AUTOMATED ANALYSIS OF ABDOMINAL AORTIC CALCIFICATION IN VERTEBRAL FRACTURE ASSESSMENT IMAGES

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF BIOLOGY, MEDICINE AND HEALTH

2020

Luke A Chaplin

School of Health Sciences
Division of Informatics, Imaging & Data Sciences

Contents

Word Count: 41420

List of Tables

List of Figures

Abstract

Cardiovascular diseases are the most common cause of death globally. For more than half of those who die of a cardiovascular event, the disease has been clinically silent until that point, indicating a need for more targeted intervention. Abdominal aortic calcification (AAC) is an independent predictor of CVD and can be used as a measure of atherosclerotic extent within the arterial system, allowing more accurate risk stratification and monitoring ahead of a major cardiovascular event. Dual energy X-ray absorptiometry (DXA) vertebral fracture assessment (VFA), performed on a densitometer can visualise calcifications in the abdominal aorta. These images represent an opportunity to obtain clinically informative data on cardiovascular risk in a noninvasive manner. Despite these advantages, AAC is time consuming to annotate and not routinely reported; it is not commonly used to affect treatment decisions.

This work investigates the automation of AAC measurement in VFA images. Approaching from the perspective of a semantic segmentation problem, two major strategies are compared to automatically identify AAC. Both random forest classification and convolutional neural networks are applied to the problem of AAC segmentation on VFA images for the first time. Additionally, an automated method to locate the abdominal aorta within VFA images using skeletal landmarks, and subdivide the aorta to produce clinically informative semi-quantitative AAC scores is presented, unique in VFA images. This is the first deep learning work in this area, and this segmentation strategy is demonstrated to outperform the random forest, and previous work on AAC segmentation in other x-ray images. This work also presents the first automated attempt at recreating a semi-quantitative clinical measure of AAC. Automated scoring shows good correlation with expert scoring of images, indicating the potential for its use as a clinically informative screening tool.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property

and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses

Acknowledgements

I would like to take the opportunity to thank the many people that have helped make this work possible.

First and foremost, my endless appreciation to my supervisor Tim for enabling me to pursue, and supporting me throughout, this life-changing opportunity. I am truly grateful for his wisdom, limitless patience and rallying good humour. I would also like to thank the late Judy Adams, whom I did not get the opportunity to know as well as I would have liked, but who was instrumental in setting me on the right path. My thanks also go to Drs. Paul Bromiley, Adrian Davison, and Josh Lewis and Prof David Suter for their advice and technical assistance. I would like to acknowledge the assistance given by Research IT and the use of the Computational Shared Facility at The University of Manchester, without which I quite literally could not have completed this work.

I would like to thank my friends and colleagues throughout the Centre for Imaging Sciences, the wider university and beyond. I am particularly grateful to Ethan, Geo, Anna Maria, Raja, Chris, Matt, Adam and Luca, for their support over the years, be it emotional, technical, delectable or convivial.

A special thank you to my partner Emma, for her love, encouragement and limitless patience during the writing of this thesis; I could not have done it without her. And to her family, for their kind assistance through difficult times.

I could not have reached this stage without the love and support of my family and all my friends. I would especially like to give my love and thanks to my mother Joanne, my father Bruce, Jake, Leo, Andrew, Jack, Hugo, Monty and Gareth, who have all helped more than I can express. And to Jimmy, Jordan, Simon, Joe and Nel, for their invaluable friendship and for never failing to put a smile on my face, I am truly grateful.

Abbreviations

AAC Abdominal Aortic Calcification

Adagrad Adaptive Gradient Algorithm

Adam Adaptive Moment Estimation Algorithm

AUC Area Under the Curve

BMD Bone Mineral Density

CAC Coronary Artery Calcification

CAIFOS Calcium Intake Fracture Outcome Study

CKD Chronic Kidney Disease

CLM Constrained Local Model

CNN Convolutional Neural Network

CT Computed Tomography

CVD Cardiovascular Disease

DSC Dice-Sørensen Coefficient

DXA Dual-Energy X-Ray Absorptiometry

FCN Fully Convolutional Network

GP Gaussian Processes

GPU Graphics Processing Unit

ICC Intraclass Correlation Coefficient

IoU Intersection over Union

***k*-NN** *k*-Nearest Neighbour

LDL Low-Density Lipoprotein

MACD Morphological Atherosclerotic Calcification Distribution

MRC Medical Research Council

MRI Magnetic Resonance Imaging

NICE National Institute for Health and Care Excellence

NSHD National Survey of Health and Development

PDM Point Distribution Model

ReLU Rectified Linear Unit

ROC Receiver Operating Characteristic

ROI Region of Interest

SGD Stochastic Gradient Descent

SMC Smooth Muscle Cells

SSM Statistical Shape Model

SVM Support Vector Machine

TPS Thin Plate Spline

VFA Vertebral Fracture Assessment

Chapter 1

Introduction

1.1 Motivation

Cardiovascular diseases (CVD) are the most common cause of death globally. The majority of these diseases are preventable and driven by atherosclerotic changes in medium and large arteries. For more than half of those who die of a cardiovascular event, the disease has been clinically silent until that point. Direct observation of the underlying atherosclerotic process is not routinely used to make treatment decisions, instead relying on clinical risk scores. Abdominal aortic calcification (AAC) is an independent predictor of CVD and can be used as a measure of atherosclerotic extent within the arterial system, allowing more accurate risk stratification and monitoring ahead of a major cardiovascular event. Dual energy X-ray absorptiometry (DXA) vertebral fracture assessment (VFA), performed on a densitometer can visualise calcifications in the abdominal aorta. Age and osteopenia increase risk of cardiovascular disease and so DXA VFA represents an excellent targeted screening tool for assessing CVD risk. Despite these advantages, AAC is rarely scored on DXA VFA reports and is currently not used to affect intervention decisions. The development of an automated system to identify and score AAC on DXA VFA images could inform targeting

of CVD interventions and improve data gathering for further study of calcification and CVD.

1.2 Aims and Objectives

This project aims to develop software that can automatically quantify the extent of abdominal aortic calcification on vertebral fracture assessment images, and elicit clinically valuable information. The first challenge is to locate the aorta within images using statistical models of the spatial relationship between the abdominal aorta and bony landmarks. After identifying the aorta, the aim is to produce a method to segment and measure calcification and demonstrate a correlation with current semi-quantitative scores. With a large bank of DXA images available with which to train and test the model, the goal is to demonstrate strong correlation with human interpretation of calcification scores.

1.3 Outline of Thesis

This chapter has summarised the motivations and broad aims of the project. It is now describing the structure of the thesis, in a way that threatens recursion. The overall design of the thesis is that the methodology, results and discussion of the thesis contributions are contained in the latter chapters, with Chapters 2 and 3 covering the clinical and technical background respectively.

Chapter 2 introduces the aetiology and pathophysiology of abdominal aortic calcification; imaging and quantification; correlations with cardiovascular disease and risk monitoring; and intervention and treatment strategies. This chapter is designed to build from a fundamental level and give context for the motivations of the thesis methodologies. However, those sufficiently familiar with the clinical background should be free

to omit Chapter 2 and return if any additional context is required.

Chapter 3 details the literature supporting the methodological decisions made in later chapters. The chapter will build up a description and justification for the techniques used in the thesis methods. Similarly to Chapter 2, it is not essential to read this chapter in its entirety if the reader is familiar with the literature, as later chapters will reference the sections of Chapter 3 relevant to each methodological choice. It is though, the intention that these chapters not be a chore to read, regardless of familiarity with the subject.

Chapter 4 provides details of the methodology that are pertinent to all approaches to this problem. The chapter will cover the source of the images used throughout, and details of annotation and scoring. It will then cover the methods, results and discussion of a method to select relevant regions of interest in images, and a method to produce semi-quantitative measures of abdominal aortic calcification from segmentation masks.

Chapter 5 covers a series of experiments utilising random decision forests, classifying image patches to segment calcification. This chapter covers the methodology, and compares the accuracy of the results with human annotation and previous work in the literature.

Chapter 6 presents experiments involving deep learning algorithms to segment aortic calcification. Focusing on the U-Net model, a fully convolutional network designed for biomedical image segmentation, this chapter details the process of model optimisation and hyper-parameter tuning for this particular problem. It then describes experiments with recent variations on the U-Net in the literature and how these segmentations compare with random forests and human annotation.

Chapter 7 briefly discusses the findings of the thesis, summarises the conclusions and reflects on avenues for future work to continue the project.

Chapter 2

Clinical Background

This chapter explores the significance and literature of the clinical research area. While future chapters concentrate on the concepts and implementation of machine learning approaches to quantifying abdominal aortic calcification, this chapter explores the clinical importance of the work and its potential value in improving patient outcomes. The following sections give context for the project, covering: the physiology and pathology of abdominal aortic calcification; how it is imaged, assessed and measured; how it relates to cardiovascular risk; and how treatment decisions are made.

2.1 Atherosclerosis and Vascular Calcification

Cardiovascular disease (CVD) is a broad class of diseases involving the heart or blood vessels. Included in this class are some of the most common causes of death globally: ischaemic heart diseases, such as myocardial infarction, and cerebrovascular diseases, such as stroke [1]. The main driving force in the development of these forms of CVD is atherosclerosis. Atherosclerosis is defined by fatty plaque formation within artery walls that narrows the artery and disrupts blood flow. It is a gradual process involving the proliferation of cells and accumulation of a lipid plaque in arterial walls. Vascular

calcification develops as a late stage in the pathology of atherosclerosis and further complicates the picture, impacting the function of the vascular system.

This section covers the pathophysiology of atherosclerosis, how it develops and impacts vascular function, then continues by exploring how this leads to the development of calcification in vessel walls, with a focus on the abdominal aorta.

2.1.1 Pathophysiology of Atherosclerosis

The development of atherosclerosis is gradual, driven by the accumulation of damage to arteries and a cascade of inflammatory reactions. Early changes can be observed in the first decade of life, demonstrating the gradual and clinically silent nature of the process [2]. Affecting a range of medium and large arteries, these changes begin in the innermost layers of an artery, and involve an increasing proportion of the vessel over time.

The fundamental structure of all arteries is similar, composed of three main layers. The outermost collagen rich layer, the tunica adventitia, and the middle muscular layer, the tunica media, are thicker in large vessels like the aorta. The innermost layer, the tunica intima, consists of an endothelial layer, a subendothelial layer of connective tissue, and an elastic membrane. Atherosclerotic changes primarily occur in the tunica intima, although in the advanced stages the inflammation and destruction can involve the tunica media [3].

During normal functioning of an artery, the endothelial cells which line the internal walls are subjected to small amounts of chemical damage, such as free radicals, and mechanical damage, from distortion of the vessel. These changes allow extravasation of monocyte inflammatory cells and lipoproteins such as low-density lipoprotein (LDL) into the subendothelial layer. After this migration the monocytes can differentiate into macrophages and lipoproteins are oxidised to form proinflammatory and

cytotoxic compounds [4]. The ready availability of LDLs in the blood and increased blood pressure are, as a result, major risk factors in CVD [5, 6]. This cytotoxic damage to the endothelium results in further recruitment of immune cells such as macrophages, via expression of adhesion molecules, and the inflammatory process continues within the artery wall.

These early changes form thick fatty streaks in the intimal wall but do not cause any loss of function, as the lumen can widen to compensate. After decades of this low-level inflammatory response however, a fibroproliferative stage can manifest. Smooth muscle cells (SMCs), which repair damage to the arterial walls and ensure their elastic nature, proliferate in response to the inflammatory changes. These cells stabilise the growing plaque by constructing a collagen-rich matrix and help prevent rupture and thrombosis. With prolonged recruitment, the SMCs proliferate to an extent that narrows the arterial lumen [7]. SMCs begin to differentiate into other cell phenotypes, including more macrophages. Apoptosis within the populations of smooth muscle cells and macrophages result in the deposition of cholesterols and further inflammatory signaling molecules, the nature of the plaque prevents removal of the resulting debris. Remaining SMCs construct a thick fibrous cap of collagen matrix over the growing plaque, forming a fibroatheroma.

In this advanced stage of atherosclerosis, the fibrous cap forms over the necrotic core of cellular debris. As the self-propelling inflammatory recruitment continues, with proliferation of cells and subsequent cell death, the plaque becomes increasingly unstable. As the number of SMCs begins to decline, due to appoptosis and differentiation, the fibrous cap begins to thin, making plaque rupture increasingly likely [8, 7]. Plaque erosion is also possible while this process is occurring, involving the breakdown of the endothelial layer, exposing the thrombogenic collagen matrix. Both rupture and erosion of the plaque lead to thrombolic events, which cause failure of the vessel and infarction [4].

As part of the complex inflammatory processes, calcium phosphate is deposited and trapped within the atherosclerotic plaque. This calcium phosphate undergoes metabolic activity and is integrated into the plaque, leading to vascular calcification. Though atherosclerosis is not the only source of calcific vascular load.

2.1.2 Vascular Calcification

Vascular calcification is defined by the accumulation of calcium phosphate within the walls of blood vessels. This pathological ectopic calcification is associated with metabolic diseases such as atherosclerosis, chronic kidney disease (CKD) and diabetes. Atherosclerotic changes in the intimal layer of the arteries can lead to the development of vascular calcification as damage to the endothelium accumulates. Alternatively, vascular calcification can occur in the medial layer, primarily associated with CKD, through mineral deposition in the smooth muscle layer of the arteries.

Intimal calcifications occur as an extension of the chronic atherosclerotic process, further limiting the elasticity of the artery. After atherosclerotic plaques have formed, calcification of the lipid rich core and connective tissues occurs [9]. While calcium is deposited as a result of cell death in the atherosclerotic plaques, the process of calcification within the arterial wall is not simply passive mineral deposition. The signalling pathways and restructuring in the vessel wall is more similar to ossification occurring in the skeletal system.

In response to the increasing concentration of cytokines, oxidised lipoproteins, calcium and phosphate from cell debris within the atherosclerotic plaque, vascular SMCs differentiate into osteoblast-like cells. This process is driven by a multitude of signalling chemicals such as bone morphogenetic protein and osteopontin within the atheroma [10]. The osteoblast-like cells begin to deposit hydroxyapatite and calcify extracellular collagen matrices in the absence of the typical collagen scaffolding

required for bone formation. The development of calcifications within artery walls further limits the functioning of the vessel. The process of calcification within the medial layer of the artery leads to the same differentiation of SMCs, though the source of the calcium phosphate is not from cell death within the arterial walls.

Vascular calcification has been observed for some time as a major factor in the morbidity and mortality of CKD. End stage CKD patients, where glomerular filtration rate has fallen below 15mL/min, and patients requiring dialysis have a high prevalence of vascular calcification and an increased risk of CVD [11]. Further investigation has demonstrated that glomerular filtration rate has an inverse relationship with cardiovascular mortality, and CVD is the leading cause of death in these patients [12, 13].

The aetiology of vascular calcification in CKD patients is different from the general population. CKD causes reduced excretion of phosphate and activation of vitamin D, limiting calcium absorption [13]. In CKD, high phosphate concentration causes it to precipitate out of solution as calcium phosphate. Not only does this precipitate cause calcification, but also drives the differentiation of smooth muscle cells into osteoblast-like cells, increasing deposition [14]. The tunica media is the middle layer of the artery wall and consists of smooth muscle and elastic tissue. Without significant atherosclerosis in the vessel wall the tunica media is the only location for vascular smooth muscle cells. As these differentiating cells are a major driving force in the calcification, CKD calcification occurs primarily in the media.

Though two separate processes, calcification in the tunica media and intima occur concurrently in individuals with CKD, creating a mixed picture. It has been demonstrated that CKD accelerates the development of vascular calcification and arteriosclerosis. As the rates of cardiovascular events are increased in this group, it is even more important to accurately assess cardiovascular risk [15]. Current imaging methods for measurement of AAC do not, and possibly cannot, assess the layer of the aorta containing the calcification but there could be variations in the distribution of the calcified

plaques that could [12]. Regardless of the source of the calcification, it is clear that the assessment of vascular calcification has clinically valuable information, and as the largest artery, calcification in the aorta may be an accessible source of this information.

Abdominal Aortic Calcification

Abdominal aortic calcification (AAC) develops as part of late stage atherosclerotic processes in the aorta. It has been found that atherosclerotic changes develop early and are most severe in the abdominal aorta [3]. Age related thickening of the aortic intimal layer can accelerate rapidly from the fourth decade of life, and calcification becomes increasingly common with age. Prevalence of AAC in over 70's was found to be 98% in men and 93% in women, compared to 55% and 50% in those aged 50-60 [16]. The effects of AAC directly affect the function of the aorta, increasing risk of CVD, and the presence of AAC is well correlated with calcification in other vessels [17], indicating a measure of general atherosclerosis in the arterial system.

There are several mechanisms through which AAC may directly contribute to CVD. The reduced elasticity of the aorta has demonstrable effects on the dynamics of blood flow through the vascular system [18]. One role of the aorta, is to convert the pulsatile flow from the heart into a more continuous flow to the peripheral arteries. The elasticity of the aorta allows it to distend during systole and slow the flow rate. During diastole, contraction of the aortic wall maintains pressure in the vasculature to prevent a drastic diastolic drop in blood pressure [19]. In particular, perfusion of coronary arteries is dependent on diastolic pressure, as they cannot be perfused effectively during contraction of the ventricles.

The changes that occur within the walls of the aorta with AAC reduce its elasticity, impacting the pattern of pressures within the vasculature. The increased aortic stiffness raises the pressure the heart is pumping against, the left ventricular afterload. If the heart cannot compensate then left sided and subsequently congestive heart failure

can develop [20, 21]. The increase in pressure in the descending aorta can redirect a higher flow rate into the branches of the aortic arch, potentially causing gradual damage and atherosclerotic changes to the arteries feeding the brain, such as the carotid arteries. With the compromise of the aorta, flow is faster during systole, leading to turbulence and increased risk of thrombotic events. During diastole the pressure drops lower, reducing perfusion of the coronary arteries. This problem is exacerbated by the calcification of the coronary arteries associated with severe AAC, leading to reduced perfusion and potential for ischaemia [22].

