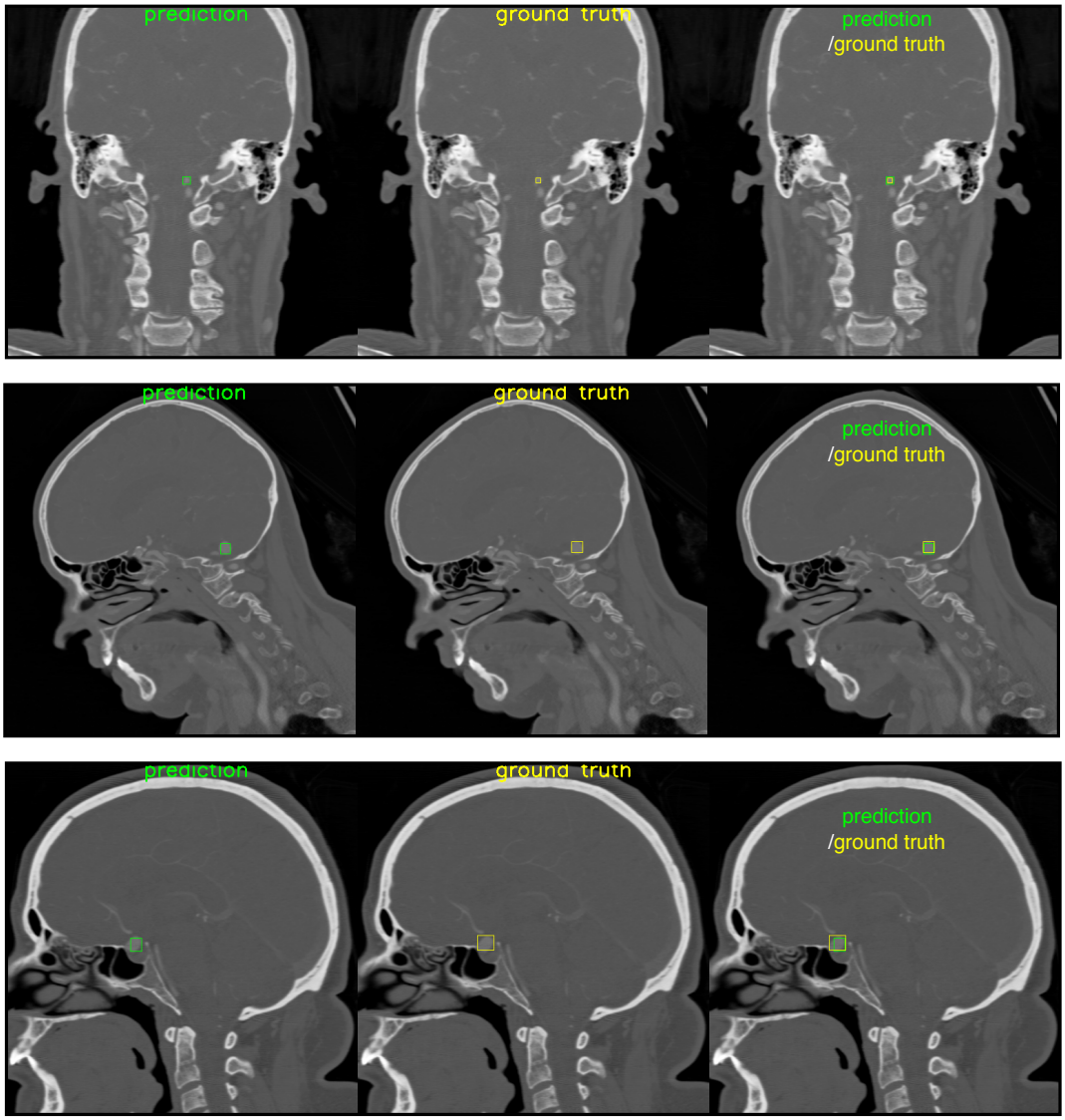


Figure 4.8: CTA slices showing the cerebral aneurysms detected and localised by the proposed method, in reference to the diagnosis and localisation by neuroradiologist.

4.8 Conclusion

Brain aneurysms are cerebrovascular diseases associated with high mortality/morbidity rates. Early diagnosis of unruptured aneurysms is essential to preventive treatments. CT angiograms are an effective and efficient imaging modality, which can detect small aneurysms with high sensitivity. It is a laborious and time-intensive task, however, to manually identify and localise aneurysms, not withstanding other

domain-related challenges. To this end, we proposed a novel CNN-based dense prediction architecture to label all the voxels in a CTA volume. A large-scale CTA dataset was constructed along with expert annotations to train the CNN architecture, in a supervised learning paradigm. Patient-level binary classification performances and localisation accuracies were reported to document the performances of the method. To the best of our knowledge, this is one of the first works to develop a CNN architecture for the detection and localisation of aneurysms from CTA images, concurrently carried out with [21].

Chapter 5

Analysing Arytenoid Cartilage Movements of Parkinson's Patients from Neck Computed Tomography Images

Parkinson's disease is a neurodegenerative disease, which can cause significant speech and voice impairment early its course. The objective of this work is to assess the speech abnormalities by identifying the abnormal vocal fold movements from neck CT images. First, an exploratory approach is proposed that utilises basic image processing techniques to extract clinically relevant feature points from the arytenoid cartilages, which support the vocal folds that are crucial for speech. It is demonstrated on a limited neck CT dataset. Subsequently, the limited dataset is augmented with additional CT scans of healthy controls to construct a large scale dataset, and a CNN-based object detector is proposed to localise the cartilage structures for further analysis of the vocal fold movements.

5.1 Introduction and Background

5.1.1 Parkinson's Disease

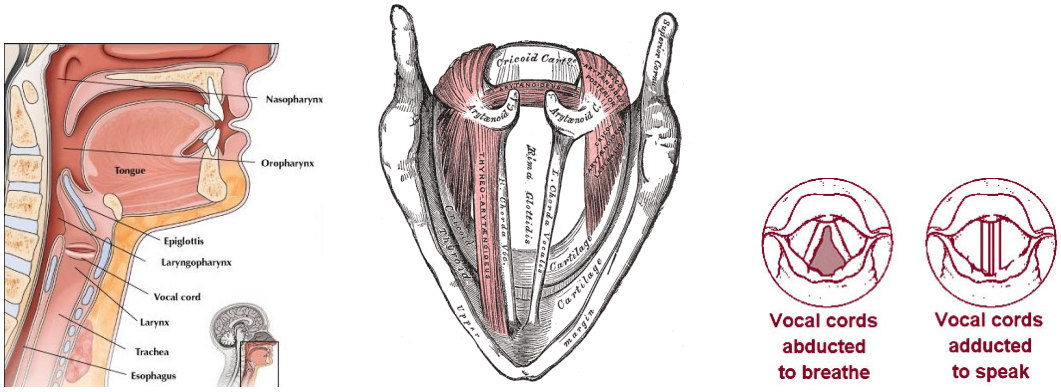
Parkinson's disease is a neurodegenerative condition that results in progressive degeneration of nerve cells. It is generally associated with depletion of dopamine in the substantia nigra. It is a hypokinetic movement disorder affecting mainly the motor system and resulting in the paucity of voluntary/involuntary movements. Parkin-

son's also has non-motor symptoms, including cognitive impairment and mood disorders. It is the second most common neurodegenerative disease that has an incidence rate of 2-3% amongst the population aged 65 or more [12] with an estimated 10 million patients in the world [13].

A critical feature of Parkinson's disease is speech impairment, which can be disabling for patients. During normal speech (phonation), vocal folds in the larynx come together (adduction) to restrict airflow and consequently produce sound (Figure 5.1). In Parkinson's disease, the regular movement of vocal folds can be impaired, leading to speech abnormalities. Poor voice quality (dysphonia), lower speech intensity (hypophonia), smaller range of articulatory movements (hypokinetic articulation), smaller pitch variability and range (dysprosodia), dysfluent speech [198–200] are some of the different forms of speech impairment reported in Parkinson's disease. In the early stages of illness, even when phonation is not overtly compromised, a small variability in pitch and loudness can be reported [15]. Perceptual voice changes (i.e., employing auditory perception to assess the voice characteristics) measured on the Unified Parkinson's Disease Rating Scale (UPDRS) were noted as far back as 9.8 years before diagnosis in [17]. Around 70-89% of the Parkinson's patients suffer from abnormal vocal fold movements [14–16]. Therefore, movement of vocal folds/speech quality may be a useful clinical feature for the early detection of Parkinson's.

5.1.2 Diagnosing the Speech Impairments

In addition to perceptual evaluation of speech quality, several other techniques can be used to analyse the dynamic movement of the vocal folds and associated structures. Electroglottography or laryngography is a non-invasive approach that can be used to measure the degree of contact between the vibrating vocal folds. The work in [202] uses laryngography and high-speed video (HSV) to study the vibration behaviour of the vocal folds in the abnormalities of phonation. Digital kymography, coupled with high-speed digital imaging, is another combination of techniques capturing vibration and vocal fold movements [203]. Laryngoscopy is an endoscopy of



(a) Neck Anatomy [201]. (b) Top view of the larynx [201]. (c) Abduction and Adduction [201].

Figure 5.1: Larynx, located in the anterior portion of the neck, constitutes vocal folds and cartilages. The three paired and three unpaired cartilages support the vocal fold movements during phonation and respiration.

the larynx that is carried out using a laryngoscope, which has a camera, directly or indirectly. Laryngeal electromyography is generally used to assess the dynamics of vocal folds and associated muscles/tissues [204]. Videolaryngostroboscopy is another widely used technique that allows direct imaging assessment of vocal fold dynamics [203]. A major drawback of most of the above-mentioned works is that they are invasive, causing significant patient discomfort. Additionally, they provide only surface visualisation of the vocal folds and fail to capture 3D movements. Further, most of them do not produce any objective measurement. Also, stroboscopic based techniques can be subjective as it relies heavily on operator skills, which can create significant inter- and intra-observer variance. Instead, non-invasive imaging techniques provide a better alternative to image the arytenoid cartilages and their low-frequency movements directly. Imaging modalities are mostly non-invasive and can capture a direct view of the desired region of interest with minimal human intervention.

Another category of the machine learning-based approaches to study analysis of the voice-related dysfunctions is based mostly on analysing the speech signals. An SVM-based classification algorithm is constructed in [205] to classify and analyse PD

patients and healthy controls. Six different articulatory aspects are described by extracting 13 features, which are used to train the classifier. A binary classification, as well as a multi-class classification amongst Parkinson's patients, is studied in [206]. Three different sets of features are extracted based on phonation, articulation, and prosody analysis, and subsequently employed to train the SVM. A neuro speech software is presented in [207] that analyses the speech recordings on different aspects and extracts various features to determine the associated abnormality and the progression of the disease. An automated approach to monitoring the neurological state of Parkinson's patients from their speech recordings is examined in [208], by investigating articulation measures and speech intelligibility. Deep neural network-based feature embeddings, known as x-vectors, are used to analyse the speech recordings from Parkinson's patients in [209].

5.1.3 Imaging the Vocal Fold Movements

Each vocal fold is attached anteriorly to the anterior commissure, which is a fixed point. Posteriorly, each vocal fold is attached to its respective arytenoid cartilage. The arytenoid cartilages are dynamic and control the movement of vocal folds during phonation and respiration. Like other movements in Parkinson's disease, the movement of the arytenoid cartilages can become disrupted. This can be a crucial feature of early vocal pathology in Parkinson's disease [200]. While an ultrasonic imaging technique has a high temporal resolution to capture the dynamic structures, it does not have the spatial resolution to delineate arytenoid structures and accurately analyse their movements. CT imaging has both the temporal and spatial resolution to capture the rapid fluctuations in vocal fold movements during phonation. A dynamic 320-slice CT enables real-time imaging of bodily structures with good spatial resolution. The laryngeal CT images acquired using a dynamic 320-slice CT are used in [200] to further analyse the movements of arytenoid cartilages and vocal fold dynamics. However, manual identification of arytenoids as in [210] remains highly subjective and an observer-dependent task, making it error-prone. Therefore, automated techniques would be useful to analyse CT images to provide reliable and

consistent scores across subjects and observers. The development of such automated techniques involves many challenges as phonation is a dynamic activity involving the constant movement of multiple structures within the image frame. Further, it involves the passage of air, which can also change the appearance of structures in the frame. The work in [211] studies the extraction of the upper airway from the cone-beam CT data, in the case of Obstructive sleep apnea (OSA) patients. In [212], the main objective is to segment the airway from the volume cone-beam CT data to compute the volumes of the airway. The application is to represent changes in the airway, before and after a medical procedure. Nevertheless, none of these works concentrate on the extraction of vocal folds. The work in [213] is the first known approach that aims to develop an automated technique to extract vocal fold planes from neck 3D CT images. However, it assumes that there is no coronal tilt during image acquisition, and hence the vertebral column is approximately orthogonal to the plane of vocal folds. The anterior intersection point of vocal folds is determined, and the orthogonal plane to the vertebral column passing through it is identified as the plane of vocal folds. However, it is possible that the vertebral columns are not always orthogonal to the vocal folds due to the anatomical diversity or due to a coronal tilt in the data. Besides, it does not provide any additional information apart from determining the location of AC.

5.2 Overview of the Proposed Approach

The objective of this work is to develop automated approaches to detect and analyse the movement of arytenoid cartilages from the CT images. We first develop an exploratory approach based on a combination of basic image processing techniques to identify clinically relevant feature points of the arytenoid cartilages. This simple, computationally inexpensive, unsupervised technique provides a mechanism to identify features from the arytenoid cartilages, which support the vocal folds that are crucial for speech. Subsequently, we construct a large scale CT image dataset consisting of healthy controls and Parkinson's patients that follow the voice examination

battery followed standard voice evaluation guidelines. It is further supplemented by CT examinations of healthy controls acquired during a regular breathing period. A supervised convolutional neural network-based object detector is then trained that can robustly detect and localise arytenoid cartilages from the CT images of neck.

5.3 Dataset

Subjects were recruited from the movement disorder clinic at the Monash Medical Center, Australia [210]. All the subjects were in the age group of 50-90 years, and some of them had Parkinson's disease for less than six years, while others did not have any neurological or laryngeal disorders. All the subjects were asked to make five short, and fast /i/ phonations and a 320-CT was used to image the entire neck. Resulting images were converted into NIfTI images with 512×512 pixels size and 12-bit grey levels with Right, Anterior, Superior (RAS) orientation. Subjects with blurred and distorted images were excluded.

5.4 Arytenoid Cartilage Feature Point Detection

The motivation of this work is to develop a rule-based approach to detect clinically relevant arytenoid cartilage feature points from the neck CT images. We define feature points to be the most anterior points on the arytenoid cartilages converging towards the airway. The feature definition is constructed, in consultation with the clinicians, as it is closely correlated to the vocal folds, and hence provides clinically useful information during phonation [210]. Steps involved in the detection of feature points are outlined in Figure 5.2 and are described in the following subsections.

5.4.1 Anterior Commissure Localisation

The primary initialisation step in the arytenoid cartilage feature detection involves locating the anterior commissure (AC), which is defined to be the junction of the two

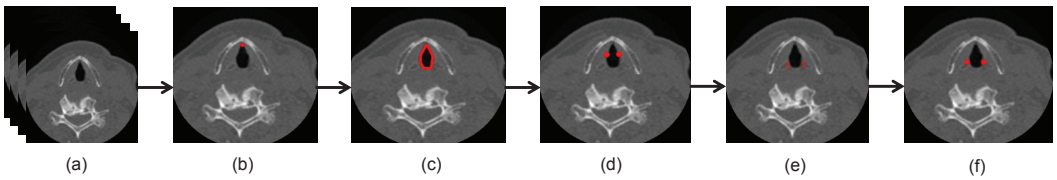


Figure 5.2: Flow diagram of the proposed work for the detection of feature points: (a) axial plane of the 3D CT Data, (b) localization of anterior commissure on the axial slice, (c) detecting the airway boundary, (d) cutoff pixels to perform filtering, (e) detecting potential feature points, (f) final feature points after post-processing.

vocal folds in the anterior portion of the larynx. The AC location acts as a starting point to scan for the arytenoid cartilages converging towards the airway, on its either side. We incorporate the approach developed in [213] to determine the AC location. Mid-sagittal plane image is extracted from the 3D CT data and used to compute a distance profile of the glottis from the posterior neck border. The farthest point of the glottis from the posterior neck border denotes AC. The anatomical constraints of larynx ensure that the pair of arytenoids cartilage are present close to the axial slice of AC in the caudal and cranial directions. Subsequently, axial slices on either side of the AC are scanned to locate the arytenoid cartilages and detect the feature points.

5.4.2 Airway Region Extraction

The axial slice image selected after the AC localisation is binarised by thresholding the grayscale image, using thresholds from a window of grayscale values. The lower and upper thresholds for the current dataset are -1200HU and -700HU, determined using Otsu's multi-thresholding technique [214] (HU represents the Hounsfield unit). An 8-neighbour connected component analysis is then carried out to determine the object regions present in the image, which are subsequently filtered out using anatomical constraints based rules to detect the airway. Some of the rules include:

- the airway region cannot be located too close to the boundary of the image

- width and height of the airway region cannot be more than 50% of the width and height of the image
- width and height of the airway region cannot be less than 2% of the width and height of the image

Moore-Neighbor Boundary Trace algorithm [215] is employed to trace and extract the airway boundary. The arytenoid cartilages are bounded by the thyroid cartilage on either side of the airway; therefore, the image containing the airway region bounded by the thyroid cartilages on either side is extracted for further processing.

5.4.3 Postprocessing

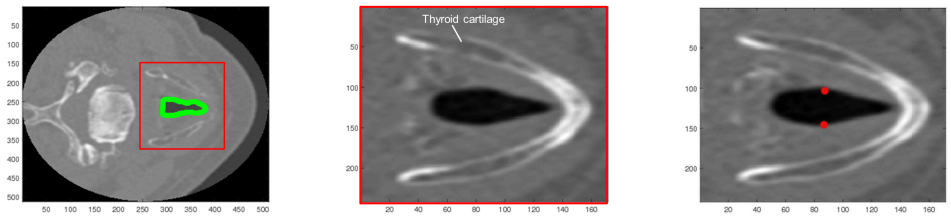
The boundary of the airway, shown in Figure 5.3, is divided into two halves, which come in contact with the two arytenoid cartilages on the left and right portions of the airway. Arytenoids are attached posteriorly to the vocal folds, and therefore, their movement is limited to posterior portions of the vocal folds during phonation. This constraint is incorporated to compute the cutoff locations (Figure 5.3c) beyond which arytenoids do not move. The pixels after the cutoff locations are filtered out. For the upper half of the airway (Figure 5.3a), cutoff pixel p satisfies the equation:

$$y(next(p)) - y(p) > 0, \tag{5.1}$$

where $y(p)$ is the y coordinate of pixel p and $next(p)$ is the pixel after p, when the airway boundary is scanned in clockwise direction. For the lower half of the airway, p satisfies $y(next(p)) - y(p) < 0$. The boundary pixels after p are ignored in subsequent processing.