Correlation with cardiovascular mortality necessitates the ability to assess these calcifications in vivo. To allow accurate comparison, both on an individual and population level, it is necessary to have reliable imaging technologies and quantitative measures of calcification.

2.2 Imaging and Measurement of AAC

A wealth of literature has explored the use of a range of non-invasive imaging techniques to assess calcification throughout the arterial system. While techniques such as ultrasound have been used to confirm the presence of calcification in superficial arteries such as the carotid arteries, assessment is subjective and qualitative [11]. The high radiopacity of calcium allows x-ray based modalities to clearly distinguish calcifications from other tissues, enabling quantification of calcific plaques.

Agatston et al. [23] quantified calcification of the coronary arteries using ultrafast computed tomography. The Agatston score was developed, calculated by multiplying the volume of a calcification by a weighted density score based on attenuation values on the Hounsfield scale. After an Agatston score has been calculated for each calcification, the scores are summed to produce a coronary calcium score. This total coronary calcium score was later improved upon with a lesion specific score that considered

shape and positioning of lesions within the coronary arteries, improving prediction of coronary artery disease [24]. These quantitative computed tomography CT techniques have been used to assess aortic calcification, with a fully quantitative measure calculated based on the same methodology using total calcific volume and attenuation [25, 26]. However, with a high radiation dose, CT imaging is not an ideal screening tool. This section covers the methods for imaging and measuring AAC, concentrating on low radiation $2D$ x-ray modalities of the lateral spine.

2.2.1 Lateral Radiography

Radiographs have also been used successfully to image AAC using a lateral view of the lumber spine. Though this modality cannot represent the calcification volumetrically, it can still visualise the extent of calcification and has several advantages over CT while mainting good correlation with the extent detected with these methods [27]. The main drawback to using CT is the high radiation exposure, typically 8mSV for an abdominal acquisition, which restricts its use in studies and as a screening tool. Radiographs involve a far lower effective radiation dose of 800μSv [16], can produce an image faster, and are less expensive. It has been demonstrated that AAC on radiographs can be used to predict CVD and that this AAC is correlated with coronary artery calcifications, another predictor of CVD [28].

The main scoring system used for AAC in lateral radiographs is a 24-point semi-quantitative scale developed by Kauppila et al. [29] on lateral radiograph images. This AAC-24 score, demonstrated in Figure 2.1, considers the anterior and posterior walls of the aorta at the levels of the lumbar vertebrae L1-4. Calcification of each wall is graded based on the total height of calcification within the aorta adjacent to each lumbar vertebra. Each of the 8 sections, two walls at 4 levels, is graded 0-3 for a total score of 0-24. If the total height of calcifications within a section is more than $2/3$ of

the vertebral height, that section is scored as 3. If the total height of calcifications is between $1/3$ and $2/3$ of vertebral height, the section is scored as 2. A score of 1 is given to sections with calcifications totaling less than $1/3$ vertebral height. And 0 for complete absence of calcification.

Schousboe et al. [30] undertook a literature review of clinical risk scoring using the AAC-24 score. Based on their meta-analysis of risk stratification, their recommendation is that AAC can be classified as mild, moderate and severe in order to inform clinical risk. This classification is used throughout this work to define severity as follows: calcification is considered mild with a AAC-24 score of 1 or 2, moderate between 3 and 5, and severe above 5.

The AAC-24 scoring system is designed to facilitate easy and quick estimation of AAC in these images. As exact measurement is difficult and time consuming, the semi-quantitative score is open to some subjective variation between interpreters. However, both inter- and intra-observer agreement is relatively high, with intraclass correlation coefficients above 0.9 [29, 31, 32, 33].

An 8-point scale was later developed by Schousboe et al. [31] in order to simplify the semi-quantitative assessment of AAC. Based on the AAC-24 score, AAC-8 measures calcification of the posterior and anterior aortic walls in the region L1-L4. However, AAC-8 does not require subdivision based on vertebral level, each wall is graded 0-4. 0 represents no calcification, 1 the aggregate height of calcification on the wall is less than one vertebral length, 2 between one and two vertebral lengths, 3 between two and three vertebral lengths and 4 if the aggregate length exceeds three vertebral lengths. This scale has an advantage in terms of speed over AAC-24, and it is less sensitive to small calcifications that are spread out [31, 33]. As demonstrated in Figure 2.2, small dispersed calcification could read as severe on AAC-24 but would be correctly identified as mild on AAC-8. Overall though, there is good correlation between the two scores in both radiographs and DXA [31, 34]. It is worth noting

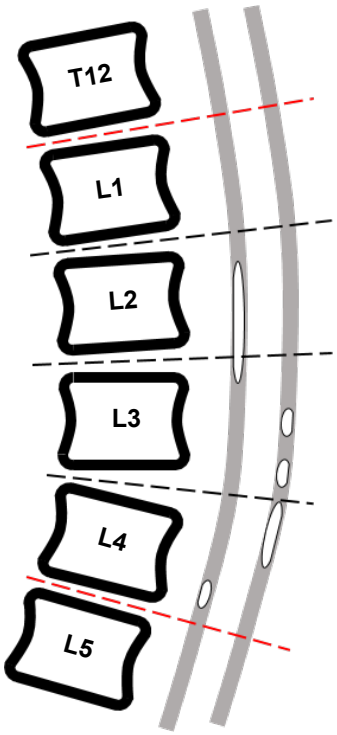

Figure 2.1: Diagram of AAC-24 score calculation for lateral radiography. Total length of calcification parallel to vertebral height is used to generate a 0-3 score for each wall adjacent to lumbar vertebrae L1-4. This example scores 9, The posterior L3 and L4 sections score 1. The anterior L3 and L4 sections score 2. The posterior L2 with more than two thirds of the section calcified scores 3.

that there is less operator precision with AAC-8, as more mental effort is required to estimate combined lengths of multiple lesions [35].

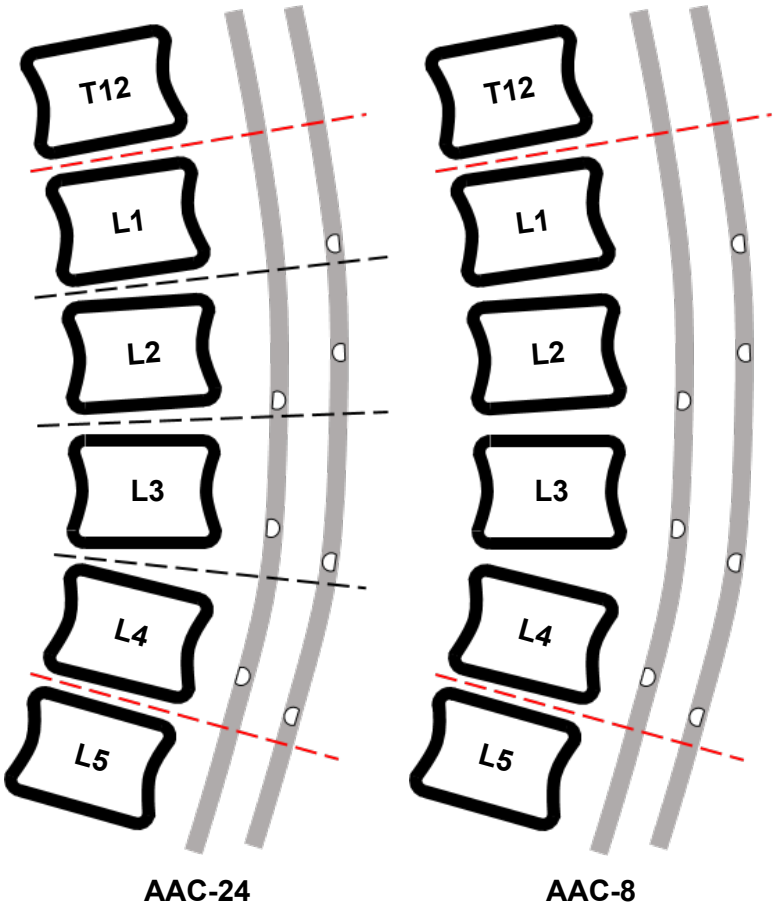

Figure 2.2: Example of the AAC-24 score mistaking mild diffuse calcification as severe. The AAC-8 score is less sensitive to calcifications of this type.

Additional measures have been developed for use in lateral radiographs. Previous to the development of the AAC-24 scoring system, a scoring system was developed while assessing the link between osteoporosis and aortic calcification [36]. This system consisted of three grades: 0 for no visible calcification, 1 for calcifications with a total length of up to 2 vertebral bodies, and 2 for total calcification above this length. Alongside this, the subjective opacity of the lesions was also incorporated. As part of the Rotterdam Study [37], investigating the predictive potential of measures of atherosclerosis, AAC was assessed with a variant scale [38]. In this system, absolute

length of calcification was used to provide a 0-5 score. A score of one for a single calcification of length 0.5-1.0cm, 2 for multiple calcifications with aggregate length < 2.5cm, 3 for < 5cm, 4 for < 10cm, and 5 > 10cm. Neither of these scores have had their reliability assessed in the literature, and have not been used beyond these studies.

The AAC-24 and AAC-8 scores are primarily aimed at providing a semi-quantitative measure of AAC in a timely manner. A more quantitative measure of AAC has been developed which incorporates information on the number and width of calcifications. Termed the Morphological Atherosclerotic Calcification Distribution (MACD) index, this scoring system demonstrates a significant improvement to prediction of cardiovascular disease mortality [39, 40]. For CVD death, the hazard ratio for each standard deviation increase in MACD score was 4.2, after adjusting for other risk factors. In contrast to the AAC-24 score, which considers only the affected wall segments, the MACD score is calculated by multiplying the number of individual calcified deposits with a simulated plaque area. The simulated plaque area estimates the extent of the atherosclerotic lesions, beyond what is visible, based on the extent and proximity of the calcifications. Although the study was limited to a niche cohort and could have benefited from a larger sample given its comparison to thoroughly researched scores such as The Framingham Score, it readdresses a potential benefit for using AAC in calculating cardiovascular risk and explores further information that can be gained from AAC distribution.

A major drawback of these semi-quantitative scores is a lack of sensitivity to small changes in AAC severity over time [35]. A discrete scale is blind to any changes too small to increase the score. With a slow but continuous process like calcification this is limiting. Consequently, follow-up times in studies must be much longer to obtain significant trends, increasing study dropout or masking the effects of potential interventions. The current scores were developed to allow ease of interpretation and consistency between observers. It is reasonable that once an automated system can

match a specialist human operator in performance, that a fully quantitative measure of AAC could be developed with exact measurements of affected aortic wall. A quantitative measure of AAC would allow robust statistical manipulation and more precise documentation of changes over time. An automated method for producing this and other scores which incorporate more detailed measures of atherosclerotic extent, could improve predictive accuracy without a large time cost for clinicians. With established and validated measures of AAC on lateral radiograph images of the abdominal aorta, these methods could be used in other modalities to quantify and compare consistency.

2.2.2 Dual-Energy X-Ray Absorptiometry

Dual-Energy X-Ray Absorptiometry (DXA) imaging is primarily used to measure bone mineral density (BMD) in osteoporosis screening. BMD is a measurement of the mineral content of bone tissues, low levels of which are associated with an increased chance of fragility fractures. Fragility fractures are a pathological variety of fracture that results from 'normal activity'. Typically, these are fractures of the spine, pelvis, neck of femur or wrist. These fractures can cost lives and independence and as a result those at risk are assessed for BMD and intervention is implemented where needed.

The National Institute for Health and Care Excellence (NICE) recommend that fracture risk is assessed in women over 65 and men over 75, as well as younger individuals with certain risk factors [41]. Risk assessment is performed using the FRAX scoring system, which uses factors such as age, weight, BMD and steroid use to produce a 10-year risk of fracture score [42]. BMD is assessed using DXA imaging to measure the calcium content of key anatomical locations, usually the hips and lumbar spine (L1-L4). BMD is compared to a reference value produced from an average BMD for a healthy 30-year-old. Osteopenia is defined as BMD lower than 1 standard

deviation below the reference mean of a normal young person of the same gender and ethnicity (T score); lower than 2.5 standard deviations below defines osteoporosis.

Often simultaneously, the spine is assessed for signs of vertebral fracture or collapse. Subclinical fractures of the vertebrae are an independent predictor of future fragility fractures and morbidity [43]. This is achieved using a lateral image of the lumbar spine, termed a vertebral fracture assessment (VFA) image. These VFA images often incidentally capture the abdominal aorta during screening, allowing access to DXA imaging of AAC without additional screening or radiation.

It has been demonstrated that increased severity of AAC is positively associated with vertebral fractures and negatively with BMD [44, 25]. A study by Szulc et al. [45] compared the severity of AAC and vertebral fractures on VFA. With a large sample size of 901 men above 50 years old, the study demonstrated that increasing severity of AAC is correlated with an increasing number and severity of vertebral fractures, even after controlling for age, BMD, comorbidities and history of falls [25]. This further demonstrates that the occurrence of cardiovascular disease, low BMD and vascular calcification are heavily interconnected. This increased risk of CVD in those with osteoporotic changes means that screening of individuals undergoing VFA imaging is already targeting an at-risk group. This will allow screening for CVD to utilise systems already in place to identify risk of osteoporosis. This is particularly important in female patients, where traditional risk factors are less predictive of CVD but prediction of cardiovascular events is more closely associated with changes in BMD.

DXA imaging is preferable to plain radiography for a number of reasons. The dosage in DXA acquisition is even lower, between 2μSv and 50μSv [46], allowing it to be more safely used for screening. DXA incorporates both a high and a low energy acquisition, each with different attenuation properties. Subtracting one image from the other produces an image less impacted by soft tissue noise for visualising and inspecting bones. Though typically, identification of AAC is performed on single-energy VFA

acquisitions. As densitometers are used for more specific tasks than radiographs, there is less variation in operating parameters leading to more inter- and intra-densitometer reproducibility [16].

Scoring of AAC in VFA images is performed using the same methods as in $2D$ radiograph images, namely the AAC-24 and AAC-8 scores. The correlation between AAC on VFA and a range of imaging modalities has been investigated. VFA has a good correlation with radiography, although VFA has a slight tendency to underestimate AAC compared to radiography due to less contrast for small calcifications and lower spatial resolution [31, 34, 47]. There is also good correlation between the severity of AAC in VFA images and calcification in the coronary arteries [22], a strong predictor of cardiovascular mortality. Comparison of AAC on VFA and the volumetric CT measurement has demonstrated significant correlation, in CKD patients [48, 49] as well as those without pathology [27]. A direct comparison of the two techniques is difficult, even with good agreement. Quantitative CT measurements are inherently more accurate than semi-quantitative VFA measurements. The high radiation dose and expense associated with CT prevents it from being used as a screening tool, allowing VFA a strong role as an alternative.

There are a range of systems and techniques used for VFA and BMD measurement. The majority of VFA images are obtained using single energy absorptiometry, particularly on Hologic densitometers [50]. In an attempt to limit the already minimal radiation exposure in VFA, some systems have a 'smartscan' feature that will limit the field of view to only the spine itself, which often excludes the aorta. It may be advantageous for future guidelines to recommend images be taken more optimally for AAC measurement. There is also some variation on patient positioning during VFA. Some densitometers can change the position of the acquisition arm and can obtain lateral images while the patient is supine. If the arm is fixed then the patient will have to lie in the lateral decubitus position. This could alter the position of the abdominal aorta

relative to the spine and the superimposition of gas in the bowels onto the aorta. The differences in equipment and techniques introduce variations in quality and artefacts for which any automated system would have to account.

2.3 Cardiovascular Risk

Cardiovascular diseases are the result of life-long gradual changes to the vasculature. The long asymptomatic phase of the atherosclerotic progression means that the first indication for intervention is often a major cardiovascular event. More than 50% of patients, 60% in women, who die of coronary heart disease have no prior symptoms [51, 52]. It has been established that these diseases are highly preventable, but a lack of clinical signs ahead of a major cardiovascular event hinders identification of at-risk individuals for intervention [53]. For this reason, the main method for assigning intervention strategies is calculating cardiovascular risk scores from clinical factors.

In the UK, the NICE guidelines recommend the use of the QRISK® score [54, 55]. Using risk factors such as smoking, diabetes status and lipid profile, the algorithm gives a risk of cardiovascular incident in the next 10 years. A percentage above 10% is used as the threshold for considering intervention such as statins. There are a number of similar scoring systems used worldwide with the majority using a similar combination of clinically assessed risk factors. Unfortunately, there are still a large number of cardiac events that occur in low risk individual. As many as 30% of individuals classified as low risk by clinical risk models go on to have cardiac events [27]. This indicates that there are other important risk factors that could help improve prediction and targeting of intervention strategies.

2.3.1 Role of AAC

With current CVD risk scoring, visualisation of atherosclerotic extent within the arterial system is not used. The severity of AAC has been found to be a predictor of future cardiovascular events, even when controlling for currently used clinical risk factors [56, 17, 57]. AAC is highly correlated with other predictors of CVD, such as coronary artery calcification [32]. AAC acts as a measure of atherosclerotic extent within the arterial system, allowing more accurate risk stratification and monitoring ahead of a major cardiovascular event. These factors have led to increased scientific interest in using AAC as a screening tool and incorporating it in current CVD risk assessment scores. A robust and well-designed meta-analysis examining the prediction of cardiovascular events using AAC was performed by Bastos Gonçalves et al. [58]. The analysis reviewed ten separate studies with more than two years of follow-up on patients not already in high cardiovascular risk groups, such as end stage renal disease. Although the study only examined AAC categorised as none/mild, moderate, and severe, as opposed to using an 8 or 24-point score, the amount of data covered and the correlation are convincing. The study shows that severity of AAC increases the risk of cardiovascular events and mortality. Controlling for other known risk factors, the paper confirms that AAC severity is an independent predictor of CVD. It is however, worthy of note that the risk of stroke was only increased with severe AAC, likely owing to the complex aetiology of stroke.

The correlation between a range of cardiovascular diseases and AAC severity has been well demonstrated. These correlations are stronger for coronary events, such as myocardial infarction [59, 60], than stroke. Wilson et al. [56] found that the risk of cardiovascular mortality was more than double in those with an AAC-24 score of more than 5. AAC severity has also been positively associated with more chronic forms of CVD. Increased rates of congestive heart failure were found as part of the Framingham

Heart Study [61]. Across all grades, AAC severity independently predicted an increasing rate of intermittent claudication, a sign of vascular insufficiency in the lower limbs [28].

Atherosclerotic changes are gradual and the progression involves a long preclinical phase, it often presents with a major incident. The inclusion of a predictor that can more directly demonstrate subclinical atherosclerotic changes could improve targeting of preventative therapies. Despite the benefits of AAC assessment and recommendations for its use in CVD risk stratification [50, 61, 29, 16, 62], it is still not routine for DXA reports to include more than qualitative comments on AAC, as its discovery is secondary to the principle investigation. There is definite potential for computer vision techniques to locate and measure calcification in the abdominal aorta and report this automatically. With well-trained models, large volumes of VFA images could be quickly and consistently assessed for AAC and facilitate earlier targeting of intervention.

This chapter has demonstrated that vascular calcification is the end stage of atherosclerosis and has a role in impeding arterial function, as well as detailing the methods through which AAC can be imaged, quantified and used to improve cardiovascular risk predictions. The following chapter details the potential techniques to automate the identification of AAC, with later chapters covering experiments to implement these techniques.

Chapter 3

Technical Background

With the clinical significance of the task established, this chapter explores the technical aspects of approaching the problem of automated analysis of abdominal aortic calcification (AAC). It introduces key literature surrounding the methodological decisions made.

The three major tasks of automated scoring of AAC in this work are: localisation of the abdominal aorta, annotation of calcification in the aorta, and conversion of annotated calcification to a clinically useful score. This chapter explores these areas of computer vision, with a significant focus on semantic segmentation techniques. The context of these tasks is first established, the approaches to solve them are then covered in more detail, ending with a review of previous attempts at automating this task.

3.1 Medical Imaging and Computer Vision

Medical imaging is a broad field, covering techniques used to create visual representations, assess the function, and diagnose disease processes of the interior of the human body, its organs and its tissues. Chapter 2 has already introduced a range of x-ray based

radiography techniques, including dual-energy x-ray absorptiometry (DXA) and computed tomography (CT). However, the modalities of the field also include techniques such as magnetic resonance imaging (MRI), ultrasound, and nuclear medicine; and extend to techniques such as electrocardiography (ECG) which while not producing an image in the traditional sense, still provide a visual representation of underlying physiology. The overall goal of medical imaging techniques is to diagnose, treat, quantify and monitor disease.

The field of computer vision seeks to enable computers to interpret digital images and videos, and use this information to infer information about the world. This process is often described as developing methods to recreate the capability of the human visual system, which has evolved to incorporate many specialised processing systems which enable effortless interpretation of the world. This task has proven to be deep and complex, requiring decades of study. The pursuit of human level perception has lead to many advancements as well as giving insight into the biological systems used in nature, and will continue to be an active area of research, as the problem remains unsolved.

Computer vision can be considered a sub-specialty of machine learning, and incorporates techniques from a range of fields including artificial intelligence. The challenges of computer vision can be roughly categorised based on the intended output. A fundamental challenge is that of classification. Classification is the inference of which class an image belongs; such as which handwritten number is in an image. A further challenge is that of segmentation, interpretation of which pixels in an image represent a given object; such as which areas of the video contain other cars or pedestrians.

The application of these tasks to medical imaging is a rich area of scientific investigation. Computer vision within medical imaging aims to increase the amount of clinically relevant information which can be acquired and the speed at which this can be achieved. Many applications seek to augment the decision making of clinicians,

allowing access to more information to affect management decisions. These applications are varied, and include: computer aided detection to reduce false-negative events in image interpretation, registration and warping between different imaging modalities or models to incorporate additional information or allow volumetric measurement of organs or tissues, and content matching to retrieve images of the same disease process for comparison.

The remainder of this section concentrates on the computer vision approaches useful in the automated analysis of abdominal aortic calcification. Techniques used to localise anatomical landmarks in images based on statistical models are explored, along with image interpolation and warping techniques to fit anatomical regions to a useful framework. An overview of medical image segmentation is then covered, giving a foundation for the detailed examination of segmentation techniques in the rest of the chapter.

3.1.1 Anatomical Shape Modelling

An important area of computer vision in medical imaging is the identification, localisation and segmentation of anatomical landmarks in an image. Shape models are a statistical method to describe and analyse a sample of shapes, representing the mean shape and quantifying the degree of variation between shapes. These shape models can then be used to localise new examples of the defined shape, to segment regions of the brain for example [63], or to predict the location of structures based on their statistical relationships. In the context of AAC on DXA images, it is useful to use skeletal landmarks to identify the location of the abdominal aorta.

A foundational model in this domain is the point distribution model (PDM) [64]. The model is built from a number of example shapes, where each shape is defined by a finite number of landmark points. Each landmark point represents a consistent location

on the modelled object, for example an extreme of the articular surface of a bone. Each shape can then be represented by n points in d dimensions. A shape in two dimensions can be expressed with a $2n$ element vector:

$$\mathbf{x} = (x_1, x_2, ..., x_n, y_1, y_2, ..., y_n) \tag{3.1}$$

With a set of s training example shapes, Procrustes analysis is used to convert these shapes into a common co-ordinate frame. The sum of distances of each shape to the mean is minimised, with the mean at the origin and unit scale. This set of shapes $\mathbf{x}_i$ can be represented as s points in a $2n$-D space. Using principal component analysis, the cloud of points is transformed into a space with fewer dimensions which encode the most significant modes of shape variation. Each image can be approximated by the mean shape:

$$\bar{\mathbf{x}} = \frac{1}{s}\sum_{i=1}^{s} x_i \tag{3.2}$$

With the covariance of the data described by:

$$\mathbf{S} = \frac{1}{s-1}\sum_{i=1}^{s}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T \tag{3.3}$$

The eigenvectors, v_i, and eigenvalues, λ_i of the covariance are calculated, and arranged in descending magnitude of λ. These sorted eigenvectors represent the principal modes of variation of the shape model, and an example of their appearance is demonstrated in Figure 4.9. With a set of eigenvectors $\mathbf{V}$ with the largest eigenvalues accounting for the required proportion of variance, an example from the training set $\mathbf{x}$ can be approximated with:

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{V}\mathbf{b}$$

$$\text{where:} \tag{3.4}$$

$$\mathbf{b} = \mathbf{V}^T\mathbf{x} - \bar{\mathbf{x}}$$

The vector $\mathbf{b}$ represents the parameters of a deformable model, which can be adjusted to yield new examples of $\mathbf{x}$. By limiting each parameter b_i based on the variance of the parameter in the training set, given by λ_i, shapes generated from the model will be similar to the training shapes.

Active shape models (ASMs) [64] use these statistical shape models to locate an example of the shape in a new image, starting with a mean shape and iteratively deforming it in compliance with the model. ASMs attempt to converge on strong edges within the image, associated with each point of the model. Later, the more robust active appearance models (AAM) were developed which used the textural information across the image [65]. In addition to landmark points, intensity variation is encoded in the statistical model and used to predict the location of objects in novel images.

The constrained local model (CLM) is an extension of AAMs and ASMs for locating landmark points [66]. When the model is constructed from training examples, patches are sampled around landmark points to incorporate texture information. With a candidate position for the shape on an image, each landmark has a local texture model which calculates a cost for placing the point at any given pixel, creating a response image. This cost can then be minimised for all landmarks while constraining with the shape model.

Regression voting was incorporated into this technique, to create Random Forest Regression-Voting Constrained Local Model (RFCLM) [67]. In this technique a random forest regressor is trained independently for each landmark point, with trees trained on patches collected at many randomly displaced positions around the point.

Haar wavelets (both random forests and Haar-like features are discussed in Section 3.2) are used from these patches to train the trees to predict the target point in a given patch. When fitting the model to a new image, each tree votes on a candidate location in the patch around a point to create the response image. The RFCLM technique has been used in a range of medical imaging applications due to its high accuracy for detection and localisation of skeletal landmarks [67, 68, 69]. RFCLM has also demonstrated accurate localisation of vertebrae on DXA VFA images [70].

These techniques have been used in a variety of applications. With robust localisation for skeletal and tissue landmarks, these models can be used to transform and register images to assist in clinical applications.

3.1.2 Image Registration and Warping

A number of imaging modalities are used in disease monitoring and diagnosis, as each modality has its own advantages and drawbacks. Often, multiple modalities may be used to image the same tissues, to combine the information made available by each. In such cases, for example when performing simultaneous CT and MRI acquisition [71], image registration is employed to maximise the utility of this data. Image registration is the process of transforming the geometry of an image to match another, allowing overlay of anatomical structures for improved visualisation. Additionally, images can be registered onto anatomical models. This is common in brain imaging, where there is substantial variation in brain size and shape. Mapping brain regions onto a consistent brain atlas allows comparison between patients or modalities [72]. Image registration techniques are applicable to the problem of automated AAC analysis as a consistent position of the spine and abdominal aorta can aid in quantifying and scoring calcifications. Image registration represents a function f which can map values from a source image s to a target image or atlas t:

$$f(\mathbf{s_i}) = \mathbf{t_i} \quad \text{where} \quad i = 1, 2, ..., n \tag{3.5}$$

As has been discussed, shape modelling allows the fitting of consistent landmark points to new images based on a statistical model. In this context, each landmark point in the source image s_i represents the same anatomical region as the corresponding point in the target image t_i. An image registration function can be defined using a range of geometric transformation functions, such as linear or affine transformations. These transformations are insufficient in any situation involving localised transformations in separate parts of the object. Non-rigid transformations of images can be defined using landmark points to allow registration of separate objects and sections within the same image.

The thin plate spline (TPS) algorithm is a commonly used landmark based registration and interpolation technique [73]. A thin plate spline has a gradient based regularisation term which controls smoothness, designed to replicate the bending of a thin sheet of metal. The TPS algorithm registers corresponding landmark points from a source image to a target image, and defines a unique interpolation for image content away from the landmarks. The function minimises the distance between corresponding landmark points s_i and t_i, while minimising the distortion of the space around the landmarks. This is achieved by finding a function which minimises the energy:

$$\mathbf{E} = \mathbf{E_f} + \lambda \mathbf{E_d} \tag{3.6}$$

Where $\mathbf{E_f}$ measures the goodness of fit for corresponding points, using sum of squared distances. $\mathbf{E_d}$ discourages distortion of the space, using the integral of the square of the second order derivative. λ is a weighting constant controlling the extent of non-rigid warping. This gives:

$$\mathbf{E_f} = \sum_{i=1}^{n} ||f(s_i) - t_i||^2$$
$$\mathbf{E_d} = \int\int \left[\left(\frac{\delta^2 f}{\delta s_x^2} \right)^2 + 2 \left(\frac{\delta^2 f}{\delta s_x \delta s_y} \right)^2 + \left(\frac{\delta^2 f}{\delta s_y^2} \right)^2 \right] ds_x ds_y \tag{3.7}$$

The interpolation function used to map source to target points minimising this energy consists of two components. An affine transformation which encompasses the global transformations across the image, and a non-linear deformation based on radial basis functions. This is represented by the equation:

$$\mathbf{t_i} = a_0 + a_1 s_i^x + a_2 s_i^y + \sum_{j=0}^{n} c_i \phi(||\mathbf{s}_i - \mathbf{s}_j||) \tag{3.8}$$

Where a represents the parameters of an affine transformation matrix. c is a mapping coefficient for each landmark, and $\phi()$ is the radial basis kernel for TPS, $\phi(r) = r^2 \log r$. Subject to these constraints and additional orthogonality conditions a unique function can be defined which maps positions in the source image to the target image while minimising the energy function. The use of the radial basis kernel ensures that points further from the landmark points are impacted more significantly by the global affine transformation, creating local transformations. The end result is a smooth transformation which is defined entirely by the landmark points with no need for manually selected parameters. The unique solution to TPS, and the lack of any tuning parameters gives it an advantage compared to other available non-rigid warping techniques.

This section has explored the value of registration and interpolation to enable further analysis of images. Once an organ or region of interest has been defined, another challenge is in segmenting areas in this region based on their function, structure or a disease process. This is known as semantic segmentation and is the focus of the next section.

3.1.3 Semantic Segmentation

Segmentation involves techniques that can separate regions of an image based on the content of those regions. In medical image analysis, segmentation can identify different organs, to locate and measure them, or tissues, to identify anomalies such as tumours. In a practical context, a segmentation technique is one which produces a class prediction for each pixel in an image based on shared characteristics. There are a multitude of techniques used to solve this problem, and it is an active field due to the complex nature of the problem.

Early segmentation techniques concentrated on the use of intensity thresholding and edge detection. Threshold techniques, where regions are classified based on their intensity, are the simplest forms of these segmentation strategies. This can be achieved using multiple thresholds and histograms to separate pixels of similar intensity into classes [74, 75]. Edge detection makes further use of intensity values by using the differential of intensity change over images to distinguish borders between regions, with the intention that sharp contrasts represent different tissues. While there exist multiple algorithms for edge detection, the general principal remains in finding a threshold for the rate of change which defines an edge and growing regions in the images bound by these edges.

Both thresholding and edge detection approaches are susceptible to noise in images. Increasing the amount of context for segmentation decisions was the strategy to combat this shortcoming. Texture based features use combinations of higher-order statistics of intensity values, and their spatial relationships. Features can then be used either by a classification algorithm or a clustering algorithm, to assign each pixel a class. Clustering approaches, such as k-means clustering, require no training data and separate pixels based on their proximity in the chosen feature space. Classifiers are supervised methods, requiring the use of training data, manually segmented images

which represent the ground truth. These methods also depend on the features available to distinguish pixels as different classes. As covered in the previous section, it is possible to register landmarks in an image onto an atlas, to segment regions of interest based on a prior shape model.

Much of the work being done in this field is in trainable models which do not require predefined crafting of features. The most recent and promising revolution in this domain is in the use of neural networks, though there is still plenty of work ongoing in ensemble methods such as random forest, and in the development of more informative image features for simpler algorithms. The aim of this chapter is to show how segmentation algorithms can solve the problem of automating abdominal aortic calcification quantification.

AAC Segmentation

In this work, automated scoring of abdominal aortic calcification (AAC) is approached as a semantic segmentation problem. While the ultimate aim is to classify AAC using the AAC-24 semi-quantitative score, there are advantages to achieving this using segmentation. Direct classification of AAC requires representation of all classes in the training data, in sufficient quantity. Severe calcification is much rarer than mild in the population, and so high AAC-24 scores are much less common, with sparse representation for all classes in the severe class (AAC-24 scores 7-24).

Additionally, the AAC-24 score was developed for the ease of a human interpreter. The exact conversion from identified calcification and class is well defined and so does not need to be approximated with another function. With segmentation masks, it is possible to change the scoring system without making the predictor obsolete, and indeed the segmentation masks can be used to better develop fully quantitative measures of AAC. Segmentation also allows demonstration of the regions identified as calcification to a clinician. This interpretability is valuable when making clinical decisions, as

it allows increased transparency and trust in decisions made by the system.

The primary focus of this work is assessing the accuracy with which segmentation strategies can separate out calcification. The success of segmentation algorithms is measured using various metrics, by comparing the produced segmentation to a ground truth. In the example of a binary segmentation, where the classes are positive or background, metrics for performance can be defined by comparing pixels assigned the positive class by the ground truth, G, and those predicted positive by the segmentation model S. A simple statistic for comparison of the overlap in predictions is the Intersection over Union (IoU), also known as the Jaccard similarity coefficient. This is defined as the total area of agreement for pixels in the positive class, intersection, divided by the total number of pixels predicted positive by either the ground truth or the segmentation model, union. This can be expressed as:

$$\textbf{IoU}(G,S) = \frac{|G \cap S|}{|G \cup S|} = \frac{|G \cap S|}{|G| + |S| - |G \cap S|} = \frac{TP}{TP + FP + FN} \tag{3.9}$$

Where true positive, TP, examples are those pixels which both ground truth and segmentation agree on the positive class, and true negatives, TN, where both agree on the background class. False positive, FP, pixels are those which have been incorrectly identified by the prediction as belonging to the positive class, and false negatives, FN, the pixels labelled background which are positive in the ground truth. Using these definitions allows simple calculation of segmentation accuracy:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.10}$$

This is the number of pixels on which the ground truth and segmentation have assigned the same class, divided by the total number of pixels in either. Similarly, the true positive rate and false positive rate of the segmentation performance can be calculated from these definitions, along with recall and precision. F_1 score is the harmonic mean

of precision and recall, in the context of segmentation accuracy it is often referred to as the Dice-Sørensen Coefficient (DSC). DSC can be expressed as:

$$\mathbf{DSC}(G,S) = \frac{2|G \cap S|}{|G| + |S|} = \frac{2TP}{2TP + FP + FN} \tag{3.11}$$

This definition makes it clear that the DSC is similar in nature to the IoU, with neither impacted by the number of true negative examples. This can give valuable insight into the performance of predictors in problems with a small number of the positive class compared to the image size, a common scenario in medical image segmentation. The metrics can also be extended to problems with multiple classes. These statistical measures of segmentation performance allow the comparison of different segmentation algorithms, to identify the most reliable for a given application, such as the identification of AAC in VFA images.

With the general problem of semantic segmentation described, the following sections cover some of the most promising solutions to the problem of segmenting abdominal aortic calcification, concentrating on random forests and convolutional neural networks.