5.4.4 Feature Point Detection

Feature point detection is carried out in the following two steps.



(a) (b) (c)

Figure 5.3: Figure shows (a) detected airway outline on an axial slice, (b) image bounded by the thyroid cartilages for subsequent processing, (c) cutoff pixels identified by clinicians to perform filtering.

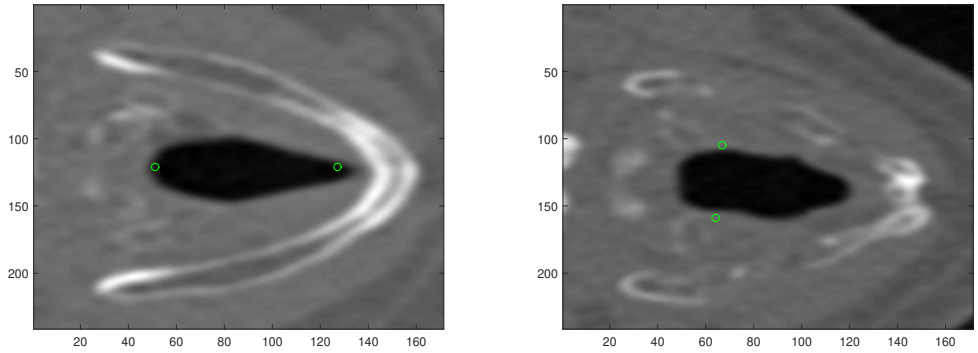
Step.1: Arytenoid cartilages are hyaline cartilages that acquire higher HU values in a CT image, resulting in slightly higher intensity image pixels. Therefore, distances of the airway boundary pixels from the first bright pixels in y direction, away from the airway are determined. For the current dataset, a bright pixel is empirically defined to be a pixel with an intensity of greater than 100HU. Let d be a function that computes city block distance between any two pixels. Let p_i be the i^{th} boundary pixel and B_i, corresponding bright pixel, then p_i is governed by:

$$p_i = \begin{cases} \text{potential feature point} & d(p_i, B_i) > T \\ \text{discarded} & d(p_i, B_i) \leq T, \end{cases} \tag{5.2}$$

where threshold T is empirically determined.

Step.2: The regions inside arytenoid cartilages are low attenuation areas, which acquire negative HU intensities on the CT image. Therefore, histogram of intensities is computed to determine the regions with the negative intensities. Let B^+ and B^- be the bins with positive and negative intensity edges, and let $N(B)$ be the number of pixels in a particular bin B. Then the i^{th} boundary pixel p_i is governed by:

$$p_i = \begin{cases} \text{potential feature point} & N(B_i^+) < N(B_i^-) \\ \text{discarded} & N(B_i^+) \geq N(B_i^-) \end{cases} \tag{5.3}$$



(a) (b)

Figure 5.4: Figure shows (a) reference pixels for clustering and (b) interest points of arytenoid cartilages.

5.4.5 Optimal Feature Point Detection

The pixels detected after the previous steps are grouped to form different clusters of feature points. Any two consecutive pixels separated by a distance of greater than two pixels in either the row or column directions are assigned to different clusters. The mean distances of all the points of a cluster, from both the references points(Figure 5.4a), are summed, and the cluster with minimum value is considered to best represent the arytenoid edge converging towards the airway. The centroid of this cluster gives the desired feature point, representing the anterior-most point on the arytenoid converging towards the airway. It can be a useful feature to get information about the movements of the arytenoids that support the vocal folds. The feature point detection can be carried out from the multiple axial slices, in which the arytenoids appear, which can be subsequently interpreted by the clinicians.

5.4.6 Results

The proposed approach was initially developed for a limited dataset and demonstrated on the CT data of 12 subjects. Ground truth feature points were generated with clinicians' support. Euclidean distance measure was used to compute the error between estimated feature point values and the ground truth. The implementations

were performed in MATLAB® 8.4 using the image processing toolbox on Windows 7 (64-bit system) equipped with an Intel® i7-4790 CPU running at 3.60 GHz. Table 5.1 shows the estimated feature point coordinates, ground truths and estimated errors using the proposed rule-based approach. Our interactions with clinicians resulted in an agreement to have an error tolerance of 15 pixels. From Table 5.1, we see that for estimating lower arytenoid feature points, the proposed approach produced a maximum error of 21.54 pixels for Subject 2, whereas the maximum error for upper arytenoid feature points was 21.8 pixels for Subject 6. Only two subjects (Subject 2 and 6) had error over 15-pixel tolerance for estimating the lower arytenoid feature points. On the other hand, only one subject (Subject 6) had error more than 15 pixels in estimating the upper arytenoid feature point. Therefore, the estimation accuracies are 83.34% and 91.67% for lower and upper arytenoid feature points, respectively. Figure 5.5 shows the feature points detected using the proposed algorithm for two CT examinations: Figure 5.5a and Figure 5.5b. From Figure 5.5, it can be observed that despite different airway width and cartilage appearances across the two images, the proposed approach was still able to detect the feature points reliably. This automated approach helps to identify initial basic feature points from the arytenoids useful to understand the arytenoid cartilages and their movements.

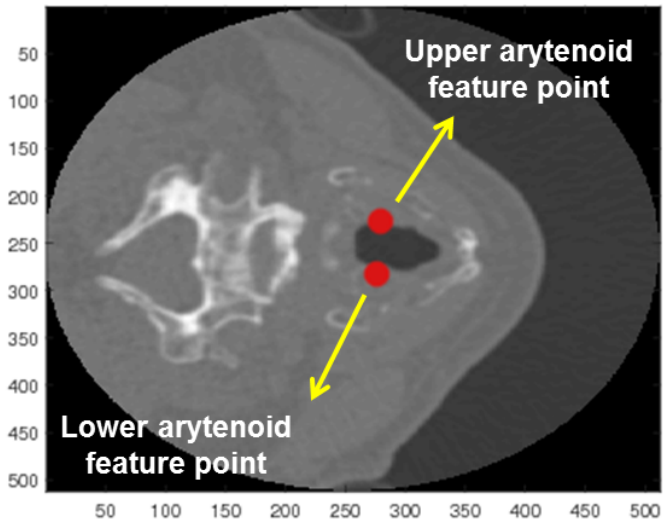
5.4.7 Limitations

It is an exploratory approach that is useful to get an initial knowledge of the arytenoids feature points. However, the accuracy of the detected feature is sensitive to accurate localisation of the anterior commissure. Besides, it may not generalise well to slightly different scanning protocols or images that may not be adequately aligned or to diverse neck anatomies. Further, it requires precise knowledge of the dataset to tune the hyperparameters. It is mainly a useful approach in the case of smaller datasets. Therefore, we develop and employ a CNN-based object detector to accurately localise the entire cartilages that form a more robust primary step to extract useful features. Further, neck CT scans of healthy controls obtained during

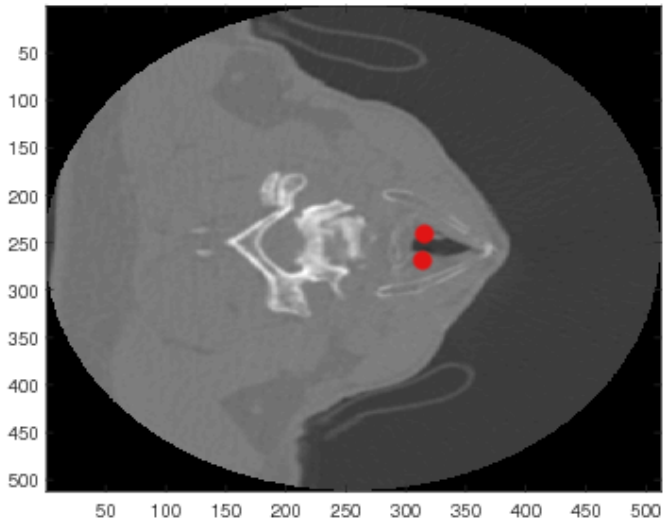
Table 5.1: Comparison of estimated feature point coordinates with ground truth for CT volumes. The lower and upper coordinates indicate the feature points detected on arytenoid cartilages (either side of the airway).

ID	**Lower Arytenoid Feature Points**			**Upper Arytenoid Feature Points**		
	Estimated [x, y, z]	Ground truth [x, y, z]	Error (pixels)	Estimated [x, y, z]	Ground truth [x, y, z]	Error (pixels)
1	293,317,76	290,310,76	7.61	234,323,76	236,314,76	9.21
2	274,324,92	266,304,92	21.54	222,320,92	225,306,92	14.3
3	265,280,52	263,279,52	2.23	230,293,52	229,287,52	6.08
4	276,392,63	274,389,63	3.60	244,390,63	243,390,63	1.00
5	274,333,69	274,330,69	3.0	232,332,69	230,329,69	3.6
6	266,300,42	284,296,42	18.43	286,302,42	265,296,42	21.8
7	266,311,53	264,312,53	2.23	233,311,53	231,312,53	2.23
8	266,314,56	264,309,56	5.38	238,314,56	236,313,56	2.23
9	280,275,30	277,266,30	9.48	226,278,30	225,269,30	9.05
10	300,247,38	294,240,38	9.21	249,253,38	246,248,38	5.83
11	284,339,52	283,333,52	6.08	249,332,52	247,328,52	4.47
12	267,315,66	265,312,66	3.60	242,316,66	242,314,66	2.0

the usual respiratory period are added to establish a large-scale neck CT dataset of 105 subjects. The large-scale CT dataset is used to train a supervised CNN object detector to localise the arytenoids, which delineates the arytenoids from the CT images for subsequent processing.



(a)



(b)

Figure 5.5: Feature points detected using the proposed algorithm for (a) CT examination: 1 (also shows the upper and lower arytenoid feature points) and (b) CT examination: 2, marked on the axial slice image of the larynx.

5.5 Localising the Arytenoid Cartilages using CNNs

A large-scale, diverse dataset is an essential component in developing a robust localisation approach using CNNs. To this end, we extend the initial dataset by adding more CT examinations of Parkinson's patients at an advanced stage of the disease. Further, CT scans of healthy controls, acquired during a regular breathing period, were also incorporated into the dataset to augment it with diverse imaging examinations. In addition to the dataset, supervised CNN training requires ground-truth annotations. The cartilages were localised and annotated from the neck CT scans independently by two trained observers. Open-source software ImageJ [191] was used to perform annotations on the axial views of the CT volumes. The cartilages were first identified and then localised by drawing a tight bounding rectangle around them such that the walls of the bounding boxes are tangential to its boundaries. The bounding boxes are parameterised by the left-lower coordinates, width, and height of the rectangle.

5.5.1 RCNNs in Object Localisation

Widely adopted and state-of-art CNN object detection methods use region-based CNN architectures that operate on the regions in images to perform classification and regression, producing output object regions along with confidence scores. A generic architecture of RCNNs is shown in Figure 5.6. The regions are usually parameterised in the form of bounding boxes, and several wide-ranging techniques are employed to generate object proposals that are subsequently fed into the CNN based detectors. Region proposal algorithms like selective search [105] are used at test time to compute region proposals with objectness scores, which are then passed through trained neural nets to extract proposal-specific features and perform object detection or localisation. The classification stages may employ classifiers ranging from class-specific linear support vector machines to a CNN classifier that could be trained end-to-end along with the feature extractor module.

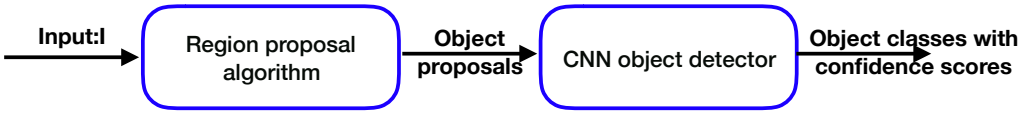


Figure 5.6: A generic RCNN architecture that has two essential components. The region proposal algorithms generate object proposals at test time, which are passed through CNN to extract the proposal-specific features and classify the object content in the proposal.

5.5.2 RCNNs with Region Proposal Networks

Faster RCNN architecture [108] is one such end-to-end trainable CNN-based object detector operating on regions in images. It is a two-stage architecture, as shown in Figure 5.7, that has a CNN based object proposal module generating object regions and a CNN based object-detection module that operates on these object proposals and produces object location coordinates with confidence scores.

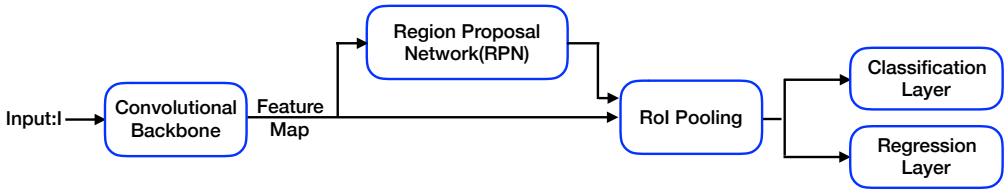


Figure 5.7: Faster RCNN is a two stage architecture that employs CNNs for object proposal and object detection modules. It is trained end-to-end to optimise the performances of both the CNN modules.

The role of object proposal algorithms is emulated by a region proposal network (RPN), an essential feature of the faster RCNN architecture that takes an image input and outputs object proposals and associated objectness scores (i.e. the probability of a proposal belonging to an object class). A fully convolutional backbone network operates on the input image to extract high-level features, which are then used by a small RPN sub-network. The RPN sub-network operates on the feature maps through sliding windows to extract high-dimensional feature vectors. Two sibling fully connected layers subsequently process the feature vectors: a regression layer and a classification layer that output the bounding box coordinates and corresponding objectness scores, respectively.

Reference boxes called anchors are incorporated at the region proposal stage allowing the parameterisation of a bounding box, with respect to reference boxes of multiple scales. The multiscale anchors allow the network to train and operate on images at multiple scales without an explicit construction of multiscale image pyramids. Therefore, K anchors at a sliding window location produce object proposals ($2K$ classification scores and $4K$ box coordinates). The reference ground-truths to train an RPN are constructed by assigning binary labels to anchors. An anchor is considered to be positive if (a) its overlap (in terms of IoU score) with any ground-truth is higher than a predefined upper threshold, (b) it has the highest overlap with a given ground-truth. A negative label is assigned to a non-positive anchor if the IoU overlap is less than a predefined lower threshold. A combination of classification and regression loss are used to construct a multitask loss used to train the RPN. A standard RCNN-based fast RCNN [107] is then combined with the region proposal network module to create two-stage architecture that consists of the RPN and detection network. The fully convolutional backbone network is shared between the RPN and detection network. Several training paradigms are employed in [108] to train the combination of RPN and the faster RCNN network, from joint training to approximate joint training of the two sub-networks to an alternating process that trains the faster RCNN and RPN in alternating steps. The alternative training process is employed in this implementation to train the entire two-stage architecture end-to-end effectively.

5.5.3 Cascade RCNN

As explained previously, the two-stage faster RCNN is trained end-to-end consisting of the region proposals generated by the RPN. An IoU threshold is used to determine positive and negative training samples for the detector. The overlap of ground-truth with the multiscale anchors is used to label whether each anchor is a positive training instance. However, IoU threshold needs to be set prior to the training and directly affects the overall detection performance of the architecture. A lower value generates a large pool of training samples; however, they can be noisy. A higher IoU threshold

may drastically improve the quality of training samples; however, it may lead to a lesser number of training samples reducing the training diversity and leading to overfitting. We implement a cascaded architecture of RCNN based on [216] that sequentially trains object detectors with an increasingly higher IoU threshold, as shown in Figure 5.8.

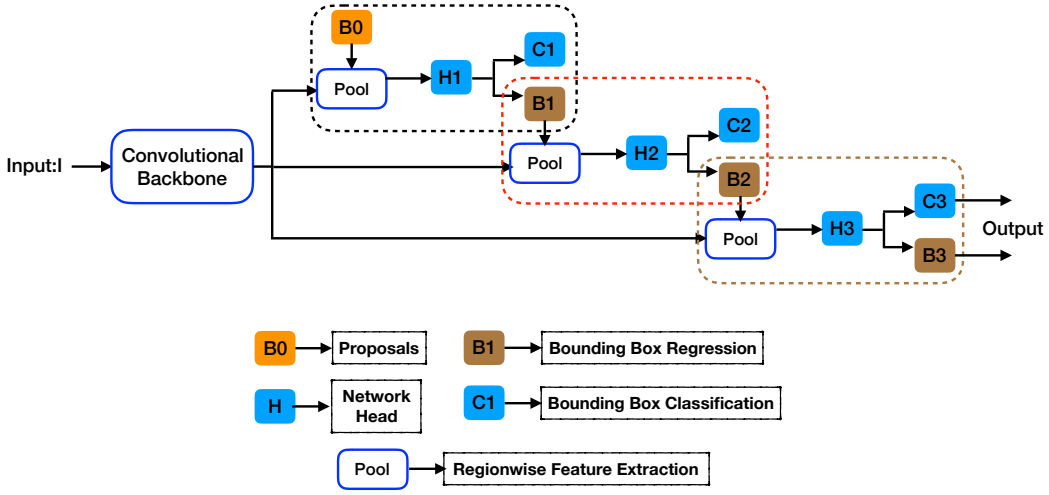


Figure 5.8: Cascade RCNN architecture. A cascade of object detectors is trained with increasing IoU thresholds. The quality of object proposals increases sequentially with later RCNN stages operating on higher quality object proposals.

In the cascade architecture, the detectors train to gradually operate on high-quality training samples without compromising on the generalisability of the detection performance. A sequence of detectors acts as a cascaded regression problem that is a cascade of specialised regressors optimised for the training sample distributions at the corresponding detector stage. The cascading architecture is employed during both the training and inference schemes to ensure there is no discrepancy between the two. Further, cascaded regression acts as a resampling technique generating different distributions of training samples for the cascaded stages. Sequentially employing detectors with increasing IoU thresholds ensures that detectors at a later stage operate on high-quality training samples; however, large amounts of training samples still exist across the stages preventing the cascaded network from overfitting. Each stage of the cascade architecture includes a classifier and regressor and is optimised using training samples generated using a preset IoU threshold that is

higher than IoU threshold for the preceding stage. The quality of detectors improves gradually during training, and similarly, quality of hypotheses improves gradually during inference. The higher quality detectors, therefore, operate on higher-quality region proposals leading to high-quality object detection.