3.2 Random Forests

Random forests are an example of an ensemble learning method, algorithms which combine a set of individual classifiers or regressors to improve decision making. In the case of random forests, the constituent parts are known as decision trees. Each individual decision tree is trained to predict a classification or continuous output. The output of the overall forest is the modal or mean prediction of the decision trees, for classification or regression respectively [76, 77].

3.2.1 Training and Inference

Decision trees are an intuitive model for classification or regression, consisting of a number of binary queries of the data to produce a prediction. The constituent parts of a typical decision tree, are: split nodes, where a test is applied to a feature of the data; branches, representing the results of the tests and linking between nodes; and leaf nodes, the final prediction at the terminal branches. The depth of a node is the number of nodes passed through to reach it. With training data T, consisting of i examples, with each example consisting of j features $(x_{i,1}, x_{i,2}, ..., x_{i,j})$, and a label y_i, an arbitrary binary test, b, can be applied:

$$b^{j,l}(\mathbf{x}) = \begin{cases} \text{true}, & \text{if } l < x_{i,j} \\ \text{false}, & \text{otherwise} \end{cases} \qquad \forall i \tag{3.12}$$

Where the choice of limit, l, and feature, j, are used to differentiate examples with different labels, y_i. This example node will then split the training data into two branches, each of which will be split further by an additional node, and so on. This continues until a predefined depth has been reached, or the remaining examples reach an acceptable uniformity, terminating in a leaf node predicting the plurality label. The choice of binary test b^{θ} (where $\theta = (j,l)$) is optimised at each node based on the partition of the training data it receives, maximising an optimality criterion. The optimality criterion encourages the choice of b which best reduces the uncertainty of the labels. This is commonly achieved using Information Gain (IG) [78], where:

$$IG(\theta, T_n) = H(T_n) - H(T_n \mid b^{\theta}(\mathbf{x})) \tag{3.13}$$

Where $H()$ is the entropy function, and T_n represents the portion of training examples available to node n. The optimal choice of parameters θ maximise the IG

function, discriminating between labels. Once the IG is below a chosen threshold on a given branch a leaf node is created, which can provide a class prediction based on the modal label, or a probabilistic estimate of y based on the proportion of each label in the training examples which end at the node.

In order to predict the label of a novel example, the binary decisions are made at each node based on the features of the novel example until reaching a leaf node, which will define the most probable label. However, decision trees do have a tendency to overfit to the training data. With enough complexity a tree will fit perfectly to the training data, but will be unlikely to generalise to novel examples [79]. To combat this inclination, a random forest uses the average decision of many trees. The number of trees used to build the forest will affect the performance of the model. Increasing numbers of trees will reduce the error rate of the model, but with diminishing returns [80]. Therefore, the number of trees can be determined experimentally, or the maximum number of trees can be chosen based on the available time and computational resources.

With a deterministic method of construction, all trees would be identical, to create the random forest each tree is trained on a bootstrapped sample of the training data [76, 79]. Examples from the original training data are sampled with replacement, until a new set is produced of equal size, T'. Known as bagging (**b**ootstrap **agg**regation), this technique improves performance as each tree will be sensitive to training data specific features, but the average of many trees will avoid this. Taking the average across the forest reduces the overall variance, without increasing the bias [76]. However, this relies on the assumption that the trees are reasonably independent.

With bagging, there is considerable correlation between trees, due to sharing a large proportion of training data. To increase independence of the trees, and the performance

of the forest, feature bagging is implemented. Feature bagging is a subsampling strategy applied to feature choices at each node [81]. A random subset of features is presented to a given node, $\mathbf{x}^*$, in order to choose the one that maximises the optimisation criteria:

$$IG(\theta, T'_n) = H(T'_n) - H(T'_n \mid b^{\theta}(\mathbf{x}^*)) \tag{3.14}$$

In the case that there are a few very strong predictors in the features, this strategy avoids the majority of decision trees being heavily dependent on these few features. There is some variation in the literature as to how many features should be used at each node, ranging from the square root of the total, to half [81, 77]. Unlike the number of trees in a forest, the extent of feature bagging is not monotonic, and should be determined experimentally [82].

With an understanding of how the model is built, the most important step is selecting features from the training data that will be used to produce the random forest.

3.2.2 Haar-like Features

The success of any machine learning method relies heavily on the selection of training data and identification of informative features. For a traditional classification task, such as identifying individuals with cardiovascular disease, available features will involve factors such as blood pressure, age, the presence of family history of CVD, and so on. The random forest will identify the most informative features and predict based on these. In the context of image segmentation, features must be extracted from the image and provided to the random forest in a usable form. This section will concentrate on features which are common in image processing tasks.

The label of the pixel in a given image is unlikely to be independent of the pixels in the rest of the image. Features gained from only the coordinate of the label do not

take advantage of the context available. An alternative is to use a patch of pixel values around the label pixel and extract features from these intensity values in order to make a prediction.

A popular feature which can be extracted from images or image patches is the Haar-like feature. Haar-like features are a simple and fundamental image feature produced by comparing the summed intensities of adjacent rectangular regions of an image. The basic arrangements of rectangles to produce Haar-like features are shown in Figure 3.1. The two rectangle arrangements are useful for detecting edges, whereas the three rectangles are sensitive to horizontal or vertical lines.

A Haar-like feature is a single number value produced from summing pixel intensities. As each feature contains relatively little information, many of these features are produced for an image. In the Viola-Jones object detection framework, each arrangement of rectangles is applied to the image in every combination of scale and translation to produce an over-complete set of features [83]. This framework was successful in the domain of real-time face detection. Random decision forests have also been successfully implemented with Haar-like features in a range of applications in medical imaging [84, 67, 85].

Integral images are used to improve the speed of Haar-like feature calculation. An integral image is a summed area table of the original image data, where each pixel is the sum of the intensities of all pixels in the original image up to that coordinate along the x and y axes. This means that a pixel of an integral image $I(x,y)$ is the sum of the intensity of each pixel in the original image $i(x,y)$ which are bounded by a rectangle with corners $(0,0)$ and (x,y):

$$I(x,y) = \sum_{x'=0}^{x} \sum_{y'=0}^{y} i(x',y') \tag{3.15}$$

This allows the rapid calculation of an integral image in a single pass across the

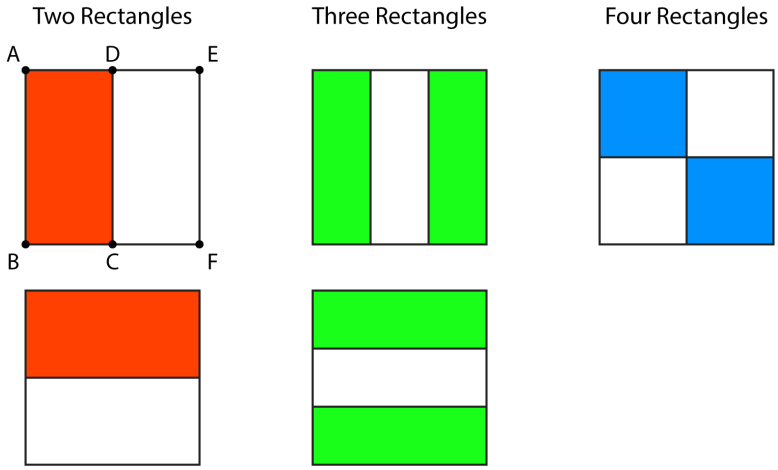

Figure 3.1: Examples of Haar-like feature arrangements. Each feature is a single value produced by subtracting the sum of the white area from that of the shaded area. Using integral images, a two rectangle Haar-like feature can be calculated rapidly as: $(A+C-D-B)-(F+D-C-E)$

original image, using:

$$I(x,y) = i(x,y) + I(x-1,y) + I(x-1,y) - I(x-1,y-1) \tag{3.16}$$

Use of integral images greatly decreases the computation time for summing regions to produce values for Haar-like features as the sum of intensities of any rectangular area of the image can be calculated from the four corner values of the rectangle instead of using all values within the rectangle. As demonstrated in Figure 3.1, a two rectangle comparison can be calculated from only six values. This allows Haar-like features to be used to rapidly gather a large number of features to train a forest with context from an area surrounding a labelled pixel. The speed at which these features are calculated is particularly beneficial when real-time testing of new examples is required, but they have been used to great effect in medical imaging applications.

Additions to, and variations of, the concept of summed areas in a region of interest are a common feature selection strategy. The next section covers the development of these additions and how random forest has been used within medical imaging.

3.2.3 Random Forests in Medical Image Segmentation

Random forest has been successfully implemented for the detection and localisation of landmarks and boundaries in medical images [67, 70, 86, 87]. These applications used Haar-like features to encode local and contextual information extracted at offsets from the landmark points. Predictions were made using regression voting on each point in the image to predict the offset of the landmark point. These independent predictions were then aggregated to give an overall prediction and a measure of confidence.

Semantic segmentation is an area where the application of random forest has made a significant impact. An early paper in this area used binary classification of individual voxels to segment myocardial tissues in 3D echocardiogram images [88]. This approach trained a random forest using features summed from integral images in random box sizes around a labelled voxel. Additionally, the coordinates of the voxel itself were fed as features to the random forest to inform the segmentation based on the location of the patches.

Additional texture information can be used depending on the image modality. Magnetic resonance imaging (MRI) can capture multi-channel data which can yield additional features [89, 90]. In the segmentation of stroke lesions, features were combined from T1-weighted, T2-weighted, fluid attenuation inversion recovery and apparent diffusion coefficient images. Outside of additional textural features, the main additional features are spatial and lesion probability priors [85, 89, 91, 92]. Stroke lesions were segmented with the addition of lesion likelihood maps from Bayesian-Markov random fields.

Segmentation of multiple sclerosis lesions in magnetic resonance images has been achieved with a random forest Haar-like feature approach and similar contextual features [85]. This approach added spatial information to attempt to include more global context, sampling features in a Haar-like manner in the patch of interest, as well as a

patch in the corresponding area of the other brain lobe (lesions being unlikely to be bilateral). Additional spatial prior information was also included by mapping image information to an atlas which informed the probability of an area containing grey or white matter. This technique of using atlas based spatial priors has been used outside of the brain, for the location of mediastinal lymph nodes in CT images [92].

These segmentation approaches highlight the strength of random forest, and its flexibility to incorporate image data as well as spatial information and the output of other models as features. In the context of AAC segmentation, a lack of consistent shape and location to the calcifications renders direct landmark based approaches unfeasible. However, successful segmentation of lesions and tumours indicate that structures can be reliably identified based on patch based texture features with the potential for additional features based on spatial relationships and the nature of the imaging modality.

Random Forest and Deep Learning

Despite the many successes of random forest approaches to segmentation, this is an area which is now dominated by deep learning algorithms. Deep learning is a branch of machine learning based on artificial neural networks. Artificial neural networks were ostensibly inspired by information processing in the brain, constructed from many simple functions termed neurons [93, 94]. Deep learning distinguishes itself from many traditional machine learning algorithms by moving away from preselected features, and toward automated synthesis of relevant features from raw data and interpretation of those features in the context of the given task. The popularity and superior performance of these models does not however, mean that random forest is without advantages over deep learning approaches.

Random forests are much faster to train and test on modest hardware. Particularly

with healthcare applications in mind, where access to GPUs with large memory capacity for neural network training is limited, random forest training and prediction is far more feasible. Alternatively, with access to multiple computers, decision trees in the forest are perfectly parallelisable, as each tree is trained independently of the others. This allows trees to be trained on subsets of training data and combined for testing, to further speed up training.

There are additional qualities of the random forest that make it more suitable to medical imaging and healthcare challenges. The demand for training data in deep learning models is far greater, in order to identify sensible and generalisable features. This allows neural networks to succeed in areas with large datasets readily available. Datasets in medical imaging however, are often far more limited as they require expert acquisition and interpretation, and their availability is more restricted. Random forests can achieve better performance in situations where there are few training images.

There has been some interest in combining the advantages of the random forest with deep learning models. Deep learning layers have been used in random forests to learn internal representations of the input data, allowing the system to encode features for making split decisions [95, 96]. Random forests can also be used as the final layer of a neural network, similarly using the features encoded by the network and identifying those with the largest influence on the label [97].

Deep learning approaches continue to be developed for the domain of image segmentation, and much of the state-of-the-art performance is achieved in this way. The sections that follow introduce the literature of deep learning and explore a number of architectures which have been used for image segmentation.

3.3 Deep Learning and Neural Networks

Neural networks have proven effective in identifying complex patterns in high-dimensional data. This has meant that the tasks to which neural networks have successfully been applied, are many and disparate. Neural networks have proved a valuable tool in areas such as language translation [98], medication design [99], and content recommendations on video streaming services, which are best left unnamed to avoid accelerating the aging of this work. This section gives a brief history of neural networks; their structure and design; and the methods for training and optimising them.

3.3.1 Overview of Neural Networks

The theoretical potential of neurons was established in the 1940s [100, 101]. With the first neural networks, including the perceptron, applied to real world tasks in the 1950s and '60s [93, 94]. During this era, the limitations of these techniques were highlighted [102], failing to match theoretical potential. At the time, it was assumed that the limitations of perceptrons applied to all neural networks, resulting in a decline in interest in this area for some time.

The basic structure of the neuron is a simple function that takes a number of input values and produces an output, y. For each input x, a neuron assigns a weight w, the weighted sum of these inputs and a bias term provide the output. This can be expressed as $y = f(\mathbf{w}^T\mathbf{x} + b_0)$. Learning within the neuron can then be done by adjusting the weights applied to each input and the bias term, to bring the output closer to the desired value. Here f represents an activation function to normalise the output, typically between 0 and 1 using a function like the sigmoid. The activation function is also used to introduce non-linearity, important for learning complex representations within the data. The input to a neural network can take many forms, in the example of image analysis, each input can be the element of a matrix containing the image pixel values.

Importantly, the input of a neuron can be the output of another neuron, allowing layers of neurons which take an input, pass it to other layers, and finally to an output layer. A resurgence of interest in neural networks came with the introduction of multi-layered networks, which were made possible with the introduction of backpropagation [103].

Backpropagation allows training of the hidden layers of multi-layer networks based on a differentiable measure of the error. The training of a neural network essentially involves adjusting weights across all neurons in the network in order to minimise an objective function across examples in a given dataset. A common example of this loss function would be the mean squared error, it can be used because it is differentiable and can be expressed as the average of loss functions for each individual training example. The gradient of the loss function can be assessed relative to each weight in the output layer of the network, indicating how these values could be adjusted to improve the loss function. With gradients calculated for the output layer, these can be used to calculate the error gradients for each neuron in the preceding layer, working backwards through the network until reaching the input layer. These calculated gradients can then be used by an optimisation algorithm, such as stochastic gradient descent to adjust weights iteratively, improving the loss function. These algorithms are discussed in more detail in Section 3.3.3.

This is the basic process of training neural networks to approximate a complex function, mapping inputs to outputs in a dataset. These networks have become increasingly complex and specialised, based on the data being used and the problem at hand. In particular, image data is challenging to implement in fully connected networks, as the number of connections between neurons becomes quickly overwhelming. To handle these data, the convolutional neural network was developed.

3.3.2 Convolutional Neural Networks

Convolutional neural networks (CNN) are specifically adapted to image recognition tasks such as classification [104]. The layers of a CNN are not fully connected, instead utilising convolutional layers. A convolutional layer consists of a number of learnable kernels or filters. In the forward pass of the network each kernel is passed over the input to that layer, producing a dot product response at each spatial location. This produces a feature map of the input with similar spatial dimensions and a third dimension equal to the number of kernels in that layer. Kernels are usually small matrices, typically 3x3. Connecting multiple convolutional layers together allows the kernels of the next layer to be passed over the feature map of the previous layer. During backpropagation, the weights in each kernel are optimised to minimise the loss function, training the network to learn informative filters that are convolved across increasingly abstract feature maps to identify high-dimensional patterns in the image data.

Pooling kernels are utilised to reduce the spatial dimensions of the abstract representations, using non-linear downsampling. Typically maximum pooling is employed, which takes the maximum value at each non-overlapping position of the kernel. This strategy is used to reduce the size of the feature maps until reaching a manageable number of features for use with traditional fully connected layers. Fully connected layers are used as the final predictive layers of the network, the feature map of the final convolutional layer is transformed into a single vector, with every element used as the input to a layer of neurons. These fully connected layers can be repeated until the final layer which gives the output. Backpropagation can then be used in the same manner as a traditional neural network to iteratively reduce the loss function and encourage the network to learn kernels which can perform successfully on novel images.

Essential to the training of CNNs, is the use of non-linear activation functions. Similar to those used in fully connected networks, the responses for each layer of

the network undergo normalisation. These functions are usually applied immediately following convolution, to each value in the feature maps. There are numerous examples of activation functions, with sigmoid and tanh functions being popular traditionally. The disadvantage of these functions is that gradients are large in the middle of the function, but decrease towards extreme values. If the gradients are very small, then the propagation of gradients between repeated layers of this function will compound this problem and the gradients calculated through backpropagation will rapidly approach 0, and training will halt. This is known as the vanishing gradient problem [105].

A commonly used approach to combat this problem in CNNs is the use of an alternative activation function, the Rectified Linear Unit (ReLU) [106]. The Relu is a simple function which takes the form: $f(x) = \max(0,x)$, outputting 0 for all negative values of x. This function is cheap to compute, improving training time. It has also been shown to improve the rate of convergence as the slope does not saturate for large values of x, minimizing the vanishing gradient problem. The flat negative side of the ReLU encourages some degree of sparsity in the network. This encourages neurons to only respond for meaningful aspects of the problem and to reducing overfitting. The disadvantage is that for large learning rates, the weights of some neurons can enter this flat area of the function and become trapped, failing to learn. CNNs have been successfully implemented on countless image processing tasks, and are particularly popular for image classification tasks since the impressive performance achieved by Krizhevsky et al. in the ImageNet challenge [107]. CNNs can be modified to estimate bounding boxes around classes for some degree of localisation [108]. However, in order for these networks to perform pixel level segmentation, a patch based approach is required [109], classifying each pixel in an image. To allow image level predictions of segmentation, fully convolutional networks were developed.