5.5.4 Results

The objective of this work is to train and implement CNN based object detection pipeline to localise the arytenoid cartilages, which is subsequently useful to extract clinically relevant feature information from the arytenoid cartilages for a better understanding of the early onset of Parkinson's disease. The following subsections describe the preliminary data characteristics and ground-truth generation, implementation details, basic performance metrics, and the significance of the work.

Data Characteristics

The CT examinations acquired during the phonation experiment design and the CT examinations acquired during a breathing period are combined to construct a dataset of 100 CT examinations of 100 unique subjects. The CT scans are acquired at different time instants throughout the phonation/breathing period, and therefore, each CT examination consists of several neck CT volumes of a subject acquired at multiple time points. The number of CT volumes in a subject range from a minimum of 20 to a maximum of 63 with a mean of 40 and a standard deviation of 10. Table 5.2 lists some of the important preliminary data characteristics.

Annotations and Inter-rater agreement

It is laborious and time-intensive to generate ground-truth annotations for the CT volumes at all the time points in a subject. It is, however, necessary that the diversities of the vocal fold movements in a vocalisation period reflect in the final data analysis. Therefore, we use the CT volumes acquired at equally spaced time-instants throughout the phonation/breathing. For each subject, we count the total number

Table 5.2: Preliminary characteristics of the neck CT dataset.

Age	70.55 ± 7.3
Gender	38% Female
Slice thickness	0.5 mm
Slice spacing	0.44 mm
Total slices	80,160
Spatial dimensions	512×512
Healhty controls	26
Parkinson's patients (onset$\leq$ 6 yrs)	16
Parkinson's patients (onset$>$ 6 yrs)	14
Breathing scans (data augmentation)	41

of CT volumes at all the time points and divide it by 5 to get n. The trained observers then annotate every n^{th} CT volume by identifying the arytenoid cartilages and delineating them in the form of bounding rectangles. The inter-observer agreement measures are computed on the two independent sets of annotations in two forms: (i) identification agreement, and (ii) localisation agreement. The identification agreement describes the agreement in reliably identifying a slice containing the arytenoid cartilages. We assign a score of 1 if the arytenoid is identified in a slice and 0 otherwise. Kohen's kappa coefficient, which measures inter-rater agreement for categorical items between two observers, is used to determine the agreement between the two trained observers in identifying a slice with the cartilages. The scores are computed over every subject and then averaged over all the subjects to obtain the final averaged agreement measure. Kohen's kappa is computed as $\frac{\bar{P}-P_e}{1-P_e}$, where $\bar{P}$ is the average observed probability of agreement and P_e is the average chance agreement. The mean Kohen's kappa for the two sets of observers across all the patients is 0.832 with a standard deviation of 0.011, confirming a reliable identification of arytenoids from the neck CT images.

The localisation agreement is computed between the bounding boxes drawn by the

two independent observers. The observers localise the cartilages by drawing enclosing bounding rectangles around them. The localisation measure describes the agreement between the two observers in localising cartilages through bounding boxes. The amount of agreement between two bounding rectangles generated by two observers can be computed using the IoU measure. IoU measures the intersecting area between the two bounding rectangles relative to the combined area of the rectangles, thus measuring the amount of agreement between the two rectangles. IoU for any two bounding rectangles A and B is computed as: $IoU = intersection(A,B)/union(A,B)$. IoU measure for a CT volume is computed by aggregating the IoUs calculated over the set of slices for which there exist bounding rectangle measurements by both the observers. These slice level IoUs are then combined to obtain a per-patient measure, which is then averaged to obtain a final agreement score amongst the observers. We compute two agreement scores by aggregating the slices level scores in the two different approaches, similar to Chapter 4. An average type-1 IoU score averaged across the left, and right arytenoid cartilages is 0.62 with a variance of 0.009. Average type-2 IoU score averaged across the left, and right arytenoid cartilages is 0.84 with a variance of 0.009.

Development and validation

To train and develop the neural network methods, we used a Ubuntu 16.04 computing machine with the Nvidia GeForce GTX 1080Ti graphics card, which has 11 Gigabyte global memory. The implementation was done using MxNET library [217], and the simpledet library [218], in Python. Only the slices containing arytenoid cartilages were used for training. The images were normalised and cropped as a part of data preprocessing pipeline. Horizontal flipping was employed as a data augmentation technique to increase the diversity of the training set. Stochastic gradient descent optimisation algorithm was used to train the neural networks with a mini-batch size of two and a momentum of 0.9 for about 125K iterations. Multitask loss functions based on regression losses for the bounding boxes as well as the classification losses are optimised during the training, with an initial learning rate of 0.0006.

The CT examinations were split into percentages of 70/15/15 subsets for training, validation, and testing the implementations. The splits were done, ensuring that all the timepoints of a CT examination belonged to only one subset.

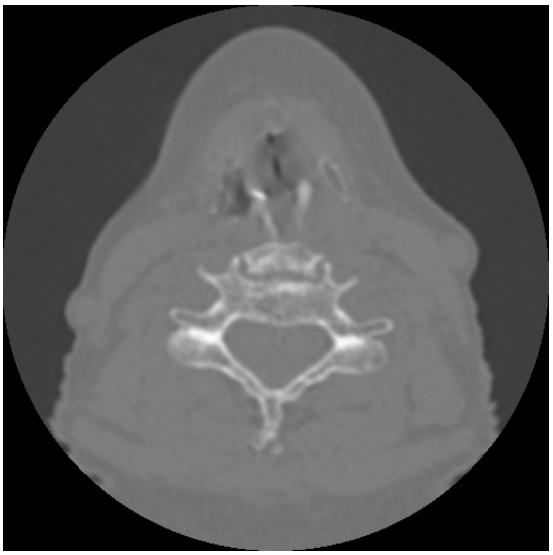
Performance Metrics

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F_1 = \frac{2*precision*recall}{precision+recall}$
- $IoU = \frac{TP}{TP+FN+FP}$

Four performance metrics were used to evaluate the implementations in localising arytenoid cartilages. Table 5.3 tabulates the performance metrics for the region based CNN object detectors. Faster RCNN [108] is a two-stage detector trained end-to-end that consists of an RPN and object detector network. Trident network [219] constructs multibranch architecture in which the branches share the parameters, however, operate at different receptive fields leading to scale specific feature maps. RetinaNet [220] is a one-stage object detector that employs a novel loss function, focal loss, to account for the extreme data imbalances. It can be seen that the faster RCNN and cascade RCNN achieve a similar recall value. Cascade RCNN, though, is slightly better in overall F1 score and the IoU score. Figure 5.9 shows the localised cartilages from the CT slices. Images on the left show the original slices and the images show the localised arytenoid cartilages along with the ground-truths.

Table 5.3: Performance metrics indicating the performances of CNN based object detectors in localising the arytenoid cartilages from the CT images slices containing the arytenoid cartilages.

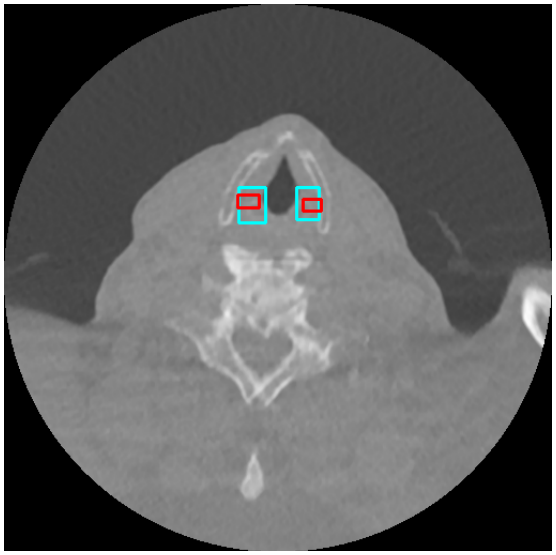
Method	**Precision**	**Recall**	**F_1 Score**	**IoU Score**
TridentNet [219]	0.841	0.938	0.886	0.387
FasterRCNN [108]	0.848	0.988	0.912	0.418
RetinaNet [220]	0.855	0.977	0.911	0.406
CascadeRCNN [216]	0.875	0.985	0.926	0.414



(a) Sample 1: CT slice containing arytenoid cartilages

(b) Sample 1: Localising the arytenoid cartilages from CT slices

(c) Sample 2: CT slice containing arytenoid cartilages



(d) Sample 2: Localising the arytenoid cartilages from CT slices

Figure 5.9: Localising the arytenoid cartilages in the form of bounding boxes using CNN based object detectors. The sky blue coloured bounding box indicates the ground truth and the red coloured bounding box is automatically detected.

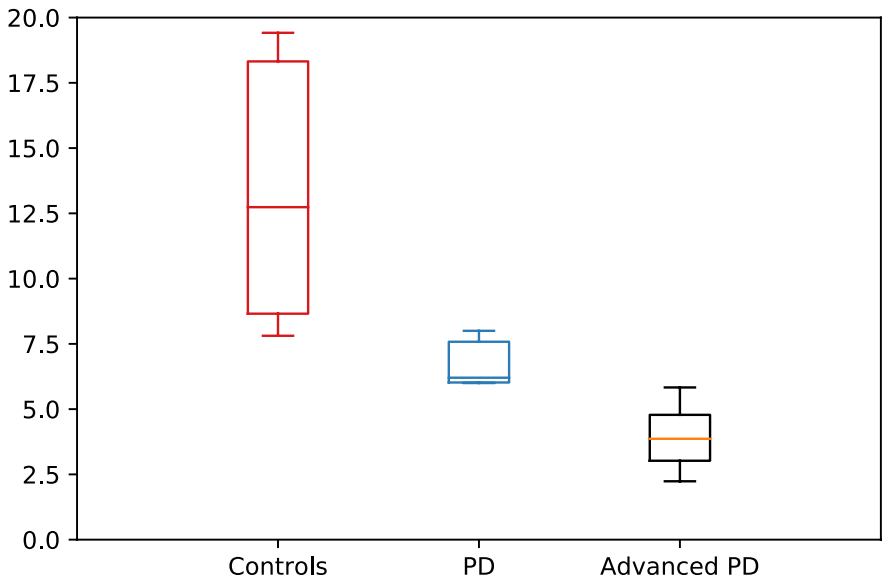


Figure 5.10: IAD for three sub-groups. The boxplots indicate the higher mean and variance in the case of healthy controls and the lower IAD variations in PD and advanced PD sub-groups.

5.5.5 Inter Arytenoid Distance

The automation of localising the arytenoid cartilages assists the clinicians to efficiently mine through CT scans and extract clinically relevant feature information for a better understanding of the data and associated diseases. A cost-efficient and straightforward measure such as distance between two arytenoid cartilages, called inter arytenoid distance (IAD), can be a piece of useful feature information [200]. We define the IAD to be the closest distance between two axis-parallel bounding rectangles enclosing the pair of arytenoid cartilages.

The current dataset consists of three distinct sub-groups of data - healthy controls without any neurological diseases, Parkinson's patients (PD) at an initial stage (≤ 6 years), and Parkinson's patients at an advanced stage(>6 years). We create a subset dataset by randomly sampling from the original dataset to create a dataset consisting of balanced sub-groups. The IAD measure is computed for five CT volumes acquired at equally spaced time points, and the absolute minimum IAD across the timepoints determines the final IAD measure for the subject. A vital hypothesis

from [200] is that the arytenoid cartilages do not move as expected with the progression of Parkinson's disease, leading to abnormal hypokinetic movement of the vocal folds. A similar observation can be made from Figure 5.10, highlighting the distinct differences in the distribution of IAD measures in the three sub-groups. Table 5.4 further clearly outlines the differences in IAD in terms of basic statistical measures. Wilcoxon's rank-sum test was used to determine the statistical significance of the differences in IAD measures for the sub-groups with p values of 0.05.

The differences in IAD values for the different sub-groups highlight the hypokinetic nature of the movements of arytenoid cartilages with the progression of Parkinson's disease. IAD can be a simple, effective measure that captures the for Parkinson's patients at different stages. Further, speech is one of the earliest affected functions before the onset of motor symptoms and therefore, imaging-based approaches to directly observe the cartilage movements and analyse their abnormalities will provide beneficial mechanisms towards detecting the Parkinson's disease at an earlier stage.

Table 5.4: IAD measures for the three sub-groups. The higher values of basic statistical measures for the healthy controls and lower measures for the disease groups can be observed.

Sub-group	**IAD mean**	**IAD mode**	**IAD median**	**IAD variance**
Healthy controls	16.15	7.81	12.73	96.50
PD	6.95	3.16	6.20	7.44
Advanced PD	3.93	2.23	3.86	1.49

5.6 Conclusion

A critical and early symptom of Parkinson's disease is speech impairment. The detection of speech abnormalities can be highly useful for early diagnosis. CT imaging is a non-invasive, and efficient diagnostic modality that helps in visualising and analysing movements of vocal folds and associated structures. This chapter proposed novel automated ways to extract clinically relevant information from the arytenoid cartilages that help in analysing the movement of vocal folds critical to phonation.

The first work provided an initial exploratory approach to identify feature points on arytenoid cartilages, which helps the user to consistently relevant points on the cartilages to analyse their movements. The dataset was subsequently augmented with CT examinations of healthy controls, and CNN-based object detection was developed to localise the cartilages in the form of bounding boxes. IAD measures were defined in terms of bounding rectangles and computed for three subgroups representing the healthy controls, Parkinson's patients at initial and advanced stages that highlighted differences amongst the sub-groups. Therefore, neck CT images during phonation can be used to derive measures that can be useful features towards identifying abnormal vocal fold movements at the onset of Parkinson's disease.

Chapter 6

Canonical Correlation Methods for Interpreting CNN Architectures

Interpretability in machine learning is an essential component assisting in a better understanding of the learning process to derive meaningful and easily understandable knowledge about a machine learning model designed for a particular task. This chapter presents novel interpretability methods that can be applied to CNN representations trained for the desired task.

6.1 Background and Introduction

Deep neural networks have been the modern machine learning architectures of choice due to their notable success in solving large-scale supervised learning tasks. Dramatic progress is witnessed in building powerful and robust neural network architectures that enable high performances in addressing several challenges such as image recognition, speech processing, and natural language understanding. They have transformed the world of pattern recognition with high-level abstractions of data through a series of linear and non-linear interacting layers that pave the way for constructing a hierarchical feature representation of the data directly from the data, instead of employing predefined handcrafted features. Despite their revolutionary success, neural networks are usually considered to be 'blackboxes' owing to their inability to generate semantically enriching knowledge representation that may provide greater understanding to the end user. Standard performance metrics like loss

function, accuracy, sensitivity-specificity, and others are useful indicators to analyse the performances of CNN architectures. However, solutions to CNN representations exist in a large space of solutions, and a gradient-based optimisation usually yields a local minimum irrespective of whether it leads to eventually optimum performance. The usual performance metrics, therefore, do not give a detailed view of the internal dynamics of the algorithm and its decision-making process. Dedicated techniques are vital that can explain the inner workings of a CNN architecture or more generally a machine learning model and decipher the complex learning representations of the model to create an end-user relevant knowledge.

The process of better understanding a machine learning model to decipher its representation into simple user understandable knowledge is usually known as *interpretability* or *explainability* of the model. There is no single formal definition for interpretability in ML that gives a comprehensive understanding of the interpretation process. It may refer to the creation of meaningful knowledge from the data or its model learned for the particular task, which is ultimately relevant to the end-user. It may involve an explainable decision-making process, creation of knowledge bank through meaningful feature representations, or generating easily understandable results. It can also be defined as the degree to which a human can understand the cause of a decision can consistently predict the model's result [221]. Further, interpretability enables testing of the causality of the features, their reliability in the decision-making process, and contributions to overall performance, in addition to assisting in the model debugging process. It is a rapidly emerging area with recent works exploring various dimensions of the interpretations under the topics of explainable ML, intelligible ML, or transparent ML, or more broadly, explainable AI. Interpretability in machine learning is important in:

- Creating meaningful knowledge representations that enrich and advance the human's knowledge.
- Creating a reliable and robust system that is easier to understand.

- Making the machine learning system more trustworthy and thereby, also increase safety in critical applications like medical machine learning, autonomous driving, and others.
- Developing meaningful systems through continuous and meaningful user feedbacks.
- Explaining the incorrect decisions and thereby, providing a detailed insight into the task for a better understanding.
- Creating a system that considers real-world biases, explains its decisions and thereby provides an opportunity for the user to understand what they are building.

Interpretability methods in machine learning can be broadly categorised into two groups: (i) Intrinsic and (ii) Post hoc [221]. Intrinsic interpretability refers to simple machine learning models that are intrinsically explainable such as decision trees, regression, sparse linear models. It is generally due to the simple structure of the model or the associated parsimonious nature that automatically explains the decision-making process and the final results. Post hoc methods, on the other hand, are external interpretability tools that are developed independently of the machine learning model and are applied to developed/trained machine learning models. The post hoc tools may be model-agnostic or model-specific and are developed independently of the model.

There are limited studies in the area of machine learning interpretability with no complete solution or a standard framework or even a standard definition for interpretation. Further, the modern deep neural networks are harder to study because of their size, and the distributed data-parallel nature of their training process. There have been, however, some interesting methods and developments that attempt to approach the overall problem by answering subsets of interpretability, addressed from different perspectives. In a broad sense, the interpretation of a model and the creation of a meaningful knowledge relevant to the end-user can be either a qualitative

approach or a quantitative one.