Fully Convolutional Networks

Fully convolutional networks (FCNs), as the name implies, are CNNs which forgo the fully connected layers when predicting their outputs, in favour of additional convolutional layers [110]. These networks were designed for the task of semantic segmentation, allowing end-to-end training on whole images with ground truth annotations. In order to achieve an output resolution which matches the input, the fully connected layers of a CNN are replaced by upsampling layers.

Various approaches to upsampling exist, traditional techniques such as interpolation (nearest-neighbour, linear, cubic) involve no learnable parameters. Transposed convolution is a convolution operation which is in essence a reversal of regular convolution, which allows a single element in a feature map to be transposed into multiple [111]. Transposed convolution is also often termed fractional stride convolution, up-convolution or deconvolution, a misleading term as it is not truly a deconvolution in the image processing sense. A kernel is effectively applied at fractional strides across the feature map to increase the number of samples and increase the spatial resolution of the feature map. The kernel used to perform this upsampling has trainable weights in the same way as any other, thus allowing backpropagation of gradients throughout the network.

The overall structure of these networks can then be described with two parts. A contracting or encoding path, which involves multiple convolutional and pooling layers akin to a traditional CNN, and an expanding or decoding path, built with transposed convolutional layers. With excessive pooling, the shrinking of feature maps can lead to almost complete loss of spatial resolution, making it difficult for the transposed convolution operations to recover informative features. Skip connections have been used in FCNs to combat this problem. Skip connections combine features from previous steps in the contracting path with features produced from upsampling the deepest feature

maps. This allows the combination of deep abstract features with spatial information. This has been achieved with element-wise addition [110] as well as concatenation [112].

The final convolutional layer outputs a feature map with the same width and height as the original image, and a number of channels equal to the number of classes. The softmax function is applied to this feature map in order to find the most likely class for each pixel. This gives the classification prediction for each pixel across the whole image and allows the calculation of a suitable loss function.

With an understanding of the structure of convolutional neural networks and their distinctions from traditional neural networks, the following section explores algorithms for adjusting the neuron weights of these networks to identify relevant features and make accurate predictions.

3.3.3 Parameter Optimisation Algorithms

With a loss function providing a measure of the performance of a given set of parameters, the next step is to adjust the parameters to minimise this loss. Finding the best performing parameters for a neural network is an optimisation problem, a search for the minimum of the objective function. As the number of parameters in a neural network is substantial, potentially millions, finding minima in such multi-dimensional functions is challenging.

The simplest approach, randomly searching through the parameter space, is unlikely to yield a good approximation of the minimum with such a complex function. As a result, there are numerous strategies to estimate weight adjustments and search the multi-dimensional space. Backpropagation allows the efficient calculation of the gradient of the loss function with respect to the parameters of the neural network, using the chain rule to compute gradients from the final layer to the first. As a natural

result of this efficiency, approaches to the problem most commonly concentrate on gradient descent algorithms. A candidate set of parameters is randomly initialised, possibly within some constraint such as a Gaussian distribution, and the derivatives of the loss function are calculated to assess the gradient. Adjustments to the parameters can then be made in the direction of the largest negative gradient, iteratively refining the parameters to find a minimum. This section describes and reviews the optimisation algorithms, or optimisers, that define exactly how these gradients are used to search the parameter space and their use in the literature.

Stochastic Gradient Descent

The simplest form of gradient descent, batch gradient descent, calculates the gradient of the loss function with respect to the parameters across all available training examples. With a summed gradient for the training data at the given point in the loss function, the value of all parameters can be updated by a small amount to move the estimate towards a minimum. This process is repeated iteratively until converging on a minimum, according to Equation 3.17.

$$\begin{aligned} v_n &= -\eta \cdot \nabla L(\theta_n) \\ \theta_{n+1} &= \theta_n + v_n \end{aligned} \tag{3.17}$$

Where θ_n is the vector of neural network parameters for the nth iteration, and $\nabla L(\theta_n)$ is the gradient vector with respect to θ_n for the loss function L. The final term η, decides the step size of updates to θ with each iteration. In the context of deep learning, this variable is commonly referred to as the learning rate, and is a hyperparameter that must be defined for training of the neural network.

Batch gradient descent requires calculation of a summed gradient across all training examples to perform each update of the weights and biases of the network. As a result, this method is slow to converge on a minimum, and unmanageable in situations

where the dataset cannot fit within memory. Stochastic gradient descent (SGD), developed from stochastic approximation methods [113, 114], alleviates these shortcomings. Instead of calculating the actual gradient for the whole dataset, SGD estimates the gradient using a random subset of the data and updates parameters more frequently. Unlike batch gradient descent, which will naturally introduce redundancy by computing gradients across similar examples in the dataset, updating parameters with a single example at a time allows SGD to perform more efficiently while keeping the training examples within memory [115].

In exchange for overcoming these limitations, each update to the parameters has higher variance, causing larger oscillations across the loss function surface. This can cause overshooting of a minimum and is therefor slower to converge than batch gradient descent, where each update is guaranteed to move closer to a local minimum in the loss function. With an appropriate learning rate, it is possible that this increased volatility in parameters can allow wider exploration of the loss function and escaping less optimal minima, though it is similarly possible to miss more optimal minima.

A compromise between the discussed approaches involves the use of larger random subsets of the data. This approach is often termed mini-batch gradient descent in the literature or included in the definition of SGD. In this work the term SGD will be used, and a batch size will be included to specify the size of the random subset used for each update of the network parameters. With larger batch sizes the variance of each update is reduced, allowing for more stable convergence but reducing the advantage of overcoming redundancy and quicker calculation. How the trade-off between these advantages can best be used to train a neural network requires experimentation, finding the optimum batch size for training a given network. This introduces batch size as an additional hyperparameter along with learning rate when using SGD in the training process.

Momentum

Momentum is an additional term that can be added to the SGD method, which encourages more consistency in the direction of exploration across the loss function surface [116]. The oscillations that occur in SGD are particularly common along saddle shaped areas of the surface, where gradients are steep along some dimensions, but slight in others. With each update to the parameters, the vector of those changes is calculated from the current gradient, which can cause the estimate to oscillate across either side of a valley, and only gradually follow the slight gradient. With momentum, a proportion of previous update vectors is also included in the current update vector, according to Equation 3.18.

$$\begin{aligned} v_n &= \alpha \cdot v_{n-1} - \eta \cdot \nabla L(\theta_n) \\ \theta_{n+1} &= \theta_n + v_n \end{aligned} \tag{3.18}$$

Where α is an exponential decay factor between 0 and 1, determining the proportion of earlier gradients in the calculation of the new update vector v. The addition of momentum allows the optimisation to accelerate toward a minimum in situations where there is a consistently small gradient, facilitating faster convergence. Additionally, in situations with a large gradient in alternating directions, the momentum term will dampen oscillations to decrease variance in parameter updates. Momentum has been successfully applied to train neural networks in a range of applications, and α typically takes a large value of 0.9 or more.

An extension of the concept of momentum is the Nesterov accelerated gradient [117], also known as Nesterov momentum. The calculation of Nesterov momentum anticipates the contribution of the momentum from previous update vectors, $\alpha \cdot v_{n-1}$, and calculates the gradient at this new position on the loss function surface, rather than the current position. This is demonstrated in Equation 3.19.

$$
\begin{aligned}
v_n &= \alpha \cdot v_{n-1} - \eta \cdot \nabla L(\theta_n - \alpha \cdot v_{n-1}) \\
\theta_{n+1} &= \theta_n + v_n
\end{aligned}
\tag{3.19}
$$

This strategy causes the updates to make large jumps along the direction of momentum and correct based on the new gradient. This increases the rate of convergence to a minima as the gradient component of updates is immediate, placing greater emphasis on newer gradients and not lagging by an iteration. As a result it has been demonstrated that Nesterov momentum has superior performance to classical momentum in many deep learning applications [118, 119].

Learning Rate Adaptation

All of the discussed optimisation strategies so far have required the use of a learning rate η, to control the size of steps taken across the parameter space. This is one of the most impactful hyperparameters in the training process. A learning rate that is too large will cause the optimisation to oscillate around minima and delay convergence, or miss minima entirely and diverge. A learning rate that is too low will converge very slowly, increasing training time, and may become trapped in a sub-optimal minimum.

Learning rate schedules are a strategy in which the learning rate is adjusted during the training process [120, 121], which has been used in medical image deep learning applications to improve training [122]. This can be achieved by scheduling a decay factor into the learning rate, decreasing the learning rate at set intervals of iterations, an exponential decrease gradually as iterations increase, or when the improvements to the optimisation slow beyond a chosen threshold. Known as learning rate annealing, these schedules introduce additional hyperparameters, as the initial learning rate, decay rate and decay intervals must be decided before training.

These learning rate schedules make use of the same learning rate for all parameter

updates across the network. The adaptive gradient algorithm (Adagrad), is an optimisation algorithm developed in order to allow adaptation of the global learning rate η to each parameter θ_i based on previous gradient updates. The key to this is to keep a running total of the sum of squares of all gradients for each parameter up to iteration n. Each parameter is updated like so:

$$\theta_{i,n+1} = \theta_{i,n} - \frac{\eta \cdot \nabla L(\theta_{i,n})}{\sqrt{S_{i,n}}} \tag{3.20}$$

Where $S_{i,n}$ is a matrix containing the sum of squares for all previous gradients $\nabla L(\theta)$ for each θ_i to iteration n. As a result of the continuous summing of gradients, learning rates gradually decrease for all parameters proportionally to the extent of their updates so far. This allows considerable flexibility in the choice of initial learning rate η, naturally including annealing of the learning rate for parameters which are updated more frequently. The downside of this process is that learning rates decrease monotonically and eventually become negligible, preventing further learning. An extension of this technique, Adadelta, was developed to combat this shortcoming.

Adadelta, adapts the Adagrad sum of squares by keeping a weighted sum of previous gradients [123]. This is achieved by summing previous gradients with a decay factor λ, increasing the weight of more recent updates. The previous gradient sum S_n is replaced with D_n:

$$D_n = \lambda D_{n-1} + (1-\lambda)(\nabla L(\theta_n))^2 \tag{3.21}$$

This exponentially decaying window of previous gradients allows Adadelta to be more robust, and avoid learning rates vanishing. Another extension of this genre of optimisation algorithms is the popular adaptive moment estimation (Adam).

Adam, is an intuitive combination of the concepts covered so far. Adam calculates adaptive learning rates on a per parameter basis using both a exponentially decaying

average of past gradients, akin to momentum, and of the squared gradients, similar to Adadelta [124]. Resembling the Adagrad and Adadelta algorithms, Adam parameter update looks like so:

$$\theta_{i,n+1} = \theta_{i,n} - \frac{\eta \cdot \hat{m}_{i,n}}{\sqrt{\hat{v}_{i,n}}} \tag{3.22}$$

Here, m and v encapsulate the decaying average of previous gradients and squared gradients respectively, each with a decay constant β. Due to being initialised as zero vectors, bias correction is implemented to avoid bias toward 0. The bias corrected $\hat{m}$ and $\hat{v}$ are calculated with these equations:

$$\begin{gathered} \hat{m}_{i,n} = \frac{m_{i,n}}{1-\beta_1^n} \\ \hat{v}_{i,n} = \frac{m_{i,n}}{1-\beta_2^n} \\ \text{where:} \\ m_{i,n} = \beta_1 m_{i,n-1} + (1-\beta_1)\nabla L(\theta_{i,n}) \\ v_{i,n} = \beta_2 v_{i,n-1} + (1-\beta_2)(\nabla L(\theta_{i,n}))^2 \end{gathered} \tag{3.23}$$

Which introduces our two decay constants β_1 and β_2. m is akin to momentum and estimates the first moment, the mean, of the gradient. The suggested value for β_1 is 0.9. v estimates the second moment, the variance, of the gradients, and its decay factor β_2 has a suggested value of 0.999. Adam is used extensively in the literature due to its stability in training a range of models and its incorporation of the strengths of both momentum and Adadelta [124].

Parameters such as learning rate and decay constants have a large impact on the performance of optimisation algorithms. These are termed hyperparameters, and large part of the challenge in training deep learning models is in the selection of suitable

values. The next section covers some of the strategies used to select optimal hyperparameters for learning.

3.3.4 Hyperparameter Optimization

In contrast to the parameters of a deep learning system, the weights and biases that are learned during training, hyperparameters are the variables of the system that are chosen before training, affecting how the parameter optimisation is performed. Hyperparameter optimisation, also known as tuning, is the process of finding an optimal set of hyperparameters for best performance of a deep learning network. Hyperparameters include the variables of parameter optimisation, such as learning rate and momentum, but can also include structural variables, such as which activation functions to use. Similar to parameter optimisation, this problem can be treated as an exploration of the hyperparameter space, with each hyperparameter as a dimension, to find a suitable maximum of the performance metric. Unlike parameter optimisation, it is not possible to assess the derivatives of the underlying function, it can only be evaluated with pointwise sampling. Hyperparameter optimisation is the main reason for the use of separate validation and testing datasets. Fitting hyperparameters to the test dataset performance would give an unrepresentative measure of model performance, as it allows overfitting to data specific to the dataset.

There are a number of techniques which can be used to search the hyperparameter space. A typical first step is to choose the hyperparameters manually. The range of reasonable values for a given hyperparameter can be ascertained from the literature, from similar models or the same model used on a different problem. Relying on the judgement of the experimenter, a sensible set of hyperparameters can be chosen and tested on the validation data. Based on the performance metrics, qualitative assessment of results and the rate of decrease and relationship between training and validation loss,

it is possible to make informed estimates of how to adjust hyperparameters, for instance an increase in regularisation to combat overfitting.

General guidelines and best practices are available within the literature to facilitate manual searching, though with a lot of variance in performance with small changes to hyperparameters and many deep learning models being highly specific to a given problem, finding an optimal solution without a structured approach is difficult. The use of a structured approach to hyperparameter tuning is also beneficial to reproducibility, manually selecting hyperparameter combinations is difficult to recreate and test.

Random and Grid Search

Each set of hyperparameters can be trained and an objective measure of performance can be acquired for each. To find the best performing set of hyperparameters, it is possible to exhaustively search the hyperparameter space within the confines of sensible values. This technique is termed grid search. Each hyperparameter is discretized between a minimum and maximum value, for example the dropout rate may be 0.1, 0.2, or 0.3 and the learning rate may be 10^x where x is -2, -3, or -4. The Cartesian product of all hyperparameters defines the number of models that must be trained, 9 in this example. Experience and judgement are still used to define the bounds of a hyperparameter and the discretisation steps. Choice of discretisation acts as the sampling frequency across the hyperparameter space, and therefor limits how closely the proposed maximum matches the theoretical maximum of the space.

It has been demonstrated that grid search is a reliable improvement over manual sequential optimisation in the same time frame, for low dimensional hyperparameter spaces [125]. The main drawback of the grid search approach is that it is computationally expensive, the number of samples increases exponentially with each additional hyperparameter. For use with deep learning applications, where training a single model can take several hours, the exhaustive search of the grid can become time consuming

very rapidly. It is however, a perfectly parallel problem, allowing any and all samples of the space to be performed simultaneously if computational resources allow.

A variation on grid search, which randomly samples the hyperparameter space instead of discretising it, is random search. Within the bounds for each hyperparameter, a random combination are selected and the objective function is sampled at that point. The random search overcomes the limitation of limited sampling frequency present in grid search, as there is a chance to sample any combination of hyperparameters infinitely close to the optimum solution. Each random combination will sample a an individual hyperparameter dimension in a different location. In the case of two dimensions that have weak correlation, grid search would inefficiently sample the same value of one dimension while adjusting the other. Random search can better find the optima of each by sampling more of each dimension. As the number of dimensions increases, the correlation between them decreases, causing random search to be more efficient compared to grid search as the number of hyperparameters increases [125].

Unlike grid search, this search is not finite, random search can continue to sample the hyperparameter space indefinitely. Random search is similarly perfectly parallel, allowing simultaneous sampling, with the effective sampling resolution increasing the more samples are taken. As a result, the sampling frequency is limited by the availability of time and computational resources. These techniques by their nature involve inefficiency, as previous samples are not used to inform the selection of future hyperparameters. This opens the way for techniques which decide where in the hyperparameter space is likely to contribute the most information about the underlying function.

Automated Hyperparameter Optimisation

The fundamental problem of hyperparameter optimisation is that the underlying objective function is not easily differentiable, can only be assessed at individual points and

is expensive to evaluate. A fruitful solution has been the development of a probabilistic model which maps hyperparameters to the expected score on the objective function [126, 127]. This approach is termed Bayesian optimisation.

The overall strategy of Bayesian optimisation is to build a probabilistic model as an approximation of the objective function. The hyperparameters which are most likely to perform best on the surrogate can then be identified and applied to the true objective function. The model is then updated to include this new observation and the process is repeated to improve hyperparameter predictions until further improvements are sufficiently minor, or time constrains. This approach is an improvement over random and grid search strategies as past results are used to inform future evaluation, accelerating convergence [128, 127].

The key components of a Bayesian optimisation based search strategies, known as sequential model-based optimisation methods, are the probabilistic model, to approximate the objective function, and the acquisition function, which compromises between exploration and exploitation to suggest the next hyperparameter choice. Variations on these methods comprise selecting different models and acquisition functions. An initial and popular probabilistic model used in the literature is based on Gaussian processes (GPs) [129].

Given a set of points $[x_1, x_2, ..., x_n] \in \mathbb{R}^d$ (where $\mathbb{R}^d$ is the d-dimensional hyperparameter space) and an objective evaluations of how these hyperparameters perform $[f(x_1), f(x_2), ..., f(x_n)]$, GPs assume that these observations are drawn from a multivariate normal distribution. The prior GP is defined by a mean function, $\mu(x)$, and covariance function, $\sigma(x)$. The covariance function assumes that observations close together are more closely correlated, and so have less uncertainty, and that observations are noisy samples from the true function [130]. As the number of observations increases, the uncertainty around the underlying function decreases, and the surrogate becomes more accurate.

The acquisition function aims to choose a region of the surrogate function which is likely to yield improvement. One of the earliest methods for achieving this is the aptly named probability of improvement [131]. Given a GP prior (with Gaussian cumulative distribution function Φ) and a current best performing observation ($f(\mathbf{x}^+)$), the aim is to choose a new candidate $\mathbf{x}$ by maximising $PI(\mathbf{x})$:

$$PI(\mathbf{x}) = p(f(\mathbf{x}) > f(\mathbf{x}^+) + \varepsilon) = \Phi\left(\frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \varepsilon}{\sigma(\mathbf{x})}\right) \tag{3.24}$$

Where ε is an additional term that encourages exploration of the surrogate function. With $\varepsilon = 0$, maximising $PI(\mathbf{x})$ will favour points that have a higher probability of negligible gains, over those with a lower probability of larger improvement. The term encourages some exploration by forcing the acquisition function to choose the new point in the hyperparameter space that has a chance to improve on $f(\mathbf{x}^+)$ by at least ε. Choice of ε, and indeed varying its value over iterations of Bayesian optimisation, impact how much exploration and exploitation the acquisition function attempts, a key challenge in any optimisation problem [131, 132].

Taking the probability of improvement concept further, it is possible to choose a point based on both the certainty of improvement as well as the magnitude of that potential improvement [126, 133]. This is a concept known as expected improvement. This acquisition function is defined as:

$$EI(\mathbf{x}) = (\mu(\mathbf{x}) - f(\mathbf{x}^+))\,\Phi\left(\frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})}\right) + \sigma(\mathbf{x})\,\phi\left(\frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})}\right) \tag{3.25}$$

Where ϕ is the Gaussian probability density function. Intuitively the addition of expected improvement is a term which encourages exploration. Candidate points are

suggested which are predicted by the probabilistic model to have high mean, $\mu(\mathbf{x}$, exploitation of potentially promising points. Alternatively, points with high variance, $\sigma(\mathbf{x}$, can be evaluated to explore regions with low certainty. Expected improvement performs well and has the advantage of not requiring its own tuning parameters, avoiding the problem of passing the hyperparametised buck.

These strategies can be used to intelligently select promising combinations of hyperparameters to optimise deep neural network performance [127, 134]. With an understanding of strategies to learn the parameters and tune the hyperparameters of deep neural networks, the next section addresses one of the most successful and popular neural networks in the domain of image segmentation, the U-Net.

3.4 U-Net

U-Net is a fully convolutional network (FCN) which has been very popular in image segmentation tasks. Networks based on the U-Net architecture have been used in a range of applications, including segmentation of structures and surfaces in satellite and aerial photography [135, 136, 137, 138], road marking detection for autonomous vehicles [139], even source separation in audio data [140]. These are an impressive leap from the biomedical application for which the original network was designed. Beyond cell segmentation in microscopy images [112, 141], U-Net has been used in a range of medical imaging applications, including: retinal imaging [142, 143]; organ and lesion segmentation [144, 145, 146, 147]; vessel segmentation [148], and tumour segmentation [149].

The U-Net architecture developed by Ronneberger et al. was applied to histological electron microscopy images to segment cell boundaries. The structure of the model is shown in Figure 3.2. It is closely structured on an FCN, with an encoding and decoding pathway. The symmetry of this pathways gives the network its name. Repeated

convolution and pooling operations are used in the decoding path to identify increasingly abstract patterns and gain insight into the image data. Transposed convolution is combined with further large channel-depth convolutions in the decoding path to recover spatial information to localise class predictions accurately. Skip connections are employed to combine features from the encoding pathway into the convolutions of the decoding pathway, allowing uncompressed propagation of spatial information across the network, these connections are implemented using concatenation across channels of the feature maps. The ReLU activation function is used throughout the network to combat the vanishing gradient problem.

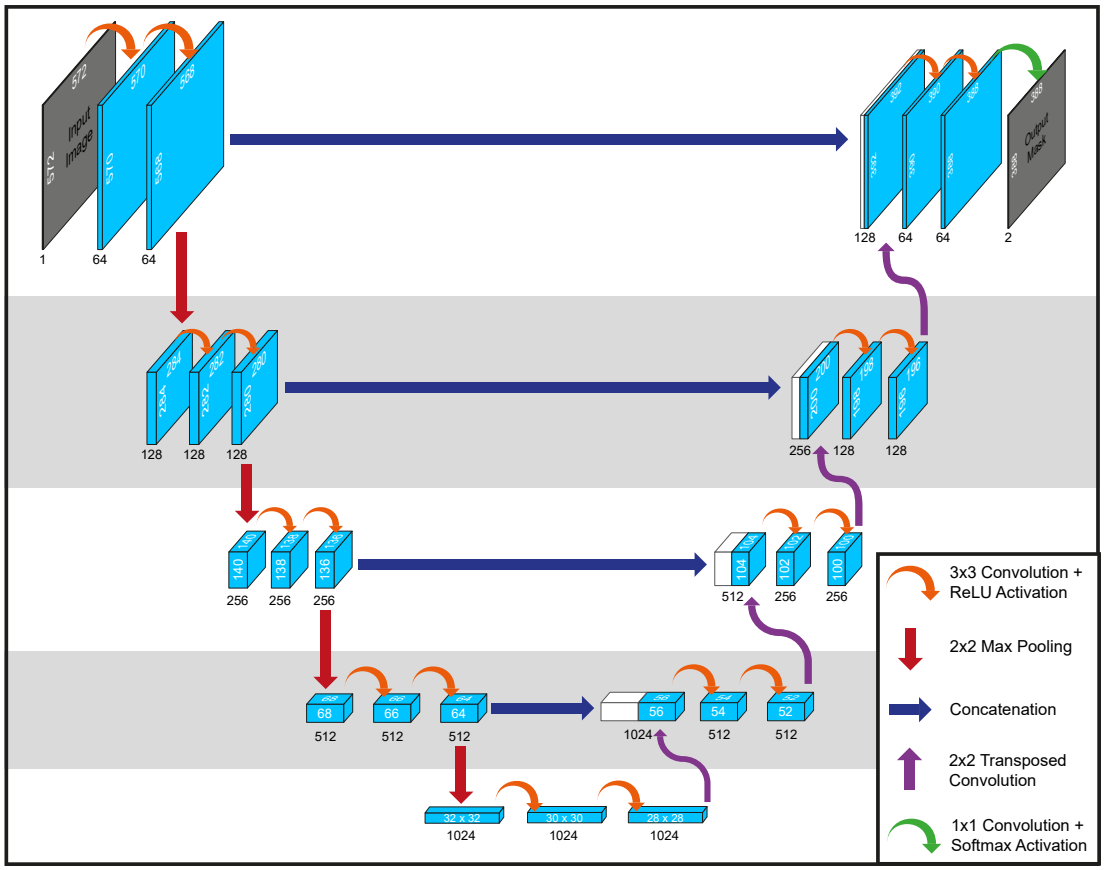

Figure 3.2: The architecture of the original U-Net. Each blue box represents a multi-channel feature map, with the feature depth below the box. Each box has the x and y dimensions of the feature map. Each white box represents the concatenated features from the decoder pathway. Adapted from Ronneberger et al. [112].

U-Net was designed for its application to large electron microscopy images. These images are very large, and are problematic to fit in memory. A tiling strategy was used to process manageable subsections of the image at any one time. Mirroring at the image edges is used to pad the shrinking field of view with sensible image content. The U-Net makes use of image deformations to effectively increase the number of training samples available. Elastic deformations, as well as flipping and rotation are implemented to create sensible variations in the data. This augmentation of the data compensates for one of the main challenges in biomedical image segmentation, the availability of training data. The requirement for high quality expert annotated data can be prohibitively expensive to acquire, use of large amounts of realistic augmentation enables fast and accurate prediction of segmentation maps with a small training set of 30 images [112].

The final challenge of the application was in the segmentation of objects with the same class, with touching borders. Ronneberger et al.[112] made use of a large weighting term in the loss function for background pixels which separate cell borders for touching cells, heavily penalising misclassification. Outside of this weighting, the loss function used was categorical cross-entropy, optimised using stochastic gradient descent with a momentum of 0.99. This loss function is covered in more detail in the next section, followed by exploration of more recent implementations and variations of the U-Net model.

Categorical Cross-Entropy

Categorical cross-entropy is a commonly used loss function in classification and segmentation tasks. Entropy, a concept introduced in information theory, deals with the idea of how much information events provide given their probability of occurring [150]. Cross-entropy can be described as the amount of informational difference

between two probability distributions, the true distribution of class labels, and the predicted distribution from the neural network.

While dealing with predictions for the probability of an example belonging to a class, the cross-entropy is equivalent to the log loss. For a binary classification task, where the class label y is either 0 or 1, and p is the probability predicted by the neural network that an example is class label 1, the binary cross-entropy is given by:

$$CE = -(y\log(p) + (1-y)\log(1-p)) \tag{3.26}$$

This means that if the true label is 1, then the loss is the log of p. As the probability is between 0 and 1, $\log(p)$ will be negative, so the negative log is taken. In the case of a multiclass classification, the cross-entropy for each example is given by a sum of the loss for each class label. For a classification with C classes, a 1 or 0 indicator y_c if the example is of class c and the predicted probability p_c that the example is of class c, the cross-entropy is given by:

$$CE = -\sum_{c=1}^{C} y_c \log(p_c) \tag{3.27}$$

As y is binary, only the predicted probability of the correct class influences the function. Additionally, due to the shape of the logarithmic curve, the function harshly punishes low predicted probability for the correct class. For use as a loss function, the mean value of the cross-entropy can be taken across the number of examples in a batch, or in the case of segmentation the cross-entropy of each pixel in a batch. Backpropagation can then be used to propagate the gradients of this function back through the layers of the network find optimal changes to parameters to reduce the cross-entropy.

To avoid the risk of the model overfitting to the most prominent class (e.g. a mask dominated by negative pixels with only some positive pixels), a weight term can be

introduced to the equation. This weight term adjusts the contribution of each class label to the overall loss function, often proportional to the inverse of the class frequency in the training data.

Clarification of Terms

Throughout the diverse literature on the implementation of U-Net, there is some variation in the terms used to describe the components of the network. For the avoidance of uncertainty, the terms used in this work are clarified here. For example the term *step* is often used as a synonym for iteration when discussing a forward and backward pass during training of a neural network. Additionally, it is common to refer to steps in the U-Net architecture, meaning the sections of the architecture which share the same feature depth, are connected across the step by a skip connection and connect to other steps via pooling and upsampling. In this work, the term iteration is used over step when describing the training process. The term level is used to describe the sections of the U-Net architecture e.g. the original U-Net has 5 levels, with the input and output on the first level, and no skip connection in the fifth.

Each level of the U-Net, with the exception of the last, is separated into two blocks, an encoding and a decoding block. The term block is used to describe the section of the U-Net with repeated convolution and activation operations and a pooling operation. Another synonym for level used in the literature is layer. This is unfortunately used as the term to describe the structure of any neural network, e.g. hidden layers, and to describe convolutional operations in a CNN. In this work, layer will be used to describe a convolution operation, as well as pooling and activation operations. For example, in the original U-Net, each block in the encoding pathway consists of a 3x3 convolution layer, a ReLU activation layer, another 3x3 convolution, another ReLU activation layer, a dropout layer, and a 2x2 max pooling layer.

3.4.1 Architectural Variations

With the highly varied applications of U-Net for image segmentation tasks, has come a considerable number of alterations to the overall structure of the network. As an example, an early and conceptually simple modification was the use of 3D input images and convolutions, to enable volumetric segmentation [151, 152]. These architectural variations still make use of the structure of the U-Net, while adding additional modules or replacing convolutional operations in order to improve performance.

A groundbreaking feature of the U-Net architecture is the use of skip connections to enable more optimal acquisition and implementation of spatial features between the encoder and decoder components of the network. Skip connections have also been used more extensively in network architectures to combat the vanishing gradient problem in deep networks. These additional connections, termed recurrent connections, are made within each convolutional block in ResNet architectures [153]. As shown in Figure 3.3, the output of each convolutional block is summed with an identity mapping of the input, allowing backpropagation of the gradient through earlier layers of the network without vanishing. This technique improves the training of deep networks, with demonstrations of increasing performance for networks with over 100 layers.

These residual blocks were improved upon with the introduction of pre-activated residual blocks [154]. This modification involved moving the activation function, ReLU, before convolutional layers within blocks so that a direct propagation of information can reach any block in the network directly. This enabled even deeper networks to train, some with over 1000 layers, apparently correcting paradoxical increases in error seen in these deeper networks. Residual blocks have been successfully implemented within the U-Net architecture to improve segmentation of retinal vessels [155] and carotid artery calcifications [156].

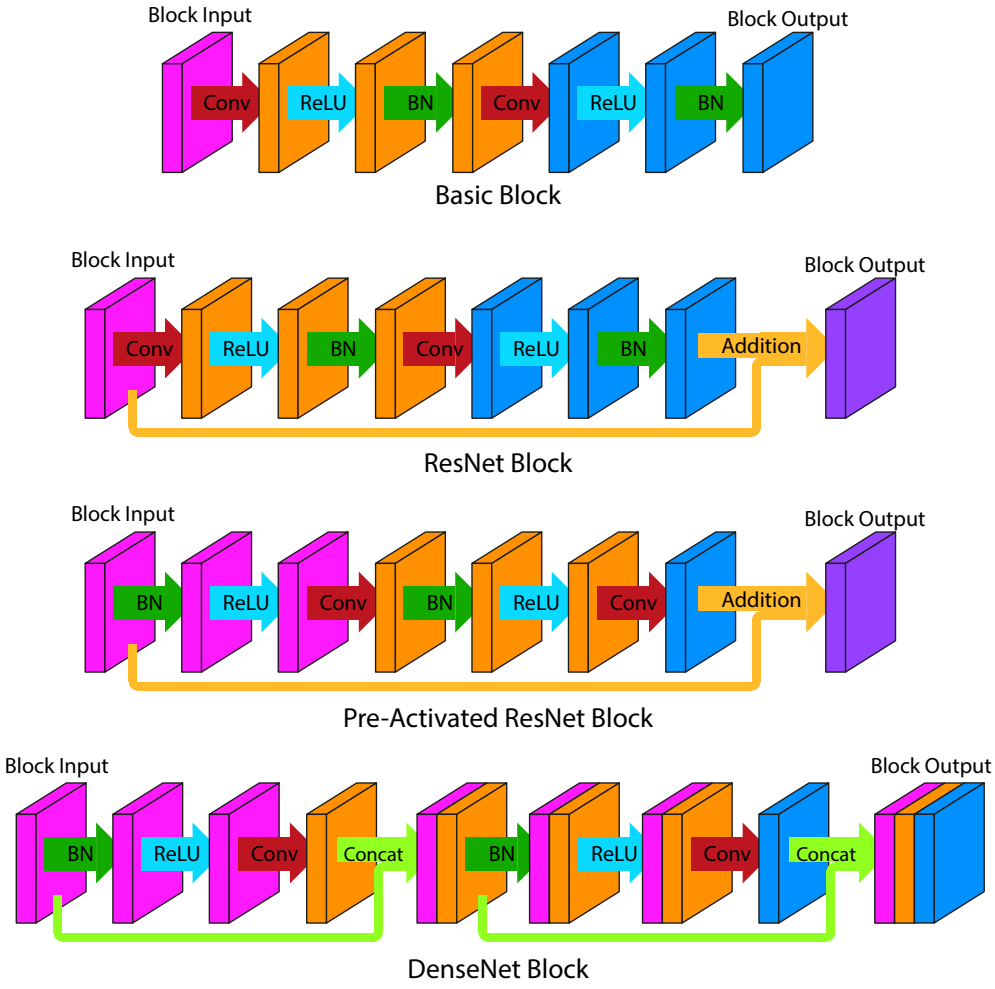

Figure 3.3: Variations on the arrangement of layers within a convolutional block. Each block consists of two convolution operations, with the order of operations varied and skip connections added. BN represents batch normalisation operations.

Taking the concept of unimpeded propagation of information between layers further, dense connections have been implemented within the U-Net architecture. An example of a densely connected block is shown in Figure 3.3, featuring concatenation of all feature maps within the block before all convolution operations. DenseNet, a CNN featuring these dense blocks, has been used to more efficiently train deep networks with hundreds of layers [157]. When used in this fashion, the input to a layer is the feature map of all preceding layers, and its output is used in the input of all subsequent layers. Unlike ResNet, concatenation is used in place of addition. This arrangement continues to combat the vanishing gradient problem, and encourages the reuse of features to reduce the number of trainable parameters required to learn.

With the features of any layer available to any other, features can be reused instead of redundant relearning of features in different sections of the network. It is theorised that the success of the DenseNet lies in an implicit deep supervision enabled by the dense connections. The architecture allows the loss function to supervise the weights throughout the network to encourage learning of discriminative features. These benefits are seen when implementing dense blocks within U-Net [157]. Allowing improved performance in a range of microscopy [145, 158, 159] and medical applications such as lung nodule segmentation, colon polyp segmentation, liver volume measurement, and multi-modal stroke lesion segmentation [145, 160].

3.5 Previous Work

Having covered the clinical and technical foundations of automating AAC quantification in this and the previous chapter, this section covers previous work in the literature aimed at this goal. Little work has been done in this area directly automating segmentation of abdominal aortic calcification in dual-energy x-ray absorptiometry images,

motivating this work. However, there is considerable literature automating the identification of calcific load on the vascular system in other areas, and with other modalities.

Computed Tomography

Computed tomography (CT) imaging allows high resolution visualisation of 3D structures. While the high radiation exposure involved in the modality does not lend itself to use as a screening tool for AAC, it is possible to glean cardiovascular information incidentally during imaging for other indications. Thoracic CT imaging is commonly implemented for screening of lung cancer, or studying chronic obstructive pulmonary disease, in heavy smokers. During these surveys, there is opportunity for imaging of coronary artery calcification and calcification of the thoracic aorta, both indicative of increased cardiovascular risk.