A qualitative approach usually involves some form of visualisation technique that allows for a visual assessment of the intermediate feature maps and inner working of a CNN. One of the earliest deep model visualisation techniques in [222] involved visualising the learned representations of a computation unit, in an arbitrary layer of a deep network. The techniques presented involved (i) finding a bounded norm pattern in the input image space that maximises the activation of a given computation unit, (ii) using a linear combination of the filters connected to characterise the filters in the higher layers. The approach in [4] visualised deep convolutional auto-encoders by reconstructing the input of each layer from its output using deconvolutional network (DeconvNet) architectures and established a high performance in the ImageNet classification challenge 2013 by proposing an architectural change to convolutional filters with smaller filter kernels. The visualisation approach in [222] is extended, in [223], to the visualisation of CNN architectures designed for classification. The first visualisation demonstrates the notion of a class learnt by CNN by generating an image that maximises the class score. It generates an image representing a specific class learned by the given trained CNN. It further demonstrates the visualisation of class-specific saliency maps describing the spatial support of a particular class in a given image. Reconstructing an image from its 'blackbox' feature representation is discussed in [224] that uses a local image descriptor SIFT to demonstrate the reconstruction of the image leading to a better understanding of the features. A similar inversion approach to reconstruct the images using only their feature representations is developed in [18] and is demonstrated on SIFT [225], HOG [226], and CNN representations. Quantitative approaches usually involve interpreting the CNNs using techniques that compute predefined quantitative metrics, which give a definitive view of the inner workings of CNN. It might involve controlling generalisation error using complexity measures adapted from statistical learning theory. Uniform stability [227–229], Rademacher complexity [230] and regularization are such measures used in practice to reduce generalization error. Pearson's correlation-based measure is incorporated in [231] to find alignment between neurons in a layer

within a network, or neurons across different networks. One-one, one-many, and many-many matching of neurons are employed to find the correlation similarity. It is further validated using entropy-based mutual information measures. Invariance, equivariance, and equivalence properties of the feature representations are studied in [232] to provide a better understanding of the hierarchical representations of neural nets and how the early shallow layers differ from the deeper layers and other such properties. Decision trees are used in [19] for quantitative interpretations of CNNs trained for image classifications tasks. A decision tree is learnt to semantically explain and interpret each prediction made by a pretrained CNN for each image category. The decision tree relates the object parts, corresponding filter activations, and their contributions to the overall prediction score. The trained neural net model is used to construct a soft decision tree in [233], which is used at test time to make explainable predictions. Decision trees are highly useful approaches that help in making hierarchical interpretable decisions. In [20], the compression techniques are studied towards interpreting the CNN representations. The neural network filters with visually redundant patterns are iteratively pruned based on corresponding classification accuracy reductions. The changes in the accuracies can be subsequently compared, and a compressed network results in a better interpretable CNN architecture. A detailed review of visualisation approaches to analyse and interpret the CNN representations is reported in [234]. Diagnosing the pretrained CNN representations, their disentanglement, and middle-to-end learning based on model interpretability are all documented. A graphical model-based interpretation is carried out in [235] by constructing a graph to represent the knowledge hierarchy inside a pretrained CNN model. Each graph node represents a pattern, and the edges represent the relationship between patterns, helping to disentangle the patterns from a CNN filter.

A canonical correlation analysis (CCA)-based population representation analysis is developed in [236] that proposes a singular value CCA method to compare layer representations within and between neural nets. A similar CCA based technique, known as projection weighted CCA is developed in [237] to provide a better distinction be-

tween the signal and noise of the feature representations during comparisons. The effect of learning rates, width of a network, and other such properties are analysed in the process.

6.2 Overview of the Proposed Work

The objective of this work is to develop automated post hoc machine learning interpretability methods that provide an efficient way to analyse the feature representations of a CNN architecture. We propose and implement computationally efficient CCA-based algorithms to analyse the latent CNN representations. CNN architectures consist of sequences of layers that learn hierarchically rich feature representations. An i^{th} convolutional layer $L_i(h_i \times w_i \times d_i)$ in a CNN consists of d_i feature maps of spatial dimensions $h_i \times w_i$. It can be organised in two broad ways: (i) as a collection of $h_i w_i d_i$ neurons, and (ii) as a collection of d_i filters, each consisting of $h_i w_i$ neurons sharing the weights.

We propose two kinds of CCA-based interpretability techniques to analyse layer representations by treating them as a collection of neurons and filter feature maps by treating layers as a collection of filter kernels. First, singular value CCA (SVCCA) [236] is extended to develop a novel generic singular value multiset CCA (SVMCCA) algorithm that is effective in comparing multiple-layer representations simultaneously by structuring them as large collections of neuron vectors. Further, a novel two-dimensional multiset CCA (2DMCCA) algorithm is proposed that is useful to analyse the 2D matrices directly. Subsequently, 2DMCCA is extended to two-dimensional SVMCCA (2DSVMCCA) to compare the 2D feature map representations of multiple filters within a CNN layer.

The following section first introduces CCA and MCCA before discussing the proposed SVMCCA and its applications. Subsequently, the proposed 2DMCCA and 2DSVMCCA and similarity analysis of 2D feature representations of the CNN filters are discussed. A CNN architecture is trained to detect and localise the cerebral

aneurysms from the CTA dataset described in Chapter 4 to demonstrate the applicability of the interpretability techniques.

6.3 Interpretability using Canonical Correlation Analysis

6.3.1 Singular Value Multiset Canonical Correlation Analysis

Canonical Correlation Analysis

CCA [238] is a data-driven technique that can effectively capture the linear relationship between two multivariate datasets. The primary objective of CCA is to determine the coordinate system, which represents the best possible linear relationship between the given multivariate datasets [238], by maximising their mutual correlation. It is defined as the problem of finding two sets of projection vectors a and b for the two sets of multivariate datasets X and Y, such that the correlation between the projection of X on a and projection of Y on b is maximised. Let X and Y be two multivariate datasets of dimensions $n \times d_1$ and $n \times d_2$, where n represents the number of data samples and d_1,d_2 are the data dimensionalities. The objective of CCA, then, is to determine projection vectors $a(d_1 \times 1)$ and $b(d_2 \times 1)$ such that the Pearson's correlation between canonical variables $U = Xa$ and $V = Yb$ is maximised. Pearson's correlation between U and V is defined as:

$$corr(U,V) = \frac{cov(U,V)}{\sqrt{var(U)}\sqrt{var(V)}} \tag{6.1}$$

where *cov* represents the covariance and *var* represents the variance. The optimisation problem usually is solved by converting it into a constrained optimisation problem with suitable constraints on the canonical variables. The constrained optimisation problem is solved by converting to unconstrained optimisation problem to obtain the first pair of canonical variables that are maximally correlated. The following pairs of canonical variables are computed such that they are uncorrelated to

the lower canonical variables, which ensures that the same relation is not captured repeatedly. CCA is an effective linear transform technique that captures the best correlation between datasets X and Y. However, it is simultaneously applicable only to two datasets, and in case of multiple datasets, it has to be repeatedly applied to pairs of datasets. Multiset canonical correlation analysis (MCCA) was proposed in [238] extending the application of CCA to three or more data sets.

Multiset Canonical Correlation Analysis

Kettenring [239] did seminal work in the development of MCCA by proposing a general framework that subsumes multiple extensions of MCCA. Five MCCA variations are discussed that have two common features:

- They all reduce to the two-set CCA theory [238] when the number of datasets is two.
- They all select one canonical variable from each dataset and optimise a function of their correlation matrix.

Let $\{X_i\}_{i=1}^m$ be m multivariate datasets, with $\{d_i\}_{i=1}^m$ number of variables and n number of samples and therefore, a dataset X_i is of dimensions n by d_i. Let $\{\Sigma_{ij} \in \Re^{d_i \times d_j}\}_{i,j=1}^m$ represent the covariance matrix between datasets X_i and X_j. Let $A_s = (a_{s1}, ..., a_{sm})$ be the canonical transforms of the m datasets at stage s, where $s \leq min(rank(X_1), ..., rank(X_m))$. Then a general MCCA objective can be written as:

$$\begin{aligned} &\underset{a_{s1},...,a_{sm}}{\arg\max} \sum_{i=1}^{m} \sum_{j=1}^{m} g(a_{si}' X_i (a_{sj}' X_j)') \\ &i.e., \underset{a_{s1},...,a_{sm}}{\arg\max} \sum_{i=1}^{m} \sum_{j=1}^{m} g(a_{si}' \Sigma_{ij} a_{sj}) \end{aligned} \tag{6.2}$$

The definition of function g in (6.2) leads to different objectives. The function $g(x) = x$ corresponds to the sum of the correlations model and $g(x) = x^2$ corresponds to the sum of the squared correlations model. Other objectives include maximising

the variance of overall correlations, minimising the residual variance of the overall covariance.

Objective. We employ a sum of correlations model that maximises the sum of overall correlations of the transformed datasets, as given in (6.3).

$$\underset{a_{s1},\ldots,a_{sm}}{\arg\max} \sum_{i=1}^{m}\sum_{j=1}^{m} a'_{si}\Sigma_{ij}a_{sj} \tag{6.3}$$

Constraint. Different constraints can be applied on the magnitudes or variances of canonical variables towards formulating a constrained optimisation problem. We incorporate the following constraint in (6.4) on the sum of variances of the transformed datasets.

$$\sum_{i=1}^{m} a'_{si}\Sigma_{ii}a_{si} = 1 \tag{6.4}$$

Constrained Optimisation Problem. The objective (6.3) is combined with constraint (6.4) to formulate a constrained optimisation problem as:

$$\underset{a_{s1},\ldots,a_{sm}}{\arg\max} \sum_{i=1}^{m}\sum_{j=1}^{m} a'_{si}\Sigma_{ij}a_{sj} \text{ s.t. } \sum_{i=1}^{m} a'_{si}\Sigma_{ii}a_{si} = 1 \tag{6.5}$$

Solution. The constrained optimisation problem is then converted into unconstrained optimisation problem using the method of Lagrange multipliers as:

$$L = \sum_{i=1}^{m}\sum_{j=1}^{m} a'_{si}\Sigma_{ij}a_{sj} - \frac{\lambda_1}{2}(\sum_{i=1}^{m} a'_{si}\Sigma_{ii}a_{si} - 1) \tag{6.6}$$

where, λ_1 is a Lagrange multiplier. The objective is to compute a_{si} in (6.6) such that the value of L is maximised. This can be achieved by computing $\frac{\partial L}{\partial a_{si}} = 0$ for i=1, ..., m that leads to,

$$\sum_{j=1}^{m} \Sigma_{ij}a_{sj} = \lambda_1 a_{si} \quad \text{for i=1, ..., m} \tag{6.7}$$

i.e.,

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} & ... & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & ... & \Sigma_{2m} \\ \vdots & \vdots & ... & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & ... & \Sigma_{mm} \end{pmatrix} \begin{pmatrix} a_{s1} \\ a_{s2} \\ \vdots \\ a_{sm} \end{pmatrix} = \lambda_1 \begin{pmatrix} \Sigma_{11} & 0 & ... & 0 \\ 0 & \Sigma_{22} & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \Sigma_{mm} \end{pmatrix} \begin{pmatrix} a_{s1} \\ a_{s2} \\ \vdots \\ a_{sm} \end{pmatrix} \tag{6.8}$$

The solution in (6.8) represents generalised singular value decomposition (gSVD) of the overall covariance matrix of the m transformed datasets. The generalised eigenvector A_s is divided into successive blocks of $d_1,...,d_m$ to obtain the projection transforms $(a_{s1},...,a_{sm})$ of the m data-sets.

Singular Value Multiset Canonical Correlation Analysis

In the first formulation, an i^{th} CNN layer $l_i(h_i \times w_i \times d_i)$ can be thought of as a collection of $h_i w_i d_i$ neurons. A neuron z_i represents non-linear scalar output for a given training sample and its output over a set of N training examples is a N dimensional vector. Therefore, the layer l_i is a collection of $h_i w_i d_i$ neuron vectors (N-dimensional) in R^N, and the layer is a subspace of R^N spanned by its neuron vectors. The neurons within a neighbourhood $h_i \times w_i$ that share the weights see different patches of images between them, and therefore, a convolutional layer $l_i(h_i \times w_i \times d_i)$ can be reshaped to form a multivariate dataset $X_i(Nh_i w_i \times d_i)$. To analyse the relationship amongst m layer representations $\{l_i\}_{i=1}^m$, corresponding m multivariate datasets $\{X_i\}_{i=1}^m$ are constructed. However, not all the layer channels contribute equally, and there may exist several dimensions that are noise with a small contribution. CCA fails to account for the data noise and may not be able to capture layer-relationships effectively. SVD is used as a preprocessing step that constructs a low-rank matrix approximation of the dataset X_i. In summary, the SVD of X_i involves finding eigenvectors corresponding to top k eigenvalues from X_i that capture maximum variation in the data and determine subspace datasets X_i'. The multivariate subspace datasets X_i' are then linearly transformed using MCCA to maximise the overall sum of their correlations. The proposed SVMCCA algorithm operates in two steps by first removing the noisy

layer directions using SVD and subsequently, employing MCCA to capture the best possible linear relationship between the m layer representations by finding the m canonical transforms $\{A_i\}_{i=1}^{m}$ for the m datasets $\{X_i'\}_{i=1}^{m}$ that represent the feature representations of the m layers $\{l_i\}_{i=1}^{m}$. The eigenvalues in 6.8 represent the group correlation similarities $\{\rho_1, \rho_2, ..., \rho_k\}$. In order to quantitatively compare the m layer representations, we define an average SVMCCA group similarity score for the group of m layers as:

$$\rho^{SVMCCA} = \frac{1}{k}\sum_{i=1}^{k}\rho_i \tag{6.9}$$

The group similarity measure ρ^{SVMCCA} represents an overall maximised correlation between the m layer representations. In addition to this group similarity measure, it is necessary to separately analyse the similarities between pairs of layers to determine their respective mutual relationships. The canonical transforms $\{A_i\}_{i=1}^{m}$ are used to project $\{X_i'\}_{i=1}^{m}$ and construct canonical variables $\{\tilde{X}_i\}_{i=1}^{m}$, which are used to determine the mutual similarity between i^{th} layer representation and j^{th} layer representation as:

$$\rho_{ij}^{SVMCCA} = \frac{1}{k}\sum_{1}^{k} corr(\tilde{X}_i, \tilde{X}_j) \tag{6.10}$$

SVMCCA compares layers across their channel dimensions by treating the parameter sharing neurons of a particular direction as additional data points. In a given layer channel, a 2D array of neuron weights are shared across the entire spatial extent of the input image to extract localised spatial features. It represents the convolution operation by a filter kernel that operates on a receptive field of size $K \times K$ and constructs feature maps. In other words, a 2D filter learns to extract a feature of a certain kind in the given direction. Therefore, it would be worthwhile to investigate the similarities between the feature map representations produced by different directions, which would provide a rich view inside layer representations. In this regard, the following section outlines the proposed 2DMCCA and 2DSVMCCA that operate on the 2D feature maps directly to analyse and compare their representations.

6.3.2 Two Dimensional Singular Value Multiset Canonical Correlation Analysis

The previously discussed SVMCCA can be employed to analyse the 2D feature map representations by vectorising them. However, vectorised data does not consider the spatial structure of feature maps. In the case of increased dimensionality of the data, it may also lead to a small sample size problem due to inadequate data samples, which may make the covariance matrix to be ill-conditioned, and make the solution unstable or non-existent. Additionally, it increases the computational complexity of the algorithm. Therefore, we propose a novel 2DMCCA algorithm that operates on the feature maps directly and subsequently extend it to 2DSVMCCA to analyse feature map representation of the convolutional filters.

Two Dimensional Multiset Canonical Correlation Analysis

Consider m matrix data sets $\{X_i\}_{i=1}^{m}$, where i^{th} dataset X_i consists of N samples with dimensions $p_i \times q_i$ ($\{X_{it} \in \Re^{p_i \times q_i}\}_{t=1}^{N}$). Dataset X_i is of dimensions $N \times p_i \times q_i$. Left linear transforms $\{l_i\}_{i=1}^{m}$ and right linear transforms $\{r_i\}_{i=1}^{m}$ are defined similar to [240], for the m sets of data, that operate along the rows and columns of matrices respectively.

Objective. The objective of 2DMCCA is to compute left transforms $\{l_i\}_{i=1}^{m}$ and right transforms $\{r_i\}_{i=1}^{m}$ such that overall correlation of transformed datasets is maximised. Similar to MCCA, we formulate a sum of correlations model that maximises the sum of overall correlations of the subspace datasets transformed using left and right transforms.

$$\begin{aligned} &\underset{\substack{l_1,\ldots,l_m \\ r_1,\ldots r_m}}{\arg\max} \sum_{i=1}^{m}\sum_{j=1}^{m} corr(l_i'\widetilde{X}_i r_i, l_j'\widetilde{X}_j r_j) \\ i.e., &\underset{\substack{l_1,\ldots,l_m \\ r_1,\ldots r_m}}{\arg\max} \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{cov(l_i'\widetilde{X}_i r_i, l_j'\widetilde{X}_j r_j)}{\sqrt{var(l_i'\widetilde{X}_i r_i)}\sqrt{var(l_j'\widetilde{X}_j r_j)}} \end{aligned} \tag{6.11}$$

The covariance between transformed datasets $l_i'\widetilde{X}_i r_i$ and $l_j'\widetilde{X}_j r_j$ can be computed in

two ways as:

$$cov(l_i'\widetilde{X}_i r_i, l_j'\widetilde{X}_j r_j) = \frac{1}{N}\sum_{t=1}^{N} l_i'\widetilde{X_{it}} r_i r_j'\widetilde{X_{jt}}' l_j \tag{6.12}$$

$$cov(r_i'\widetilde{X}_i' l_i, r_j'\widetilde{X}_j' l_j) = \frac{1}{N}\sum_{t=1}^{N} r_i'\widetilde{X_{it}}' l_i l_j'\widetilde{X_{jt}} r_j \tag{6.13}$$

Right covariance matrix and left covariance matrix are then defined between datasets X_i and X_j as:

$$\Sigma_{ij}^r = \frac{1}{N}\sum_{t=1}^{N} \widetilde{X_{it}} r_i r_j' \widetilde{X_{jt}'} \tag{6.14}$$

$$\Sigma_{ij}^l = \frac{1}{N}\sum_{t=1}^{N} \widetilde{X_{it}'} l_i l_j' \widetilde{X_{jt}} \tag{6.15}$$

The covariance matrix computations in (6.12) and (6.13) can be restructured using the definitions in (6.14) and (6.15) as:

$$cov(l_i'\widetilde{X}_i r_i, l_j'\widetilde{X}_j r_j) = l_i'\Sigma_{ij}^r l_j \tag{6.16}$$

$$cov(r_i'\widetilde{X}_i' l_i, r_j'\widetilde{X}_j' l_j) = r_i'\Sigma_{ij}^l r_j \tag{6.17}$$

Constraints. The objective function (6.11) is invariant with respect to the scaling of individual variances of the transformed datasets. Therefore, several constraints can be incorporated on variances of the transformed datasets. The overall sum of variances of the transformed datasets is constrained to unity, as given in (6.18) and (6.19).

$$\sum_{i=1}^{m} l_i'\Sigma_{ii}^r l_i = 1 \tag{6.18}$$

$$\sum_{i=1}^{m} r_i'\Sigma_{ii}^l r_i = 1 \tag{6.19}$$

Constrained Optimisation Problem. The constraints in (6.18) and (6.19) and the covariance matrix definitions in (6.16) and (6.17) can be combined to formulate the constrained optimisation problems.

$$\underset{l_1,...,l_m}{\arg\max} \sum_{i=1}^{m}\sum_{j=1}^{m} l_i'\Sigma_{ij}^r l_j \text{ s.t. } \sum_{i=1}^{m} l_i'\Sigma_{ii}^r l_i = 1 \text{ (fixed } r_1,...,r_m) \tag{6.20}$$

$$\underset{r_1,...,r_m}{\arg\max} \sum_{i=1}^{m}\sum_{j=1}^{m} r_i'\Sigma_{ij}^l r_j \text{ s.t. } \sum_{i=1}^{m} r_i'\Sigma_{ii}^l r_i = 1 \text{ (fixed } l_1,...,l_m) \tag{6.21}$$

Solution. The constrained optimization problem 1 in (6.20) is then converted into unconstrained optimization problem using the method of Lagrange multipliers as:

$$L = \sum_{i=1}^{m}\sum_{j=1}^{m} l_i' \Sigma_{ij}^r l_j - \frac{\lambda_1}{2}(\sum_{i=1}^{m} l_i' \Sigma_{ij}^r l_i - 1) \tag{6.22}$$

$$L = \sum_{i=1}^{m}\sum_{j=1}^{m} r_i' \Sigma_{ij}^l r_j - \frac{\lambda_2}{2}(\sum_{i=1}^{m} r_i' \Sigma_{ij}^l r_i - 1) \tag{6.23}$$

where, λ_1 and λ_2 are the Lagrange multipliers.