Coronary artery calcification (CAC) can be quantified using the Agatston score (discussed in Section 2.2), and there is a wealth of work automating this process. Early work concentrated on identification of a region of interest containing the heart, with the use of thresholding to select candidate regions, which could be further refined by excluding continuous regions which were either too small or too large [161, 162, 163]. Classification was achieved using spatial and texture features [161], and additional 3D local features inspired by Haar-like features [162] with a k-nearest neighbour (k-NN) classifier. These techniques were extended with additional spatial features, using coordinate systems localised around the heart [164] and segmentation of coronary arteries [165, 166, 167], allowing the development of a multi-atlas *a priori* probability map of calcification [168].

These processes were also applied to allow segmentation of the thoracic aorta, thresholding and subsequent classification of calcified regions, and a total volume of calcific load within the thoracic aorta was reported [169]. Additional methods were later developed specific to the thoracic aorta, with fully automated landmark detection

and shape model fitting [170]. With this method, an approximate circular shape is localised on each slice to form a smooth 3D shape, then the boundary of this shape is grown using a energy function using smoothness and edge detection subject to a regularisation restraint. For segmentation of calcification, a similar simple threshold is applied to find high intensity regions. A region growing algorithm is used on these areas to identify and eliminate those which represent vertebrae or the trachea.

In recent years, focus has concentrated on deep learning approaches to segmentation of vascular calcification in CT images. Agatston scoring of CAC has been automated with CNNs, using a patch based classifier [171]. This method used registration of a cardiac atlas to create a region of interest, sampling positive and negative training patches from each slice according to manual annotation of classes. A simple CNN with 7 convolutional layers was then trained to classify each patch as belonging to the calcification or background class. This method demonstrated high sensitivity and specificity, and calculated Agatston scores showed excellent agreement with manual scores. The use of CNNs was later extended to make use of 3D convolution operations, where a large dataset of CT images was used to train the network to directly output a Agatston score [172]. This approach demonstrated good agreement between predictions and manual scoring, though this was compared in a scaled down low resolution input.

Lessmann et al. made use of CNNs to classify calcification from multiple locations throughout low-dose chest CT images [173]. Patch based sampling was used without the need for prior segmentation, sampling a 2D image at each orthogonal plane and feeding them into a CNN classifier. Patches were classified based on the structure each voxel belonged to, allowing detection of calcification in the coronary arteries, the aorta or each of the heart valves. A second CNN was then used to classify candidate voxels as calcification, totalling 13 convolutional layers for each of the three orthogonal images of each patch. This method achieved moderate agreement with manual annotation

of calcification, and could reliably distinguish calcifications in each region.

The most recent work in automating aortic calcification segmentation have made use of the instance segmentation network Mask R-CNN [174, 175]. This work provides an end-to-end method to detect, classify and segment plaques, using a 50 layer pre-trained model which is fine-tuned using manual annotations of calcification. Further refinement of false-positive regions was performed by clustering pseudo-calcifications in an unsupervised manner using the features of the final convolutional layer. This approach included automatic segmentation of the end plates of the vertebrae and allowed quantification of the calcific load on the abdominal aorta. A sensitivity of 85.0% with a mean of 10 false positives per patient throughout the entire aorta was achieved. This is a relatively high rate of false positives for a potential screening technique, given the clinical risk difference between none and mild AAC. Further work on calcification specifically in the abdominal aorta has been done in lateral radiographs.

Lateral Spine Imaging

With a lateral view of the spine and aorta, precise quantification of AAC is no longer possible. As the only visible calcifications are those parallel to the direction of the x-ray path, only plaques on the anterior and posterior walls of the aorta can be measured. This change in perspective requires different machine learning approaches to automate quantification.

As the aorta is not visible on lateral radiographs unless calcified, work has concentrated on approximating the location of the aorta and probable regions for calcification using shape modelling of the vertebrae. Lauze and de Bruijne use such a technique to produce a spatially varying prior to refine pixel classifications [176, 177]. The probability of each pixel belonging to the calcification class was estimated with Gaussian derivative filters to produce features for a k-NN classifier. The prior aorta shape model is combined with these probabilities to iteratively adjust the shape to find the most

probable location with the foundation that calcification can only be found in the aortic wall. This technique achieved an IoU measure of 0.4214, demonstrating good overlap with the ground truth annotations, though this was done on candidate regions identified by the shape model, not the overlap score of automatically segmented whole images. All of the images in this study contained at least some calcification, and true evaluation of its segmentation performance and use of semi-quantitative scores are left for future work.

Automated scoring of AAC has been demonstrated by Petersen et al. [178] in lateral radiographs. Using a Bayesian framework, the method is able to calculate an AAC-24 score completely automatically. A likelihood function uses Bayesian inference and prior information in the form of shape predictions for the vertebrae and aorta. Appearance features from 10,000 patches were sampled to train a random forest with 200 decision trees, which along with spatial priors, estimated the distribution of calcifications within the aorta. Using 5-fold validation on 800 images, the agreement between the automated method and manual annotations by radiologists was assessed. This comparison achieved a IoU measure of 0.28. Agreement between automated and manual AAC-24 scoring was not high, owing to high image noise and errors in vertebral level, achieving a correlation coefficient $r = 0.7$. The automated analysis also goes a further stage and calculates a CVD risk score by comparing prior and follow-up image sets. Automatically segmenting patients into high and low risk (AAC-24 score above or below 3.5), the correlation demonstrated encouraging improvements to risk prediction in CVD, with a mean hazard ratio for CVD mortality within 5 years of 2.4.

As has been covered in Section 2.2, the low radiation dose of DXA imaging gives it an advantage as a screening tool, with the downside of increased noise and lower resolution. A paper by Elmasri et al. [179] assessed automated quantification of AAC in VFA images. Single energy VFA images were used to manually train an AAM to identify the spine and aorta. After segmenting the aorta, multi-level thresholding

was applied to leave likely candidate calcifications. Image histogram features were extracted and used to develop k-NN and support vector machine classifiers. AAC-24 scores were not directly assessed, as the classes for training and output were mild, moderate and severe. The ground truth for these classes was based on the AAC-24 scores provided by manual annotation.

The method uses a relatively small image set of 73 VFA images with mild to moderate AAC, of these 20 were used for training. With the aorta itself being invisible on VFA, the model was trained using images with clear calcification, likely severe AAC. Heavy calcification in the aorta is likely to have a less typical location and shape owing to the increased rigidity. Combined with the small sample, the training data may not have contained a representative sample of anatomical variation. Despite these limitations, the results of the study show an impressive agreement between automatic and manual classification, particularly with a k-NN classifier. The accuracy of the automated method compared to manual ranged from 83.3% and 95.2%. As is to be expected based on the training data, the highest accuracy was seen in severe AAC. This study is an encouraging first step into automated AAC measurement using VFA.

To further the work on automated measurement of AAC severity, a different approach is required. AAMs cannot accurately segment the aorta when there is mild or no calcification, the models depend on ample texture information. This highlights the importance of including images containing no calcification to evaluate false positives and misclassification of risk. With an understanding of the performance of previous techniques and potential machine learning approaches used for the problem of segmentation, the remainder of this work presents relevant experiments to achieve automated AAC scoring. There are several machine learning techniques for image segmentation which have not been evaluated on this problem in VFA images. The following chapters explore some of these techniques, namely random forests and fully convolutional networks, as well as a shape modelling approach to image augmentation and registration.

Chapter 4

Automated Localisation and Scoring

This chapter discusses the data and methodologies that are common to all the approaches to semantic segmentation of abdominal aortic calcification used in this work. Firstly, the image and annotation data used throughout this and future chapters is described. The chapter then explores the methods for selecting suitable regions of interest containing the abdominal aorta, and the automated methods for converting label masks of AAC into the semi-quantitative scores used in the literature. The results of these methods are presented and discussed.

4.1 Data

The data used throughout this work were from two main sources. The first was the Medical Research Council (MRC) National Survey of Health and Development 1946 (NSHD) [180]. The second is the Calcium Intake Fracture Outcome Study [181].

The MRC NSHD is a large cohort of 5362 men and women born in England, Scotland and Wales in one week in March 1946. The survey has collected a range of information on individuals from their birth to the current day, including cardiovascular, respiratory and reproductive health as well as socioeconomic factors. In a recent

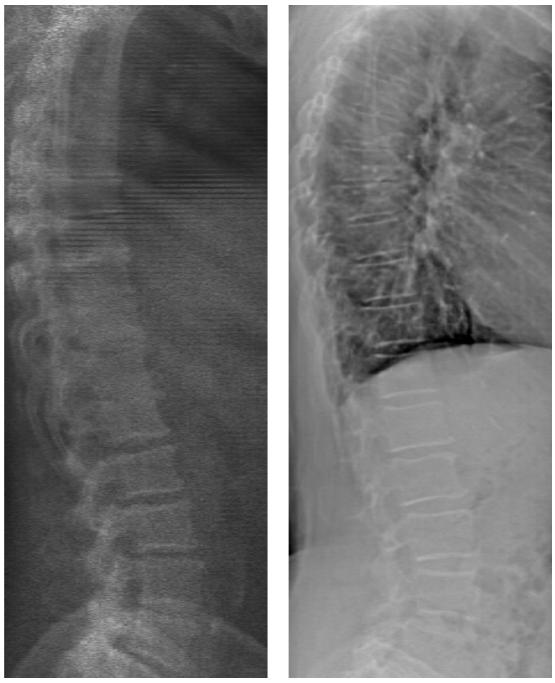

Figure 4.1: Examples of DXA images from the NSHD and CAIFOS datasets. The left and right images show examples from the NSHD and CAIFOS data sets respectively, with contrasting image quality.

data collection, completed in 2011 [180], imaging studies of the skeletal system were performed on a subset of participants, including DXA VFA.

1601 participants underwent DXA VFA imaging on a Hologic QDR 4500 Discovery in single-energy mode, while cohort members were in the 60-65 year-old range. The images provided had a spatial resolution between 378x1127 and 399x1160 and an 8-bit bit depth. After eliminating images that were of unusable quality, and those which did not fully image the abdominal area between the 1st and 4th lumbar vertebrae, 210 images containing AAC were available with expert annotation using the AAC-24 score. Figure 4.1 shows an example of a DXA image from this dataset alongside an image from the Calcium Intake Fracture Outcome Study.

The Calcium Intake Fracture Outcome Study (CAIFOS) was a 5-year prospective, randomised, controlled trial to prevent osteoporotic fractures using oral calcium supplements. 5,586 participants were recruited from the Western Australian general

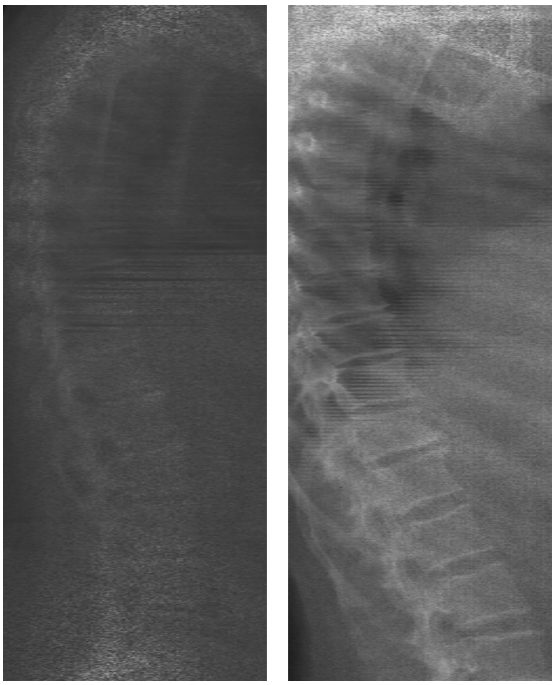

Figure 4.2: Examples of images that were not included from the datasets. The left and right images demonstrate data that was rejected due to poor quality and incomplete imaging of the target area respectively.

population of women aged over 70 years. Participants were given a daily calcium supplement or placebo and followed up for 5 years. Baseline or Year 1 DXA VFA images were taken for a subset of study participants. A total of 1083 VFA images were available from this cohort, imaged using a Hologic QDR 4500A in single-energy mode. Images have a spatial resolution of 287x800 and a 12-bit depth, in DICOM format. Expert annotation of AAC-24 score was available for 747 of these images, once unsuitable images had been eliminated. Figure 4.2 provides examples of rejected images from the studies.

In total, the data available consisted of 721 DXA VFA images with evidence of AAC, scored by expert readers and given an AAC-24 score. The quality of images varies considerably between the two image sources, as demonstrated in 4.1. Though imaging occurred on similar models of scanner, there will be considerable variation in time available for scanning, the skill of the operators and the degree of consideration

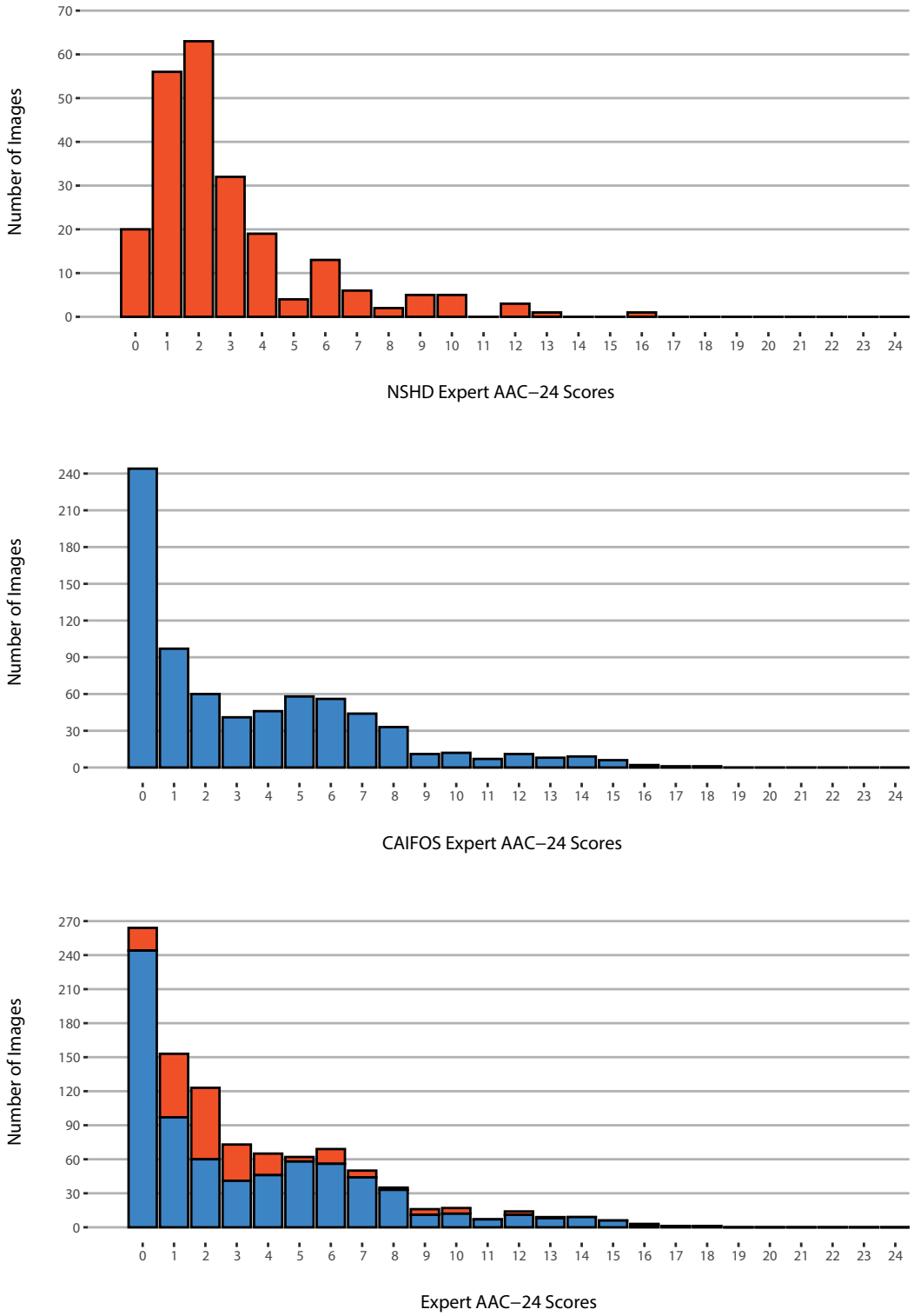

Figure 4.3: Distribution of AAC-24 scores across the data set as annotated by domain experts. The top chart shows the distribution for the NSHD data, with the CAIFOS data in the middle, and the combined distribution on the bottom chart.

for AAC inclusion.

Additionally, 256 images were read by expert annotators as containing no calcification. These images were also included in the data as it was essential to be able to accurately score VFA images without AAC. Figure 4.3 shows the distribution of AAC-24 scores in the NSHD and CAIFOS data and across the dataset as a whole. Noticeably, there was a heavy bias towards lower AAC-24 scores, with few examples from the extremely high scores.

This distribution of AAC scores made a classification task difficult, with under-representation and complete absence of many classes. This is typical of the population examined for osteoporosis screening and would be even more pronounced if screening were to occur across a broader age range. As a result, the general approach of previous work [178, 179] and in this work has been the use of segmentation strategies, with additional processing of these segmentations to produce classes, rather than direct classification of images.

4.1.1 CT Sagittal Projection Images

An additional image dataset was included to assess the spatial relationship between the aorta and spine. The dataset was composed of CT images from a study automating vertebrae localisation for osteoporotic fracture detection [182]. The 3D CT images underwent sagittal projection to construct 2D images through the midline of the vertebrae. A slice thickness was approximated by choosing the number of sagittal slice rasters to sum to produce the 2D sagittal projection images.