The objective is to compute l_i in (6.22) (or equivalently r_i in (6.23)) such that the value of L is maximized. This can be achieved by setting $\frac{\partial L}{\partial l_i} = 0$ (or equivalently $\frac{\partial L}{\partial r_i} = 0$) for i=1, ..., m.

$\frac{\partial L}{\partial l_i} = 0$ gives,

$$\sum_{j=1}^{m} \Sigma_{ij}^r l_j = \lambda_1 \Sigma_{ii}^r l_i \quad \text{for i=1, ..., m} \tag{6.24}$$

i.e.,

$$\begin{pmatrix} \Sigma_{11}^r & \Sigma_{12}^r & ... & \Sigma_{1m}^r \\ \Sigma_{21}^r & \Sigma_{22}^r & ... & \Sigma_{2m}^r \\ \vdots & \vdots & ... & \vdots \\ \Sigma_{m1}^r & \Sigma_{m2}^r & ... & \Sigma_{mm}^r \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{pmatrix} = \lambda_1 \begin{pmatrix} \Sigma_{11}^r & 0 & ... & 0 \\ 0 & \Sigma_{22}^r & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \Sigma_{mm}^r \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{pmatrix} \tag{6.25}$$

$\frac{\partial L}{\partial r_i} = 0$ gives,

$$\sum_{j=1}^{m} \Sigma_{ij}^l r_j = \lambda_2 \Sigma_{ii}^l r_i \quad \text{for i=1, ..., m} \tag{6.26}$$

i.e.,

$$\begin{pmatrix} \Sigma_{11}^l & \Sigma_{12}^l & ... & \Sigma_{1m}^l \\ \Sigma_{21}^l & \Sigma_{22}^l & ... & \Sigma_{2m}^l \\ \vdots & \vdots & ... & \vdots \\ \Sigma_{m1}^l & \Sigma_{m2}^l & ... & \Sigma_{mm}^l \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix} = \lambda_2 \begin{pmatrix} \Sigma_{11}^l & 0 & ... & 0 \\ 0 & \Sigma_{22}^l & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \Sigma_{mm}^l \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix} \tag{6.27}$$

The solution in (6.25) represents gSVD of the right covariance matrix. By assum-

ing the right linear transforms $\{r_i\}_{i=1}^{m}$, right covariance matrices can be computed. These are used to compute SVD of the above system to obtain the eigenvector. This is divided into successive blocks of $p_1,...,p_m$ to obtain the left transforms of the m datasets. The left transforms can then be used to construct the left covariance matrix, which can be decomposed according to SVD as given in (6.27). This eigenvector is then divided into successive blocks of $q_1,...,q_m$ to obtain the right transforms of the m datasets. This process is repeated until the right linear transform, and the left linear transforms converge. This procedure can also be carried out by first assuming the left linear transform to be fixed. Algorithm 1 outlines the detailed steps for the 2DMCCA computation.

Two Dimensional Singular Value Multiset Canonical Correlation Analysis

As described previously, $L_i(N \times h_i \times w_i \times d_i)$ represents the response of an i^{th} convolutional layer of dimensions $h_i \times w_i \times d_i$ to N training images. The layer $L_i(N \times h_i \times w_i \times d_i)$ can also be represented as a collection of $K \times K$ filters $\{F_j(N \times h_i \times w_i)\}_{j=1}^{d_i}$. That is, $F_j(N \times h_i \times w_i)$, in turn, represents the response of a filter, in a channel d_i, on N training images producing $h_i \times w_i$ dimensional feature maps. The objective of 2DSVMCCA is to directly compare the similarities between feature map representations $\{F_j\}_{j=1}^{d_i}$ by finding their projected subspaces using 2DSVMCCA.

The feature map $h_i \times w_i$ may consist of neuron activations with small noisy magnitudes that impact the signal-to-noise ratio of the feature map. In the first step of 2DSVMCCA, two-dimensional singular value decomposition (2DSVD) [241] is employed that determines the lower dimensional subspaces that capture maximum variations present in the feature map data. The 2DSVD determines $(U_k, V_s, \{M_i\}_{i=1}^{N})$ for the feature map dataset $F_j(h_i \times w_i \times N)$, wherein U_k and V_s provide the two common subspaces to project the original feature map datasets to, along the row and column dimensions. Using 2DSVD, the feature map datasets $\{F_j\}_{j=1}^{m}$ are projected onto the subspaces to construct the subspace datasets $\{F_j'\}_{j=1}^{m}$ that capture maximum vairations in the data. The K most important directions along the row and

Algorithm 1
2DMCCA Algorithm

Procedure: 2DMCCA
Input:
Image data, $\{X_{it} \in \Re^{p_i \times q_i}\}_{t=1}^{N}$ for $i = 1, ..., m$.
Number of left and right canonical variates, $d_1 \times d_2$.
Output:
Left transforms, $\{L_i \in \Re^{p_i \times d_1}\}_{i=1}^{m}$.
Right transforms, $\{R_i \in \Re^{q_i \times d_2}\}_{i=1}^{m}$.
Procedure:
Center the image data $\{X_i\}_{i=1}^{m}$ to get $\{\widetilde{X}_i\}_{i=1}^{m}$.
Initialize the right transforms, $\{r_i\}_{i=1}^{m}$.
Repeat:
Compute the right covariance matrices, $\{\Sigma_{ij}^{r}\}_{i,j=1}^{m}$ to construct the matrix,

$$\Sigma^{r} = \begin{pmatrix} \Sigma_{11}^{r} & \Sigma_{12}^{r} & ... & \Sigma_{1m}^{r} \\ \Sigma_{21}^{r} & \Sigma_{22}^{r} & ... & \Sigma_{2m}^{r} \\ \vdots & \vdots & ... & \vdots \\ \Sigma_{m1}^{r} & \Sigma_{m2}^{r} & ... & \Sigma_{mm}^{r} \end{pmatrix}$$

Compute the gSVD of Σ_r to get the d_1 largest eigenvectors.
Divide the d_1 largest eigenvectors into successive blocks of sizes $p_1, ..., p_m$ to get d_1 left transforms for the m datasets.
Use the previously computed left transforms to compute left covariance matrices, $\{\Sigma_{ij}^{l}\}_{i,j=1}^{m}$ and hence, to construct the matrix,

$$\Sigma^{l} = \begin{pmatrix} \Sigma_{11}^{l} & \Sigma_{12}^{l} & ... & \Sigma_{1m}^{l} \\ \Sigma_{21}^{l} & \Sigma_{22}^{l} & ... & \Sigma_{2m}^{l} \\ \vdots & \vdots & ... & \vdots \\ \Sigma_{m1}^{l} & \Sigma_{m2}^{l} & ... & \Sigma_{mm}^{l} \end{pmatrix}$$

Compute the gSVD of Σ_l to get the d_2 largest eigenvectors.
Divide the d_2 largest eigenvectors into successive blocks of sizes $q_1, ..., q_m$ to get d_2 right transforms for the m datasets.
Until the left and right transforms converge

column directions are considered to construct $\{F'_j\}_{j=1}^{m}$ of size $K \times K \times N$. In the second step, the m subspace datasets are processed by 2DMCCA to determine the best aligned directions along the rows and columns of the feature maps and maximise their correlation. The canonical left linear transforms $\{l_i\}_{i=1}^{m}$ and right linear transforms $\{r_i\}_{i=1}^{m}$ are computed that construct projected subspaces $\{\tilde{F}_j\}_{j=1}^{m}$ that maximise overall sum of the correlations between the m datasets. The eigenvalues of the right and left covariance matrices represent the left group correlation similarity $\rho_l = (\rho_{l1}, \rho_{l2}, ..., \rho_{lk}$ and right group correlation similarity $\rho_r = (\rho_{r1}, \rho_{r2}, ..., \rho_{rk}$, which converge after the iterative optimisation process outlined in Algorithm 1.

We define overall group similarity score for a group of m filter feature map datasets $\{F_j\}_{j=1}^{m}$ as:

$$\rho^{2DSVMCCA} = \frac{1}{2k} \sum_{i=1}^{k} (\rho_{li} + \rho_{ri}) \tag{6.28}$$

The mutual similarity score between i^{th} and j^{th} filters datasets F_i and F_j is computed as:

$$\rho_{ij}^{2DSVMCCA} = \frac{1}{2k} \sum_{i=1}^{k} corr(\tilde{F}_i, \tilde{F}_j) \tag{6.29}$$

6.4 Results

To demonstrate the applications of the proposed interpretability methods, we train a simple encoder-decoder based UNet architecture [139] for the dense prediction of cerebral aneurysms from the CTA images. The objective of the task and dataset are the same as described in chapter 4. Positive aneurysm slices from all the CTA volumes are split in the ratio of 70/15/15 subsets for constructing the training/validation/test subsets. The network is trained to optimise a combination of dice loss and cross-entropy loss, and the optimisation is carried out using stochastic gradient descent. The following sections outline some simple interpretability questions and their analysis using the proposed methods.

6.4.1 Evolution of Feature Representations during Training

Developing a useful CNN machine learning model involves effective training of the architecture to extract feature representations that are optimised for the desired objective. The optimisation process is a gradient-based technique that iteratively updates the feature representations, until convergence. Therefore, the evolution of the feature representations can be tracked throughout the training to gain a better understanding of their convergence process. Similar to [236], we analyse the evolution of a feature representation during training by comparing it with its final representation at the end of the training process.

SVMCCA. The tracking of a feature representation during training involves an analysis of the same layer at different time instants, and therefore, m layers are of same dimensions $h \times w \times d$. As described previously, the response of convolutional layer $(N \times h \times w \times d)$ on N training images is reshaped into layer dataset $L_i(Nhw \times d)$ for the SVMCCA analysis. The trained UNet consists of five convolutional layers in the downsampling encoder path, four convolutional layers in the upsampling decoder path, and a final convolutional logit layer producing the dense aneurysm prediction map for the input image.

At the first stage, we analyse the feature representations of the nine layers. The network is trained for 50 epochs, and the feature representations are sampled at seven training time instants (epochs: 0, 10, 20, 30, 40, 45, 49). It leads to 7 layer datasets for a particular layer. Therefore, the input to the SVMCCA algorithm consists of 63 datasets (9 layer representations acquired at 7 training time instants), with dataset $L_{i,j}(Nhw \times d)$ representing dataset of layer i at training time j. The SVMCCA algorithm constructs the canonical variable datasets for the 63 layers, maximising the sum of their overall correlations. The mutual correlation-based similarity scores $\{\rho_{ij}^{SVMCCA}\}_{i,j=1}^{i,j=9}$ between the nine layers at the sampled training time instants are then determined, as shown in Figure 6.1. It can be observed that the early layers in the network start converging first towards their final representations, an observation that is consistent with the observation from [236]. It can also be

noted from Figure 6.1 that the final layers attain a relatively higher similarity score compared to the middle stage of the network. It is because the skip connections in the UNet propagate the representations from the earlier layers in the network and combine them with the representations from the decoding layers at the later stages. This observation can also be inferred from the higher off-diagonal ρ_{ij}^{SVMCCA} entries in the Figure 6.1. The off-diagonal similarity scores represent the relationship between layers in the earliest and latest stages of the network (layer 1 with layer 9, layer 2 with layer 8, etc.).

2DSVMCCA. The feature map representations produced by the filters within a layer are further analysed using the proposed 2DSVMCCA algorithm. As described previously, $L_i(N \times h \times w \times d)$ is the dataset of layer$i$ consisting of d filter kernels. The value of d ranges from 64 to 1024. Therefore, to reduce the computational complexity, we randomly subsample 50 filter representations from a layer to construct 50 filter datasets of dimensions $h \times w \times N$ at a given training time instant, for the given layer. The sampling of filter datasets at the seven training instants for layer L_i leads to 350 feature map datasets of dimensions $h \times w \times N$, which are subsequently analysed using the proposed 2DSVMCCA. The evolution of a filter kernel, during the training process, with respect to its final representation can be seen from the Figure 6.2. It can be noted that the filters in the early stage layers of the network converge faster than the filters in the later stages of the network. It can also be observed that the skip connections induce higher similarity values in the final stage layers of the network. Another interesting to be noted is that not all the filter representations in a layer converge to fully to their final representations. It indicates that the layer may learn the representation in an effective subspace, as also evidenced from [236].

To further analyse the layerwise representations, the SVMCCA is applied to analyse the datasets acquired at the different training time instants for a specific layer. The SVMCCA similarity measure represents the group similarity measure for the layer during the training process. Similarly, the randomly sampled 50 filter representa-

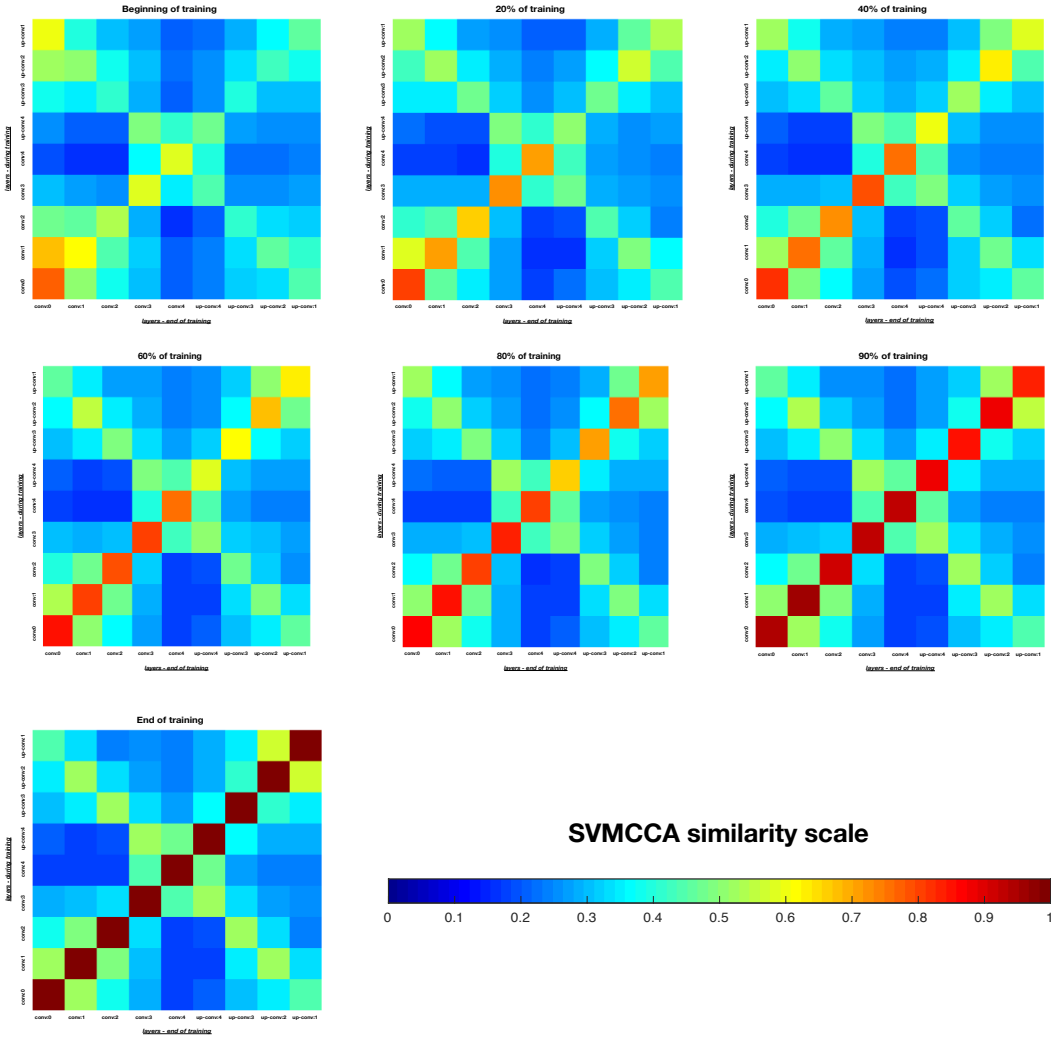


Figure 6.1: Analysing the evolution of a layer representation during training, with reference to its final representation, using SVMCAA. The entries in cell (i, j) represent ρ_{ij}^{SVMCCA} similarity score between layer i and layer j. The higher scores on the diagonal and off-diagonal entries can be observed from the figure.

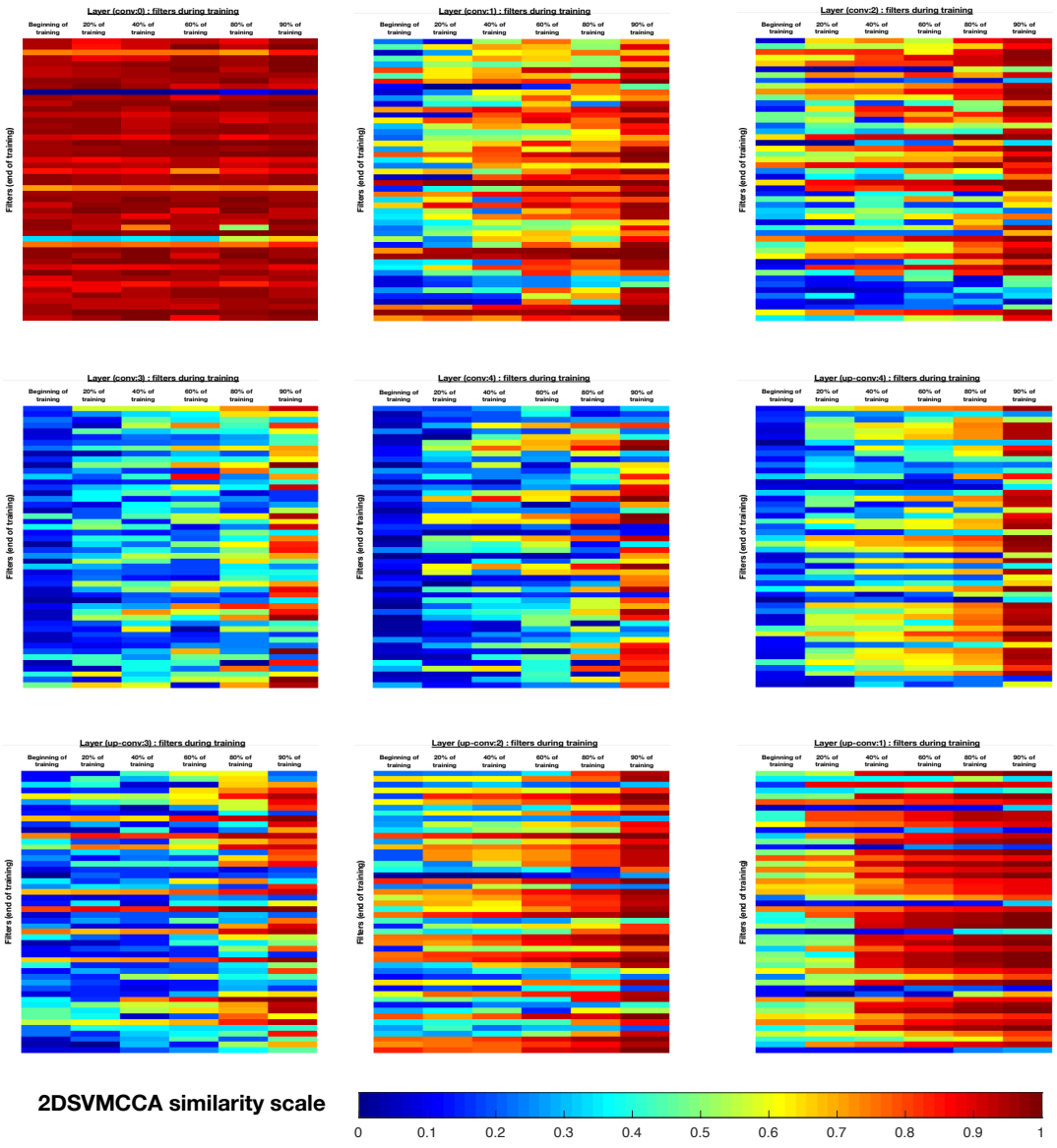


Figure 6.2: Analysing the evolution of feature map representations of a filter in a layer during training, with reference to its final representation, using 2DSVMCAA. The entries in the cell (i, j) show the $\rho_{ii}^{2DSVMCCA}$ similarity score of i^{th} filter at j^{th} training time instant. The higher similarity scores for the filters in the initial and deeper layers can be observed.

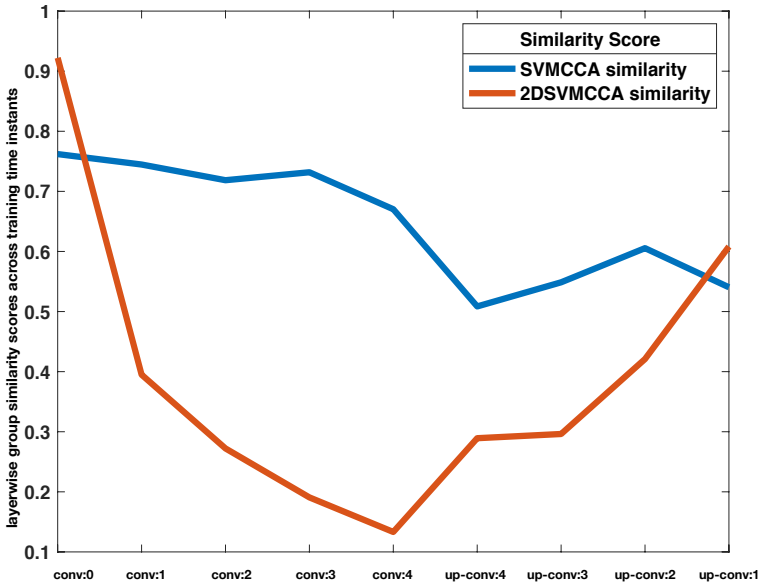


Figure 6.3: The group similarity scores ρ^{SVMCCA} and $\rho^{2DSVMCCA}$ of representations of a specific layer and its filters, across all the training time instants, using SVMCCA and 2DSVMCCA.

tions, at the different training time instants, for a specific layer are analysed using 2DSVMCCA to determine the group similarity measure from the filters. The Figure 6.3 shows group similarity measures computed for all the layers using SVMCCA and 2DSVMCCA. It can be observed from the Figure 6.3 that the early layers attain a high similarity convergence score and thereby, converge faster than the layers in the later stages of the encoder network, with the skip connections reversing the convergence in the decoder stage. It can bee seen that the trend remains the same using SVMCCA and 2DSVMCCA, and the discrepancy between the two is because the 2DSVMCCA is an approximation computed using 50 randomly sampled filter representations.

6.4.2 Comparing Intermediate Feature representations with Final Prediction Representations

Comparing the layer representations with the final logit predictions involves measuring the representational similarities between the feature representations in a layer and the feature representations of the final convolutional logit layer and gives a

good indication of the progression of the sensitivity of a layer to the final prediction logits.

SVMCCA. The SVMCCA computation involves analysis between a layer i dataset $L_{i,j}(Nhw \times d)$ acquired at training instant j and the final convolutional logit layer $L_1 0$ acquired at the same training instance.

2DSVMCCA. The 2DSVMCCA computation involves analysis between the feature map datasets of randomly sampled 50 filters of layer i at training instant j and the final convolutional logit dataset representing the prediction maps for the aneurysm voxels.

Figure 6.4 shows the evolving relationship of a convolutional layer and a group of its filters, with the final logit layer and aneurysm prediction filter, respectively. The filter similarities are computed using 2DSVMCCA representing the group correlation similarity ($\rho^{2DSVMCCA}$) between the 51 filters (50 for a layer and 1 final aneurym logit). The layer similarities are computed using SVMCCA representing the similarity ρ_{i10}^{SVMCCA} between i^{th} convolutional layer and the final prediction layer. It can be seen that the layers closer to the output and skip connections produce the maximum similarity score underlying the importance of skip connections to the final performance. The lower similarity score for middle layers of the encoder-decoder architecture further highlights their bottleneck nature and where overall performance can be improved.

6.4.3 Cross Model Comparisons

Another important application is to determine the representational similarities between the different architectures that might provide more insights into generic properties about the representational powers of a network. To this end, we trained five UNet architectures. Three networks were trained with three different initial learning rates, and the two were initialised with the same learning rate. The representations were compared after the training process.

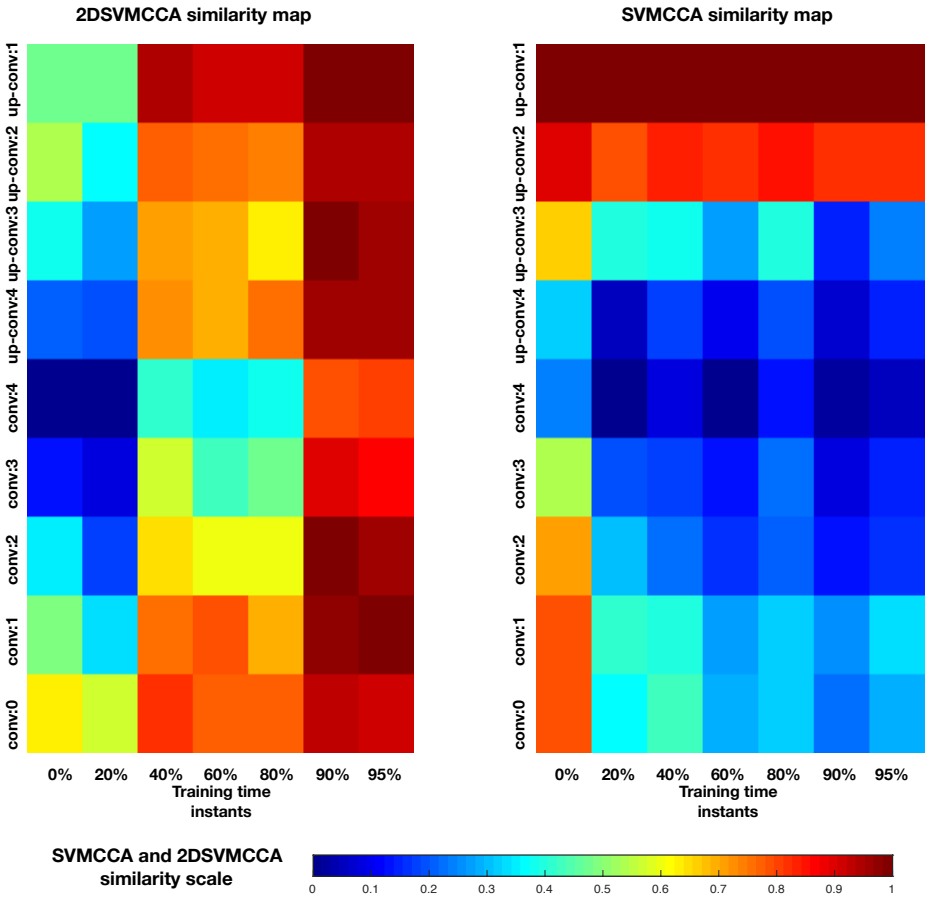


Figure 6.4: Comparing the layer representations and the respective feature map representations of the filters, during training, to the final prediction layer and feature representation of its filter predicting the aneurysm voxels. Figure in the left pane represents the SVMCCA similarity scores. The entries in the cell (i, j) represent the ρ_{i10}^{SVMCCA} similarity between i^{th} layer and the 10^{th} prediction layer at j^{th} training time instant. Figure in the right pane represents the 2DSVMCCA similarity scores. The entries in the cell (i, j) represent the $\rho^{2DSVMCCA}$ group similarity score between randomly sampled 50 filters of layer i and the aneurysm predicting filter of layer 10.

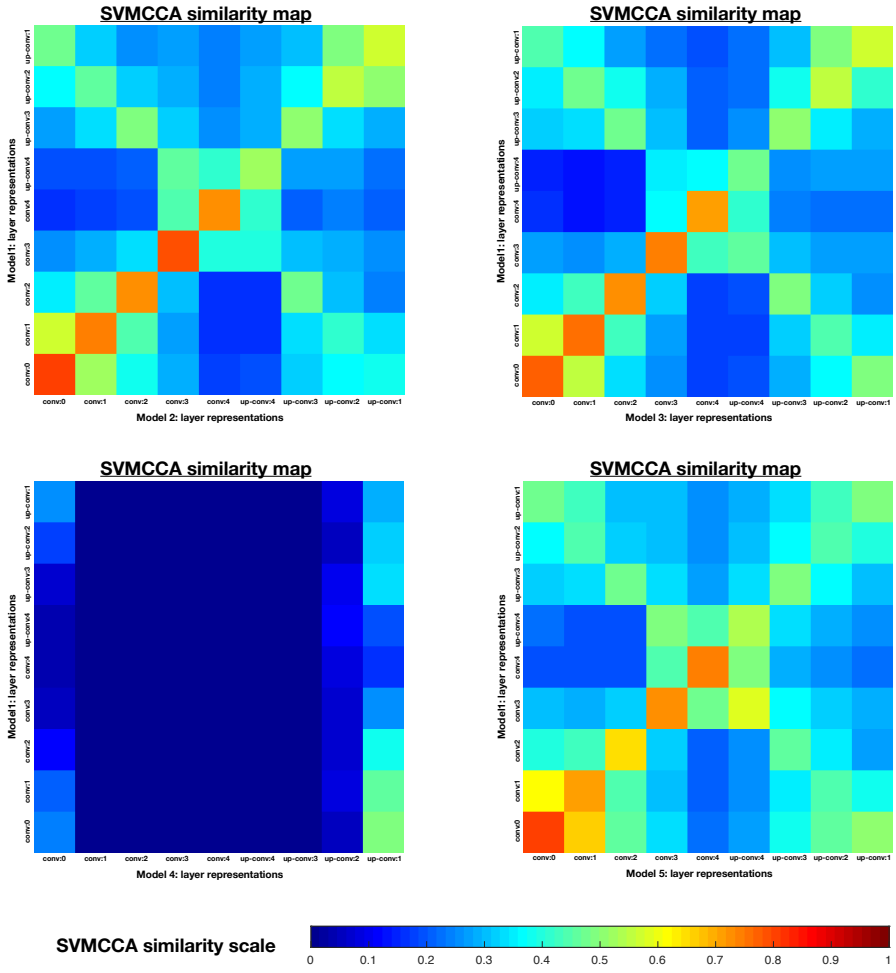


Figure 6.5: Comparing the layer representations of multiple trained UNet models using SVMCAA. The entries represent ρ_{ij}^{SVMCCA} similarity score between layer i of a model x and layer j of model y. The higher scores on the diagonal and of-diagonal entries can be observed from the figure, highlighting the similar and distinct natures of the lower and higher layers, respectively.

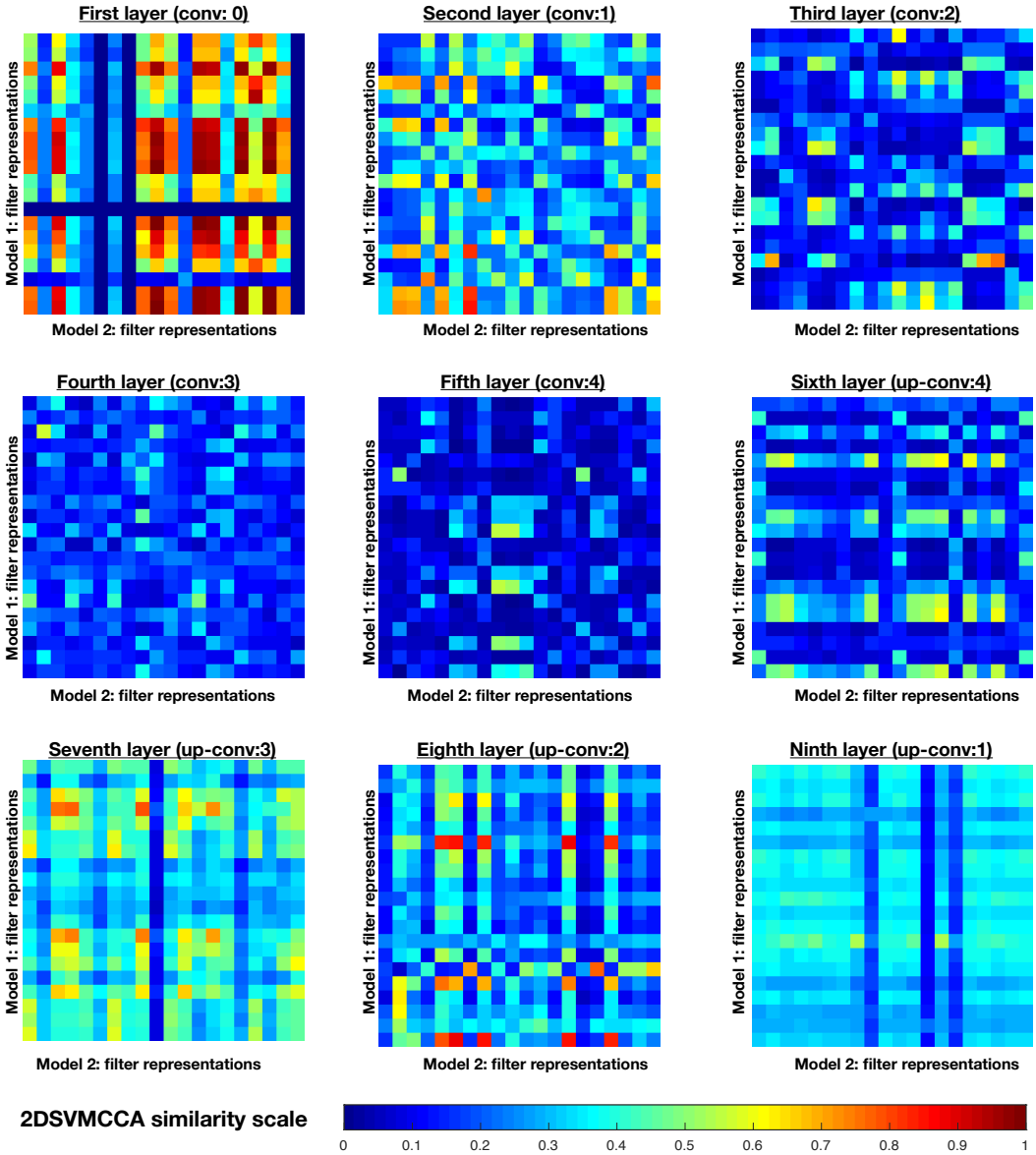


Figure 6.6: Comparing the feature map representations of randomly sampled 50 filters, of a particular layer, between model 1 and model 2. The entries in cell (i, j) of a similarity map represent $\rho_{ij}^{2DSVMCCA}$ similarity scores between the filters i and j, of a specific layer, between models 1 and 2. The distinct nature of the filters in the middle layer that differentiate the representations can be observed from the figure.

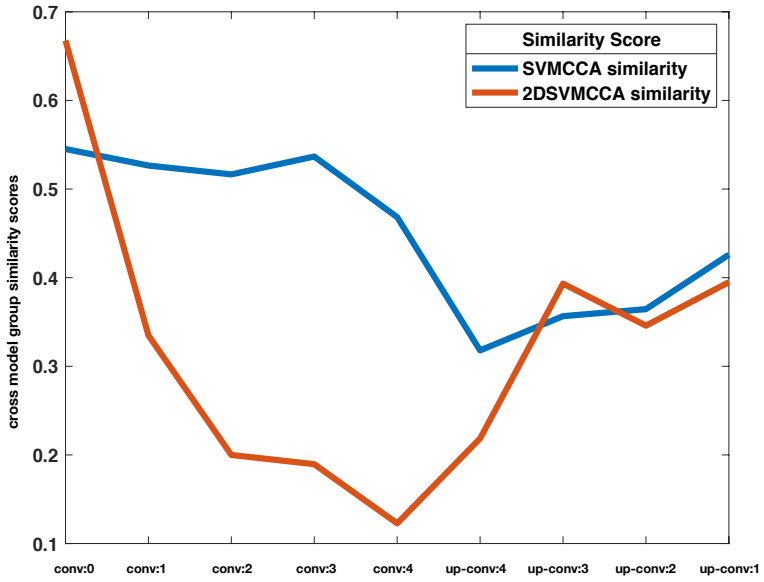


Figure 6.7: Layerwise group similarity scores of the layer representations and corresponding filter feature map representations, using ρ^{SVMCCA} and $\rho^{2DSVMCCA}$, aggregated from the five trained models.

SVMCCA. As described previously, the UNet architecture consists of 9 convolutional layers apart from the final logit layer. The 9 layers of a network and the five different models lead to 45 layer datasets. The comparison involves different layers, and therefore, all the layers are interpolated to a predefined fixed size. The 45 layer datasets are analysed using the SVMCCA to determine the inter-model similarities, as shown Figure 6.5. It can be observed that the layers in the earlier stages of the network are the most correlated across the networks, implying that the essential features remain the same independent of the final solution. As the network progresses, it can be seen that the similarity reduces significantly highlighting the distinctive nature of the higher layers. An interesting observation from the similarities between models 1, 2, 3, and 5 is that the higher layers are less correlated, indicating the convergence of networks towards different high-level feature representations and solutions that are not correlated. However, the validation performances of these networks are at a similar level despite the higher layers achieving separate solutions. It indicates the presence of multiple optima solutions that a network can achieve, an observation that is consistent with the inference from [236]. Also, a low

similarity score between models 3 and 4 can be noted from Figure 6.5, which can be independently verified by a poor validation performance of model 4.

2DSVMCCA. The 2DSVMCCA can be used to further analyse the layerwise relationships between two models. Figure 6.6 shows the similarities between feature map representations of model 1 and model 2. Randomly sampled 50 filters from a layer of each architecture are combined to create a dataset of 250 filters. The 250 datasets are analysed using 2DSVMCCA to construct the correlated subspace. Each image gives a detailed insight into the relationship between the two layers. It can be observed that the filters in the lower and the higher layers have higher similarity scores, with the higher encoding layer representing the distinction between the filters of the two architectures. Figure 6.7 shows the layerwise group similarity between all the five trained models, computed using SVMCCA and 2DSVMCCA.

6.5 Conclusion

Machine learning methods employing CNNs have dramatically advanced performance boundaries of image recognition challenges and also, medical image analysis tasks. However, the 'blackbox' nature of the CNNs is often a limitation in several applications as it constrains them from providing a greater understanding of the feature representations and a meaningful knowledge of the task. To this end, this chapter presented some simple, novel post hoc interpretability techniques that can be applied to interpret the representations of a trained CNN model. CCA-based novel methods, SVMCCA and 2DSVMCCA were proposed that can be applied to analyse the feature representations of the layers and filters of the CNN architecture. The proposed approaches were demonstrated by applying them to a UNet trained to predict aneurysm pixels from CTA slices that highlighted some basic properties such as the layer dynamics during the training, filter dynamics during training, inter-model comparisons.

Chapter 7

Conclusion and Future Research Direction

This chapter summarises the contributions to the thesis, outlines future research directions, and concludes overall thesis objectives.

7.1 Summary of Contributions

Neurological diseases are diseases of the nervous system and are significant contributors to human mortality and morbidity. Several diagnostic choices exist from indirect measurement techniques such as EEG, EMG to the non-invasive and direct medical imaging modalities. Medical imaging modalities, in particular, have greatly enhanced diagnostic accuracy with improved spatial-temporal resolutions. This thesis presented automated approaches for the early diagnosis of neurological diseases from CT modality medical images. The application included cerebral aneurysms and Parkinson's disease. Further, interpretability techniques were proposed that assist in analysing and understanding the inner workings of CNN-based automated approaches.

7.1.1 Cerebral Aneurysms

A cerebral aneurysm is a localised dilation of the brain vessels due to weakness in the vessel wall. They may have fatal consequences with their ruptures, leading to possible subarachnoid haemorrhage. CTA is an emerging modality to diagnose the

unruptured aneurysms with high-sensitivity to plan preventive treatments. However, it is challenging to manually diagnose complex aneurysm structures from the CTA images with a potentially high cost for a missed diagnosis. To this end, this research makes contributions to assist with the early diagnosis of aneurysms paving the way for timely clinical interventions.

First, a large-scale brain CTA dataset of aneurysms is constructed that is useful to explore and analyse the aneurysm characteristics in details. The dataset consists of aneurysms annotated by experienced neuroradiologists, with a high inter-observer agreement. Next, a novel CNN architecture is proposed to diagnose and localise the aneurysms from the orthogonal views of the volumetric CTA. The CNN architecture is trained using the fully annotated dataset with an out of sample independent test set to verify its performance. The proposed approach works at two levels by first diagnosing the presence of an aneurysm in the given CTA examination and subsequently, showing its actual location in the form of a bounded rectangle. It forms a handy assistive tool for radiologists to efficiently investigate for the presence of aneurysms from the CTA images.

7.1.2 Parkinson's Disease

Parkinson's disease is a neurodegenerative disorder affecting mainly the motor system resulting in the paucity of voluntary/involuntary movements. Speech impairment can be a critical feature at an early stage of the illness accompanied by abnormal vocal fold movements. Therefore, analysing the speech issues and associated vocal fold structure can pave the way for earlier detection of the disease. Advantages of non-invasive, direct assessments in assessing vocal fold movements have been shown in prior research. To this end, this research contributes simple, automated approaches to explore and analyse vocal fold movements from neck CT images.

First, a large-scale neck CT dataset was constructed consisting of neck images acquired during a phonation experiment. An exploratory approach consisting of a sequence of basic image processing operations was proposed to identify clinically

relevant feature points of the arytenoid cartilages, which support the movements of vocal folds. The dataset was subsequently augmented with neck images of healthy controls captured during a breathing period supplemented by annotations of the arytenoid cartilages. A CNN-based object detector was developed to fully localise the arytenoid cartilages, followed by an inter arytenoid distance feature differentiating the subgroups of Parkinson's patients from that of healthy controls. It demonstrates that neck imaging is a useful modality to analyse and assess speech abnormalities towards early diagnosis of Parkinson's disease.

7.1.3 Interpretability Methods

The 'blackbox' nature of the machine learning automation approaches and the CNN architectures, in particular, is a bottleneck to understanding the automated approaches advancing the frontiers of medical diagnostics. The opaque nature of these approaches prevents the end-user from extracting more meaningful knowledge of the inner workings of these technologies and especially, inhibiting them from analysing failures in greater details. Interpretability in machine learning is an emerging topic that provides useful tools to analyse and interpret the decision making the process of machine learning approaches providing better clarity in their internal dynamics. To the end, this research contributes simple data-driven techniques that assist in analysing the learned representations of the CNNs and the relationships between different CNN architectures.

CCA is a data-driven technique that captures the best possible relationship between sets of multivariate data. SVMCCA and 2DSVMCCA algorithms are proposed that apply to multiple sets of vector and image datasets, simultaneously. The proposed algorithms are demonstrated on a dense prediction CNN architecture trained to detect aneurysms from CTA images. The interpretations include analysing the evolution of a layer over time, capturing the relationship amongst a set of layers, and analysing groups of trained CNN architectures. They form utility tools useful to analyse and interpret the CNN training performance and gain an insight into their

performances.

7.2 Limitations and Future Research Directions

This thesis proposed automated approaches for the early diagnosis of neurological diseases from computed tomography images. The first two contributions focused on two neurological diseases: cerebral aneurysms and Parkinson's disease. The final contribution was to develop automated interpretability techniques to interpret and understand the automated machine learning approaches for image analysis tasks. The following paragraphs outline future research directions for each contribution made in the thesis.

Cerebral aneurysm. The cerebral aneurysm was the first application considered in the present work. A CNN based automated approach was designed and implemented to detect and localise the aneurysms from the CTA imagery. An apparent future study for this work would involve boosting the diagnostic accuracy further. Besides, the specificity needs to be improved to reduce false diagnosis. Another possible extension is to improve the time and computational complexities of the automated approaches as CNN methods, in general, are computationally expensive to train and validate. Furthermore, efficient automated approaches can be developed to harness limited data power more effectively.

Parkinson's disease. The second application involved analysis of speech abnormalities in case of Parkinson's disease. A large-scale neck CT dataset was constructed, and automated approaches were demonstrated to study the vocal fold movements from the CT image data. This work could pave ways to explore clinically relevant feature information with knowledge diversity from the image data and better understand the speech quality in Parkinson's patients. Using the proposed contribution as a platform, more comprehensive experiments could be designed to precisely investigate the usefulness and accuracies of these features in classifying disease patients from healthy controls. A primary reason for using CT images is that they provide

improved resolution, spatially and temporally. A better alternative would be to investigate the same problem using other image modalities with no radiation effects. Challenges, however, would arise if the spatial resolution is not enough to accurately delineate the vocal folds during phonation.

Machine learning interpretability. A common drawback of the automated approaches based on machine learning is the lack of knowledge that enhances the relevant user knowledge. The final contribution to the thesis attempted to develop machine learning interpretability techniques in this regard. CCA was employed to propose novel interpretability tools that provide simple interpretation windows in the inner working of a CNN method. Interpretability in machine learning is a widely emerging topic that is essential to provide a greater understanding of the dynamics of a machine learning system. The proposed methods could be extended in simple ways to seamlessly integrate them into the neural network training process and build performance indicators that assist in understanding the training process better, rather than only interpreting them post-hoc. The computationally expensive and memory-intensive models exploit the energy and resources drastically, which can limit their hardware deployment despite their extraordinary performance. The studies to reduce the size and cost of a deep network are in demand for energy conservation and hardware development. Generally, only a small percentage of features actively contribute to the prediction. Possible feature engineering studies could involve fine-tuning the model by removing low-correlation, rare features inside (network pruning) as well as combining multiple models by concatenating the selective features (with no overlap). The post-hoc ensembles with a greater model or feature diversity could be tested at the training stage to improve ensemble performance.

7.3 Conclusion

The research contributions documented in this thesis revolved around developing automated assistance methods for the early diagnosis of neurological diseases from computed tomography images. Medical image analysis generally two key steps of

data acquisition and data analysis. A third step is increasingly becoming essential that involves interpretation of the data analysis techniques. This thesis focused on all the three aspects by constructing large-scale CT datasets through retrospective data acquisition, automated methods employing CNNs, and interpretability techniques using CCA. Cerebral aneurysm and Parkinson's disease were the two applications considered in this work. The approaches developed would assist the clinicians in effective diagnosis of aneurysms, efficient analysis of speech abnormalities in Parkinson's patients. The interpretability techniques would further provide a platform to extract meaningful knowledge from the automated approaches, in addition to only the final results.

Bibliography

[1] T. Dua, M. G. Cumbrera, C. Mathers, and S. Saxena, "Global burden of neurological disorders: estimates and projections," *Campanini B (Editora). Neurological disorders: Public Health Challenges. Ginebra: Banco Mundial*, pp. 27–39, 2006.

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

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

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

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

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

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

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[11] J. I. Suarez, R. W. Tarr, and W. R. Selman, "Aneurysmal subarachnoid hemorrhage," *New England Journal of Medicine*, vol. 354, no. 4, pp. 387–396, 2006.

[12] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann *et al.*, "Parkinson disease. Nature reviews Disease primers. 2017; 3: 17013."

[13] Parkinson's Disease Foundation, "Statistics on Parkinson's," http://www.pdf.org/en/parkinson_statistics, 2017.

[14] R. J Holmes, J. M Oates, D. J Phyland, and A. J Hughes, "Voice characteristics in the progression of Parkinson's disease," *International Journal of Language & Communication Disorders*, vol. 35, no. 3, pp. 407–418, 2000.

[15] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of

Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[16] W. J. Mutch, A. Strudwick, S. K. Roy, and A. W. Downie, "Parkinson's disease: disability, review, and management." *Br Med J (Clin Res Ed)*, vol. 293, no. 6548, pp. 675–677, 1986.

[17] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and cognition*, vol. 56, no. 1, pp. 24–29, 2004.

[18] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[19] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting cnns via decision trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6261–6270.

[20] R. Abbasi-Asl and B. Yu, "Interpreting convolutional neural networks through compression," *arXiv preprint arXiv:1711.02329*, 2017.

[21] A. Park, C. Chute, P. Rajpurkar, J. Lou, R. L. Ball, K. Shpanskaya, R. Jabarkheel, L. H. Kim, E. McKenna, J. Tseng *et al.*, "Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Usin the HeadXNet Model," *JAMA network open*, vol. 2, no. 6, pp. e195 600–e195 600, 2019.

[22] Z.-H. Cho, *Foundations of medical imaging.* Wiley-Interscience, 1993.

[23] R. A. Robb, *Three Dimensional Biomedical Imaging (1985): Volume II.* CRC Press, 2017.

[24] T. M. Deserno, Ed., *Biomedical Image Processing.* Springer Berlin Heidelberg, 2011. [Online]. Available: https://doi.org/10.1007%2F978-3-642-15816-2

[25] J. Tavares and R. N. Jorge, *Advances in computational vision and medical image processing: methods and applications.* Springer Science & Business Media, 2008, vol. 13.

[26] I. Bankman, *Handbook of medical image processing and analysis.* Elsevier, 2008.

[27] G. Dougherty, *Medical image processing: techniques and applications.* Springer Science & Business Media, 2011.

[28] A. C. Kak, M. Slaney, and G. Wang, "Principles of computerized tomographic imaging," *Medical Physics*, vol. 29, no. 1, pp. 107–107, 2002.

[29] J. Hsieh, *Computed tomography: principles, design, artifacts, and recent advances.* SPIE press, 2003, vol. 114.

[30] T. M. Buzug, "Computed tomography," in *Springer Handbook of Medical Technology.* Springer, 2011, pp. 311–342.

[31] D. C. Hatcher, "Operational principles for cone-beam computed tomography," *The Journal of the american dental association*, vol. 141, pp. 3S–6S, 2010.

[32] R. C. Orth, M. J. Wallace, M. D. Kuo, T. A. C. of the Society of Interventional Radiology *et al.*, "C-arm cone-beam CT: general principles and technical considerations for use in interventional radiology," *Journal of Vascular and Interventional Radiology*, vol. 19, no. 6, pp. 814–820, 2008.

[33] K. T. Bae, "Intravenous contrast medium administration and scan timing at CT: considerations and approaches," *Radiology*, vol. 256, no. 1, pp. 32–61, 2010.

[34] A. Sodickson, P. F. Baeyens, K. P. Andriole, L. M. Prevedello, R. D. Nawfel, R. Hanson, and R. Khorasani, "Recurrent CT, cumulative radiation exposure, and associated radiation-induced cancer risks from CT of adults," *Radiology*, vol. 251, no. 1, pp. 175–184, 2009.

[35] G. Boulouis, A. Morotti, A. Charidimou, D. Dowlatshahi, and J. N. Goldstein, "Noncontrast computed tomography markers of intracerebral hemorrhage expansion," *Stroke*, vol. 48, no. 4, pp. 1120–1125, 2017.

[36] U. Tayal, L. King, R. Schofield, I. Castellano, J. Stirrup, F. Pontana, J. Earls, and E. Nicol, "Image reconstruction in cardiovascular CT: Part 2–Iterative reconstruction; potential and pitfalls," *Journal of cardiovascular computed tomography*, 2019.

[37] T. Ogawa, T. Okudera, K. Noguchi, N. Sasaki, A. Inugami, K. Uemura, and N. Yasui, "Cerebral aneurysms: evaluation with three-dimensional CT angiography." *American journal of neuroradiology*, vol. 17, no. 3, pp. 447–454, 1996.

[38] W. A. Kalender, *Computed tomography: fundamentals, system technology, image quality, applications*. John Wiley & Sons, 2011.

[39] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*. Cambridge university press, 2017.

[40] W. Hollingworth, C. J. Todd, M. I. Bell, Q. Arafat, S. Girling, K. R. Karia, and A. K. Dixon, "The diagnostic and therapeutic impact of MRI: an observational multi-centre study," *Clinical radiology*, vol. 55, no. 11, pp. 825–831, 2000.

[41] G. B. Chavhan, P. S. Babyn, B. Thomas, M. M. Shroff, and E. M. Haacke, "Principles, techniques, and applications of T2*-based MR imaging and its special applications," *Radiographics*, vol. 29, no. 5, pp. 1433–1449, 2009.

[42] P. A. Rinck, "A short history of magnetic resonance imaging," *Spectroscopy Europe*, vol. 20, no. 1, pp. 7–10, 2008.

[43] C. L. Dumoulin and H. Hart Jr, "Magnetic resonance angiography." *Radiology*, vol. 161, no. 3, pp. 717–720, 1986.

[44] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[45] M. D'Esposito, E. Zarahn, and G. K. Aguirre, "Event-related functional MRI: implications for cognitive psychology." *Psychological bulletin*, vol. 125, no. 1, p. 155, 1999.

[46] R. S. Cobbold, *Foundations of biomedical ultrasound*. Oxford university press, 2006.

[47] F. A. Duck, A. C. Baker, and H. C. Starritt, *Ultrasound in medicine*. CRC Press, 1998.

[48] L. Bricker, J. Garcia, J. Henderson, M. Mugford, J. Neilson, T. Roberts, M. Martin *et al.*, "Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's views," 2000.

[49] R. Badea and S. Ioanitescu, "Ultrasound Imaging of Liver Tumors–Current Clinical Applications," in *Liver Tumors*. IntechOpen, 2012.

[50] C. Merritt, "Ultrasound safety: what are the issues?" *Radiology*, vol. 173, no. 2, pp. 304–306, 1989.

[51] D. L. Bailey, M. N. Maisey, D. W. Townsend, and P. E. Valk, *Positron emission tomography*. Springer, 2005.

[52] T. A. Henderson, "The diagnosis and evaluation of dementia and mild cognitive impairment with emphasis on SPECT perfusion neuroimaging," *CNS spectrums*, vol. 17, no. 4, pp. 176–206, 2012.

[53] R. C. Gonzalez, R. E. Woods *et al.*, *Digital image processing*. Prentice hall Upper Saddle River, NJ, 2002.

[54] A. K. Jain, *Fundamentals of digital image processing.* Englewood Cliffs, NJ: Prentice Hall„ 1989.

[55] J. Goutsias, L. Vincent, and D. S. Bloomberg, *Mathematical morphology and its applications to image and signal processing.* Springer Science & Business Media, 2006, vol. 18.

[56] R. M. Rangayyan, *Biomedical image analysis.* CRC press, 2004.

[57] A. Meyer-Baese and V. J. Schmid, *Pattern recognition and signal analysis in medical imaging.* Elsevier, 2014.

[58] P. Suetens, *Fundamentals of medical imaging.* Cambridge university press, 2017.

[59] W. E. Brant and C. A. Helms, *Fundamentals of diagnostic radiology.* Lippincott Williams & Wilkins, 2012.

[60] Y. Zhang and L. Wu, "Optimal multi-level thresholding based on maximum Tsallis entropy via an artificial bee colony approach," *Entropy*, vol. 13, no. 4, pp. 841–859, 2011.

[61] M. P. De Albuquerque, I. A. Esquef, and A. G. Mello, "Image thresholding using Tsallis entropy," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1059–1065, 2004.

[62] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[63] L.-K. Huang and M.-J. J. Wang, "Image thresholding by minimizing the measures of fuzziness," *Pattern recognition*, vol. 28, no. 1, pp. 41–51, 1995.

[64] R. Pohle and K. D. Toennies, "Segmentation of medical images using adaptive region growing," in *Medical Imaging 2001: Image Processing*, vol. 4322. International Society for Optics and Photonics, 2001, pp. 1337–1346.

[65] F. Y. Shih and S. Cheng, "Automatic seeded region growing for color image segmentation," *Image and vision computing*, vol. 23, no. 10, pp. 877–886, 2005.

[66] G. Tsechpenakis, "Deformable model-based medical image segmentation," in *Multi modality state-of-the-art medical image segmentation and registration methodologies*. Springer, 2011, pp. 33–67.

[67] T. Heimann and H. Delingette, "Model-based segmentation," in *Biomedical Image Processing*. Springer, 2010, pp. 279–303.

[68] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[69] S. Lankton and A. Tannenbaum, "Localizing region-based active contours," *IEEE transactions on image processing*, vol. 17, no. 11, pp. 2029–2039, 2008.

[70] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[71] T. F. Cootes and C. J. Taylor, "Active shape models'smart snakes'," in *BMVC92*. Springer, 1992, pp. 266–275.

[72] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: a review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[73] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[74] M. Droske, B. Meyer, M. Rumpf, and C. Schaller, "An adaptive level set method for medical image segmentation," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2001, pp. 416–422.

[75] S. Balla-Arabé, X. Gao, and B. Wang, "A fast and robust level set method for image segmentation using fuzzy clustering and lattice Boltzmann method," *IEEE transactions on cybernetics*, vol. 43, no. 3, pp. 910–920, 2013.

[76] X. Chen and L. Pan, "A survey of graph cuts/graph search based medical image segmentation," *IEEE reviews in biomedical engineering*, vol. 11, pp. 112–124, 2018.

[77] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Departmental Papers (CIS)*, p. 107, 2000.

[78] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1768–1783, 2006.

[79] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[80] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[81] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE transactions on medical imaging*, vol. 24, no. 3, pp. 371–380, 2005.

[82] Y. Zheng, B. Georgescu, and D. Comaniciu, "Marginal space learning for efficient detection of 2D/3D anatomical structures in medical images," in *International Conference on Information Processing in Medical Imaging*. Springer, 2009, pp. 411–422.

[83] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT vol-

umes using marginal space learning and steerable features," *IEEE transactions on medical imaging*, vol. 27, no. 11, pp. 1668–1681, 2008.

[84] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[85] D. Mahapatra, "Analyzing training information from random forests for improved image segmentation," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1504–1512, 2014.

[86] T. M. Mitchell *et al.*, "Machine learning," 1997.

[87] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological cybernetics*, vol. 20, no. 3-4, pp. 121–136, 1975.

[88] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern recognition*, vol. 15, no. 6, pp. 455–469, 1982.

[89] M. C. Mozer, *The perception of multiple objects: A connectionist approach*. The MIT Press, 1991.

[90] T. Matsumoto, T. Yokohama, H. Suzuki, R. Furukawa, A. Oshimoto, T. Shimmi, Y. Matsushita, T. Seo, and L. Chua, "Several image processing examples by cnn," in *IEEE International Workshop on Cellular Neural Networks and their Applications*. IEEE, 1990, pp. 100–111.

[91] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Transactions on circuits and systems*, vol. 35, no. 10, pp. 1257–1272, 1988.

[92] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[93] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[94] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, 2013, pp. 1139–1147.

[95] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2015.

[96] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[97] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in neural information processing systems*, 2008, pp. 161–168.

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

[99] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[100] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[101] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[102] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in neural information processing systems*, 2013, pp. 2553–2561.

[103] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[104] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik, "Recognition using regions," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1030–1037.

[105] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[106] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[107] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[108] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[109] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012.

[110] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *European conference on computer vision*. Springer, 2014, pp. 345–360.

[111] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[112] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[113] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[114] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[115] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[116] L.-C. Chen, G. Papandreou, and et. al., "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[117] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[118] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[119] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE transactions on Medical Imaging*, vol. 15, no. 5, pp. 598–610, 1996.

[120] S.-C. B. Lo, J.-S. Lin, M. T. Freedman, and S. K. Mun, "Computer-assisted diagnosis of lung nodule detection using artificial convolution neural network," in *Medical Imaging 1993: Image Processing*, vol. 1898. International Society for Optics and Photonics, 1993, pp. 859–869.

[121] W. Zhang, K. Doi, M. L. Giger, Y. Wu, R. M. Nishikawa, and R. A. Schmidt, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Medical Physics*, vol. 21, no. 4, pp. 517–524, 1994.

[122] S.-C. Lo, S.-L. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE Transactions on Medical Imaging*, vol. 14, no. 4, pp. 711–718, 1995.

[123] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

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

[125] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis:

Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[126] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[127] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[128] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[129] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1170–1181, 2015.

[130] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1332–1343, 2016.

[131] B. D. de Vos, J. M. Wolterink, P. A. de Jong, T. Leiner, M. A. Viergever, and I. Isgum, "ConvNet-based localization of anatomical structures in 3-D medical images," *IEEE Trans Med Imaging*, vol. 36, no. 7, pp. 1470–1481, 2017.

[132] F. C. Ghesu, E. Krubasik, B. Georgescu, V. Singh, Y. Zheng, J. Hornegger, and D. Comaniciu, "Marginal space deep learning: efficient architecture for volumetric image parsing," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1217–1228, 2016.

[133] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, "3D deep learning for efficient and robust landmark detection in volumetric data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 565–572.

[134] J. M. Wolterink, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum, "Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks," *Medical image analysis*, vol. 34, pp. 123–136, 2016.

[135] N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Išgum, "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions," *IEEE transactions on medical imaging*, 2017.

[136] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2016.

[137] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[138] M. T. McKenna, S. Wang, T. B. Nguyen, J. E. Burns, N. Petrick, and R. M. Summers, "Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence," *Medical image analysis*, vol. 16, no. 6, pp. 1280–1292, 2012.

[139] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[140] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[141] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[142] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE transactions on medical imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[143] Z. Jia, X. Huang, I. Eric, C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[144] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz *et al.*, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 674–683, 2017.

[145] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[146] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.

[147] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[148] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[149] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[150] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazurowski, "Hierarchical convolutional neural networks for segmentation of breast tumors in mri with application to radiogenomics," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 435–447, 2018.

[151] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[152] L. Mou, L. Chen, J. Cheng, Z. Gu, Y. Zhao, and J. Liu, "Dense Dilated Network with Probability Regularized Walk for Vessel Detection," *IEEE transactions on medical imaging*, 2019.

[153] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating Fully Convolutional Networks with Spatial and Channel 'Squeeze and Excitation' Blocks," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 540–549, 2018.

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

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

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

[157] A. Riaz, M. Asad, E. Alonso, and G. Slabaugh, "Deepfmri: End-to-end deep learning for functional connectivity and classification of adhd using fmri," *Journal of Neuroscience Methods*, vol. 335, p. 108506, 2020.

[158] A. M. Aradhya and A. Ashfahani, "Deep network optimization for rs-fmri classification," in *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2019, pp. 77–82.

[159] S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P. K. Pal, and M. Ingalhalikar, "Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri," *Neuroimage: Clinical*, vol. 22, p. 101748, 2019.

[160] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[161] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, p. 158, 2018.

[162] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[163] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[164] W. Kuo, C. Hne, P. Mukherjee, J. Malik, and E. L. Yuh, "Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 45, pp. 22 737–22 745, 2019.

[165] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[166] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[167] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study," *The Lancet*, vol. 392, no. 10162, pp. 2388–2396, 2018.

[168] A. Nielsen, M. B. Hansen, A. Tietze, and K. Mouridsen, "Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning," *Stroke*, vol. 49, no. 6, pp. 1394–1401, 2018.

[169] J. Merkow, R. Lufkin, K. Nguyen, S. Soatto, Z. Tu, and A. Vedaldi, "DeepRadiologyNet: radiologist level pathology detection in CT head images," *arXiv preprint arXiv:1711.09313*, 2017.

[170] M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan, "Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 281–284.

[171] T. Jerman, F. Pernuš, B. Likar, and Ž. Špiclin, "Computer-aided detection and quantification of intracranial aneurysms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 3–10.

[172] C. M. Hentschke, K. D. Tönnies, O. Beuing, and R. Nickl, "A new feature for automatic aneurysm detection," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. IEEE, 2012, pp. 800–803.

[173] A. Lauric, E. Miller, S. Frisken, and A. M. Malek, "Automated detection of intracranial aneurysms based on parent vessel 3D analysis," *Medical image analysis*, vol. 14, no. 2, pp. 149–159, 2010.

[174] R. Cárdenes, J. M. Pozo, H. Bogunovic, I. Larrabide, and A. F. Frangi, "Automatic aneurysm neck detection using surface Voronoi diagrams," *IEEE transactions on medical imaging*, vol. 30, no. 10, pp. 1863–1876, 2011.

[175] R. C. Helmich, D. E. Vaillancourt, and D. J. Brooks, "The future of brain imaging in parkinsons disease," *Journal of Parkinson's disease*, vol. 8, no. s1, pp. S47–S51, 2018.

[176] M. F. Dirkx, H. den Ouden, E. Aarts, M. Timmer, B. R. Bloem, I. Toni, and R. C. Helmich, "The cerebral network of parkinson's tremor: an effective connectivity fmri study," *Journal of Neuroscience*, vol. 36, no. 19, pp. 5362–5372, 2016.

[177] J. B. Rowe, L. E. Hughes, R. A. Barker, and A. M. Owen, "Dynamic causal modelling of effective connectivity from fmri: are results reproducible and sensitive to parkinson's disease and its treatment?" *Neuroimage*, vol. 52, no. 3, pp. 1015–1026, 2010.

[178] S. Prange, E. Metereau, and S. Thobois, "Structural imaging in parkinsons disease: new developments," *Current Neurology and Neuroscience Reports*, vol. 19, no. 8, p. 50, 2019.

[179] J. L. Brisman, J. K. Song, and D. W. Newell, "Cerebral aneurysms," *New England journal of medicine*, vol. 355, no. 9, pp. 928–939, 2006.

[180] J. Van Gijn, R. S. Kerr, and G. J. Rinkel, "Subarachnoid haemorrhage," *The Lancet*, vol. 369, no. 9558, pp. 306–318, 2007.

[181] W. I. Schievink, "Intracranial aneurysms," *New England Journal of Medicine*, vol. 336, no. 1, pp. 28–40, 1997.

[182] S. Juvela, M. Porras, and O. Heiskanen, "Natural history of unruptured intracranial aneurysms: a long-term follow-up study," *Journal of neurosurgery*, vol. 79, no. 2, pp. 174–182, 1993.

[183] N. Chalouhi, B. L. Hoh, and D. Hasan, "Review of cerebral aneurysm formation, growth, and rupture," *Stroke*, vol. 44, no. 12, pp. 3613–3622, 2013.

[184] D. O. Wiebers, I. S. of Unruptured Intracranial Aneurysms Investigators *et al.*, "Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment," *The Lancet*, vol. 362, no. 9378, pp. 103–110, 2003.

[185] Commons.wikimedia.org., "Cerebral aneurysm NIH - Wikimedia Commons," https://commons.wikimedia.org/wiki/File:Cerebral_aneurysm_NIH.jpg, 2010, [Online; accessed 13-Jan-2020].

[186] MayoClinic, "Aneurysmal Rupture," https://www.mayoclinic.org/hidden-dangers-brain-aneurysm-infographic/ifg-20404403, [Online; accessed 13-Jan-2020].

[187] P. A. Turski, W. J. Zwiebel, C. M. Strother, A. B. CrummY, G. G. Celesia, and J. F. Sackett, "Limitations of intravenous digital subtraction angiography." *American Journal of Neuroradiology*, vol. 4, no. 3, pp. 271–273, 1983.

[188] Z. L. Yang, Q. Q. Ni, U. J. Schoepf, C. N. De Cecco, H. Lin, T. M. Duguay, C. S. Zhou, Y. E. Zhao, G. M. Lu, and L. J. Zhang, "Small intracranial aneurysms: diagnostic accuracy of CT angiography," *Radiology*, vol. 285, no. 3, pp. 941–952, 2017.

[189] C. M. Hentschke, O. Beuing, R. Nickl, and K. D. Tönnies, "Detection of cerebral aneurysms in MRA, CTA and 3D-RA data sets," in *Medical Imaging 2012: Computer-Aided Diagnosis*, vol. 8315. International Society for Optics and Photonics, 2012, p. 83151I.

[190] A. Firouzian, R. Manniesing, Z. H. Flach, R. Risselada, F. van Kooten, M. C. Sturkenboom, A. van der Lugt, and W. J. Niessen, "Intracranial aneurysm segmentation in 3D CT angiography: Method and quantitative validation with and without prior noise filtering," *European journal of radiology*, vol. 79, no. 2, pp. 299–304, 2011.

[191] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nature methods*, vol. 9, no. 7, p. 671, 2012.

[192] Z. Gao, L. Wang, and G. Wu, "LIP: Local Importance-based Pooling," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3355–3364.

[193] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[194] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *In NIPS Workshop*, 2017.

[195] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[196] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[197] J. Zhuang, "Laddernet: Multi-path networks based on u-net for medical image segmentation," *arXiv preprint arXiv:1810.07810*, 2018.

[198] A. K. Ho, J. L. Bradshaw, and R. Iansek, "For better or worse: The effect of levodopa on speech in parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 4, pp. 574–580, 2008.

[199] K. Tjaden, "Speech and swallowing in Parkinson's disease," *Topics in geriatric rehabilitation*, vol. 24, no. 2, p. 115, 2008.

[200] L. Perju-Dumbrava, K. Lau, D. Phyland, V. Papanikolaou, P. Finlay, R. Beare, P. Bardin, S. Stuckey, P. Kempster, and D. Thyagarajan, "Arytenoid cartilage movements are hypokinetic in Parkinsons disease: A quantitative dynamic computerised tomographic study," *PloS one*, vol. 12, no. 11, p. e0186611, 2017.

[201] H. Gray, *Anatomy of the human body*. Lea & Febiger, 1918.

[202] X. Qin, S. Wang, and M. Wan, "Improving reliability and accuracy of vibration parameters of vocal folds based on high-speed video and electroglottography,"

IEEE Transactions on Biomedical Engineering, vol. 56, no. 6, pp. 1744–1754, 2009.

[203] C. R. Krausert, A. E. Olszewski, L. N. Taylor, J. S. McMurray, S. H. Dailey, and J. J. Jiang, "Mucosal wave measurement and visualization techniques," *Journal of Voice*, vol. 25, no. 4, pp. 395–405, 2011.

[204] K. Bevan, M. Morgan, and M. Griffiths, "The role and techniques of laryngeal electromyography," *Clinical Otolaryngology & Allied Sciences*, vol. 13, no. 4, pp. 299–305, 1988.

[205] M. Novotnỳ, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in parkinsons disease," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366–1378, 2014.

[206] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Noth, "Unobtrusive monitoring of speech impairments of parkinson's disease patients through mobile devices," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6004–6008.

[207] J. R. Orozco-Arroyave, J. C. Vasquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei *et al.*, "Neurospeech: An open-source software for parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, 2018.

[208] J. R. Orozco-Arroyave, J. Vdsquez-Correa, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Noth, "Towards an automatic monitoring of the neurological state of parkinson's patients from speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6490–6494.

[209] L. Jeancolas, D. Petrovska-Delacretaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehericy, and H. Benali, "X-vectors: New quantitative

biomarkers for early parkinson's disease detection from speech," *arXiv preprint arXiv:2007.03599*, 2020.

[210] L. Dumbrava, Lau.k, D.Phyland, P.Finlay, R.Beare, P.Bardin, P. Stuckey, P.Kempster, and D.Thyagarajn, "Vocal cords are hypokinetic in parkinson's disease," Departments of Neuroscience, Respiratory Medicine, Radiology Monash Medical Centre, Clayton, Victoria, Australia, Tech. Rep., 2014.

[211] T. Ogawa, R. Enciso, A. Memon, J. K. Mah, and G. T. Clark, "Evaluation of 3D airway imaging of obstructive sleep apnea with cone-beam computed tomography," *Studies in health technology and informatics*, vol. 111, pp. 365–368, 2005.

[212] I. Cheng, S. Nilufar, C. Flores-Mir, and A. Basu, "Airway segmentation and measurement in ct images," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 795–799.

[213] S. Hewavitharanage, J. Gubbi, D. Thyagarajan, K. Lau, and M. Palaniswami, "Estimation of vocal fold plane in 3d ct images for diagnosis of vocal fold abnormalities," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 3105–3108.

[214] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[215] P. R. Reddy, V. Amarnadh, and M. Bhaskar, "Evaluation of stopping criterion in contour tracing algorithms," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 3, pp. 3888–3894, 2012.

[216] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[217] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[218] Y. Chen, C. Han, Y. Li, Z. Huang, Y. Jiang, N. Wang, and Z. Zhang, "SimpleDet: A Simple and Versatile Distributed Framework for Object Detection and Instance Recognition," *Journal of Machine Learning Research*, vol. 20, no. 156, pp. 1–8, 2019. [Online]. Available: http://jmlr.org/papers/v20/19-205.html

[219] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6054–6063.

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

[221] C. Molnar, *Interpretable machine learning*. Lulu. com, 2019.

[222] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

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

[224] P. Weinzaepfel, H. Jégou, and P. Pérez, "Reconstructing an image from its local descriptors," in *CVPR 2011*. IEEE, 2011, pp. 337–344.

[225] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[226] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[227] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin, "Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization," *Advances in Computational Mathematics*, vol. 25, no. 1-3, pp. 161–193, 2006.

[228] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of machine learning research*, vol. 2, no. Mar, pp. 499–526, 2002.

[229] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, "General conditions for predictivity in learning theory," *Nature*, vol. 428, no. 6981, p. 419, 2004.

[230] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[231] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. E. Hopcroft, "Convergent Learning: Do different neural networks learn the same representations?" in *Iclr*, 2016.

[232] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 991–999.

[233] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint arXiv:1711.09784*, 2017.

[234] Q.-s. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.

[235] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu, "Interpreting cnn knowledge via an explanatory graph," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[236] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Advances in Neural Information Processing Systems*, 2017, pp. 6076–6085.

[237] A. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," in *Advances in Neural Information Processing Systems*, 2018, pp. 5727–5736.

[238] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[239] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[240] S. H. Lee and S. Choi, "Two-dimensional canonical correlation analysis," *IEEE Signal Processing Letters*, vol. 14, no. 10, p. 735, 2007.

[241] C. Ding and J. Ye, "2-dimensional singular value decomposition for 2d maps and images," in *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 2005, pp. 32–43.