For the purposes of assessing the relationship between the lumbar vertebrae and abdominal aorta, images with an effective slice thickness of 10mm were calculated, summing the sagittal slice rasters from 5mm either side of the plane through the midline of the vertebrae. This region should contain the abdominal aorta from L1-4. Although

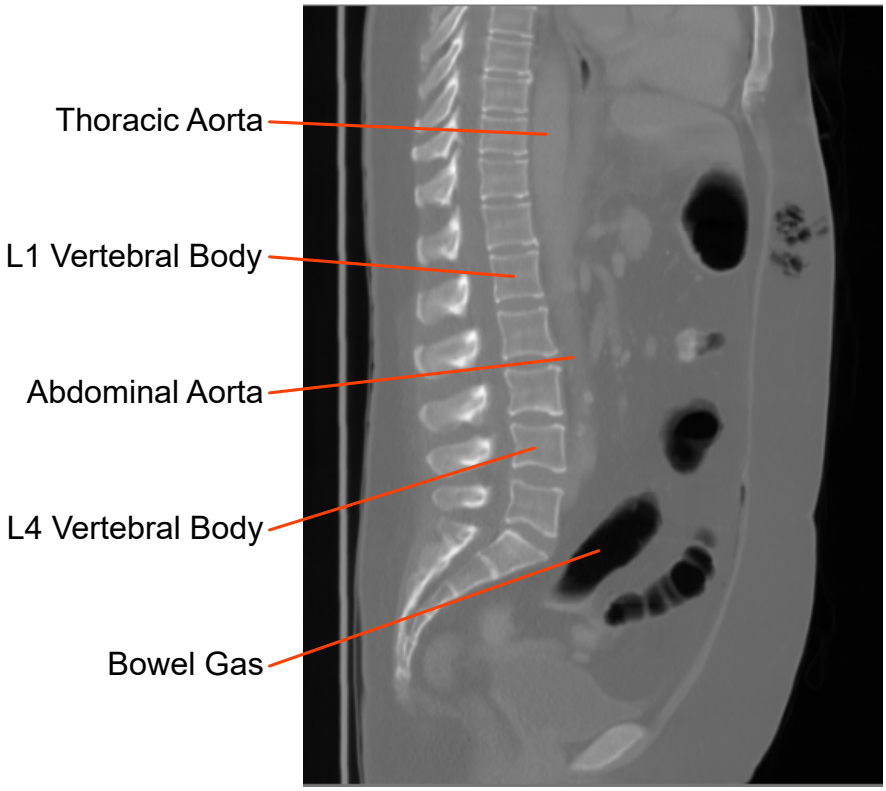

Figure 4.4: Example of a CT sagittal projection, showing the close relationship between the abdominal aorta and the lumbar vertebrae.

there are differences in the acquisition of DXA and CT images, magnification of structures closer to the source for example, the spatial relationship between vertebrae and aorta should be relatively consistent between the two imaging modalities. Figure 4.4 shows an example of the sagittal projections with the relative positions of the aorta and spine.

Images with obvious aortic pathology such as abdominal aortic aneurysms or stents were excluded. While identification of these pathologies may be an important research area for future automation, the primary goal was to model the spatial relationship between the vertebrae and aorta in normal and calcified aortas. Due to the nature of the dataset, many images included vertebral pathology, such as fractures. This was useful to include in the model, as many of the DXA images also include pathology. The randomly selected CT images contained a range of levels of calcification.

4.1.2 AAC Annotation

While the available datasets contained expert radiologist scores for AAC on an image level, the development and evaluation of segmentation algorithms require pixel-level annotation of images. The clinical gold standard for assessment of AAC is through the use of CT imaging. As paired CT and DXA data was not available, identification of AAC on DXA is an approximation of the underlying ground truth. Annotation of the DXA VFA images provides the standard against which segmentation performance will be judged, any reference to ground truth segmentation performance throughout this work, is intended to refer to this pseudo-ground truth.

Images were read by the author, scored according to the AAC-24 scale described in Section 2.2, and used to annotate pixel-level segmentation masks. The author is not a domain expert in DXA VFA or AAC, but does have formal training in interpretation of radiographs, and clinical and academic experience of osteoporosis and CVD. The author was trained by a consultant radiologist, a specialist in DXA VFA imaging and bone density assessment [183, 184, 185], to identify and score AAC in these images in line with clinical practice. Additionally, feedback and instruction was provided by domain expert radiologists researching the cardiovascular risk implications of AAC [186].

In total 350 images were annotated for AAC. All 210 available NSHD images with expert identified AAC were included (with an additional 20 images with no AAC) and 120 randomly selected images from the CAIFOS dataset. This allows assessment of segmentation performance generalisability with changes in image quality and imaging hardware, and training to improve performance across the datasets. These images were first read and given an AAC-24 score while blind to the expert scores, to give a measure of the inter-rater reliability. Scoring was performed twice by the reader with a delay of at least two months between any image being repeated, allowing the calculation of

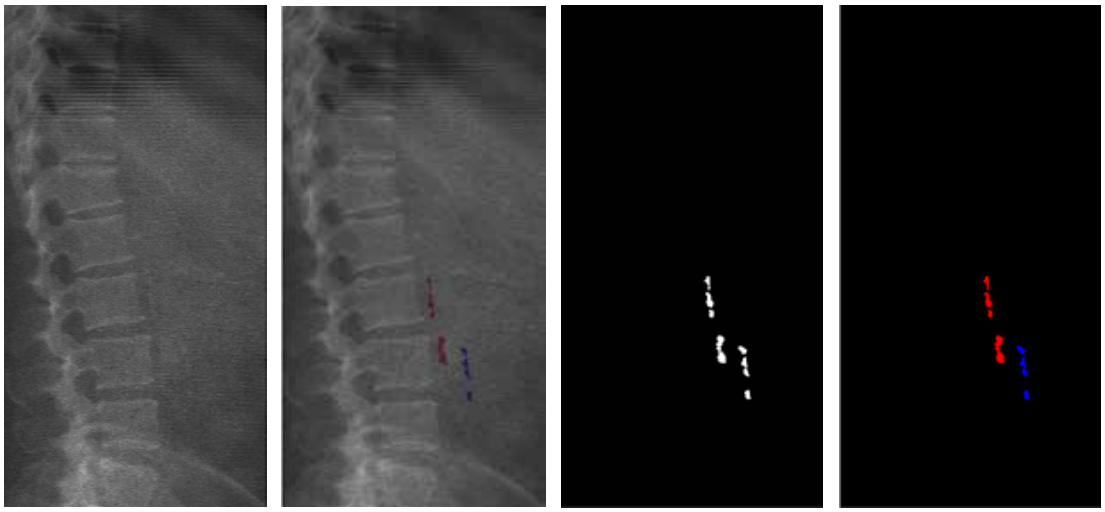

Figure 4.5: An example of AAC annotation to produce segmentation masks. The left image is the original DXA image. The centre left image is the overlay of annotated calcification pixels. The right two images are the resulting single and two class annotated masks.

intra-rater reliability. Separately to AAC-24 scoring, pixel annotation of AAC within images was performed. Figure 4.5 shows an example of an annotated image.

This AAC annotation produced pixel mask images for each VFA image with the same dimensions. These masks had a value of zero everywhere except where there was calcification present, where the value was one. Calcifications were also annotated as either anterior or posterior wall calcifications, creating an additional 2-channel mask for each image, with a channel for each of the classes. Figure 4.5 also shows these mask annotations of a VFA image, with separate channels for anterior and posterior wall calcifications.

4.2 Methods

The first challenge in automating AAC assessment was reliably locating the aorta within an image. Given the variability in patient positioning within images and resolution, consistent identification of a region of interest can focus segmentation algorithms

on the informative parts of the image. With the aim of calculating calcification scores for the abdominal aorta, segmenting the area of the image likely to contain the aorta was an important first step. Once a region of interest had been identified, pixel-level annotations of calcification were used to calculate AAC-24 scores for images.

4.2.1 ROI Prediction

As the aorta is invisible in DXA images unless calcified, location of the aorta on DXA VFA images could only be performed when calcification is severe. Due to differences in the shape and rigidity of the calcified aorta, this could introduce bias, reducing sensitivity to minor calcifications. Instead, the location of the aorta was estimated from consistently visible bony landmarks. As the abdominal aorta is reliably positioned anterior to the lumbar vertebrae, its location can be approximated from vertebral shape information.

The CT sagittal projection image dataset, described in Section 4.1.1, was used to build a statistical shape model (SSM) which encodes the spatial relationship between the abdominal aorta and lumbar vertebrae. Once a model of the relationship between vertebral annotations and aortic annotations was trained, it was possible to predict the position of the aorta on a DXA image using only vertebral annotations.

100 of the projection images were annotated to build an SSM. Each of the 100 images was annotated with 30 2D points. Figure 4.6 demonstrates the annotations on a sagittal projection image. Each of the vertebrae L1-4 were annotated with four points at the corners of the vertebral bodies. The inferior corners of T12 and superior corners of L5 were also annotated, allowing the model to include shape information on these intervertebral spaces. The anterior and posterior walls of the aorta were annotated at the level of each intervertebral space from T12-L5, giving 10 aortic points.

As the relative positions of the landmark annotation points were used, and not the

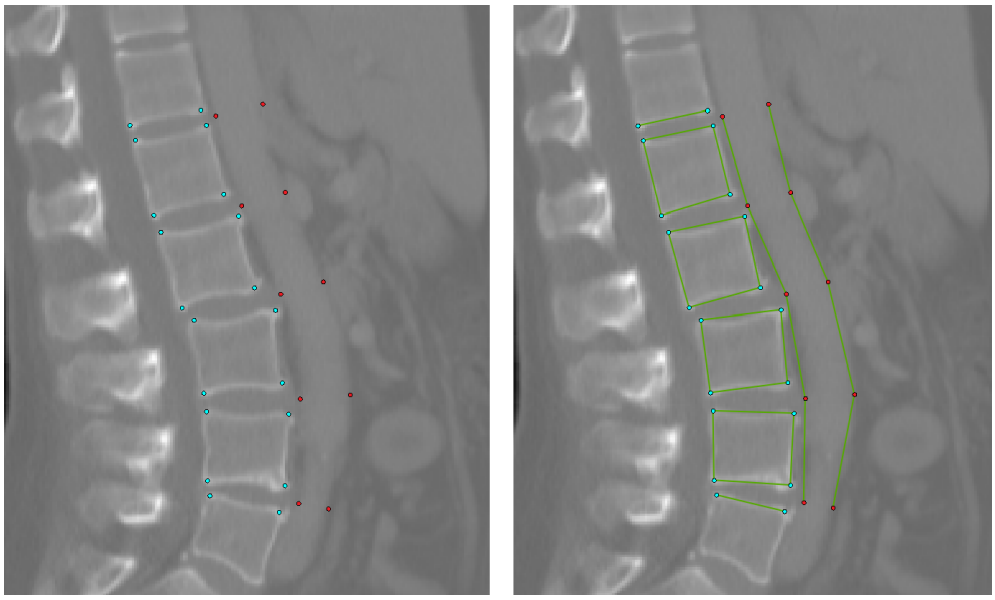

Figure 4.6: Example of a CT sagittal projection with point annotations. The 20 vertebral points are shown in blue, with the 10 aortic annotations in red.

image data itself, a point distribution model (discussed in Section 3.1.1) was trained to capture the spatial relationships. Each set of 30 points could be represented as a point in a 60-dimensional space. The mean of those points represents the mean shape of the model and the principal components that contained the most shape variation were found. These orthogonal modes of shape variation were calculated in decreasing order until 95% of the variation had been included.

Each mode included position information for multiple points and so if the PDM was given a subset of the points, the modes contained information about the location of other points. Using all the shape modes may not have been ideal though. After a certain number of modes are added to the model, the additional modes may have corresponded to variations that did not contribute information about the aortic points and instead added small variations specific to the training data, overfitting the model. The aim was to measure how accurately the model could place the aortic points when presented with the vertebral points. To measure this, the model was first tested on the CT data, as this had annotations to act as the ground truth.

The accuracy of the PDM was measured using cross-validation, where the 100 annotated projection images were separated into training and test datasets. This ensured that the model was tested on images outside of the training process, ensuring it was robust. The cross-validation used 20 folds, 95 training images and 5 test images in each fold. For each test image, the model was given the vertebral points and returned the predicted aortic points.

A measurement of error was produced using the Euclidean distance between the predicted points and the annotated ground truth points for each point across all images. By building the model and testing it with a limited number of modes, a mean error was produced for each number of modes. The optimum number of modes was found by repeatedly calculating the mean error and adding modes of shape variation until the error began to increase. Once the number of modes with the lowest error was known, this was used to calculate a prediction for the region of interest.

Across all 20 folds of the cross-validation, each aortic point was predicted 100 times, once in each image. In the reference frame, each prediction could be treated as a vector displacement from the ground truth, along the intervertebral line defined by the vertebral points. Assuming that the distribution of these displacements was approximately Gaussian with a mean of 0, the standard deviation of the error around each point could be calculated. This deviation was measured separately for each point as it was likely the model could predict some points better than others.

With a calculated standard deviation for the error on each point across the CT images, a region of interest was produced on the DXA VFA images. On each of the 350 images in the DXA VFA dataset, the same 20 point annotations were made at the corners of the vertebral bodies. A semi-automated approach was used, with the superior corners of L5 placed manually. A random forest regression-voting constrained local model (discussed in Section 3.1.1) trained on the CT annotations was used to fit the remaining points, with manual adjustment in the few cases of large misalignment.

Regarding the production of a fully automated system, the automation of vertebral landmark annotation is an active area of research and any automated approach can be integrated into an end-to-end AAC scoring system [182, 184].

A final PDM was built using all 100 annotations. This model was used on DXA VFA vertebral annotations, to predict the aortic landmark points. Though the DXA VFA images had no ground truth aortic annotations, by treating the predicted points as samples of the distribution around the true location, the probability that the true location lies within the ROI could be estimated. With an approximately Gaussian error an ROI produced from these points would expect to include 50% of true aortic points. Translating all predicted points outward by one standard deviation could create the bounds of an ROI containing 68% of ground truth points, and so on.

The predicted anterior and posterior aortic points were extended toward and away from the perpendicular of the intervertebral space. As this ROI prediction was used as part of a tool on large datasets, the risk of excluding part of the abdominal aorta should be very low. For this reason, the ROI used for the pixel classifier consisted of the predicted points from the PDM extended by three standard deviations, including approximately 99.7% of ground truths.

4.2.2 AAC-24 Scoring

With a method to produce a ROI prediction for each VFA image, the next challenge was to use label masks to automatically calculate an AAC-24 score for each image. The predicted aortic points were used to separate the walls of the aorta for AAC-24 scoring, discussed in Section 2.2. Figure 4.7 shows the 8 sections used for scoring, and an example of these sections separating a VFA image.

The pair of aortic points at each vertebral level defined the midline of the intervertebral space. These 5 midlines separated the sections of the aorta adjacent to vertebrae

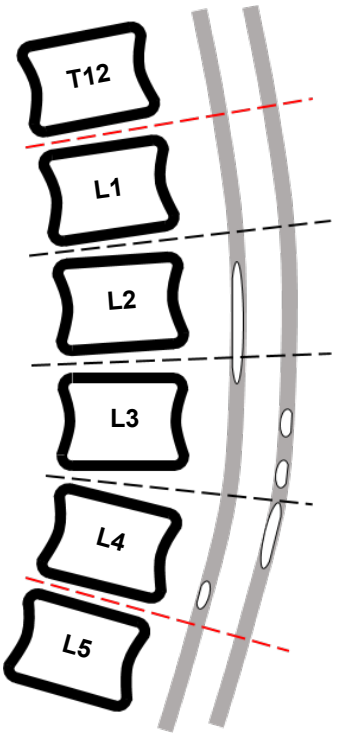

Figure 4.7: Diagram of AAC-24 score calculation for lateral radiography. Total length of calcification parallel to vertebral height is used to generate a 0-3 score for each wall adjacent to lumbar vertebrae L1-4. This example scores 9, The posterior L3 and L4 sections score 1 each. The anterior L3 and L4 sections score 2 each. The posterior L2 with more than two thirds of the section calcified scores 3.

L1-4. Each pair of aortic predicted points, before any translation to create the ROI, were averaged to give 5 mean points which define a prediction of the midline through the aorta. Once these sections of the aorta had been defined, the label mask was transformed to remove curvature in the aorta, standardise the height of the vertebral sections, and create a common orientation for scoring.

These transformations were achieved using thin plate splines (TPS), discussed in Section 3.1.2. Each mask was transformed using a TPS defined by the landmark points, into a common image space. There was some variation in the size of the area covered by the aortic region in each image and the space that the image could be transformed into could take any resolution. For the purposes of scoring images, a resolution of

256x64 was used as this was enough detail for scoring small calcifications without a prohibitive amount of stretching.

The new image space was defined by having the posterior aortic points with an x-coordinate of 0, anterior points an x-coordinate of 64, and midline points 32. The y-coordinates for points at each vertebral space were evenly spread along the height of the image. The TPS created a mapping between the two images defining which pixel values in the source mask were used to calculate each pixel value in the target image. Nearest neighbour interpolation was used to calculate values for coordinates that fell between pixel values, as values can only be 0 or 1. With each mask transformed to a common orientation, minimizing curvature in the aorta, a simple approach to calculation of the AAC-24 score was performed using the proportion of positive pixels in each section of the image. The number of positive pixels in each column of the images were summed to find the column with the maximum value on each side of the midline. This defined an estimated position of the anterior and posterior walls. These estimates were used to calculate the AAC-24 score, by summing the number of positive pixels in each vertebral section of each wall. The AAC-24 score was calculated for each image according to this equation:

$$AAC24 = \sum_{v=1}^{4} \left\lceil \left(3\frac{a_v}{h_v} \right) \right\rceil + \left\lceil \left(3\frac{p_v}{h_v} \right) \right\rceil \tag{4.1}$$

Where c_v is the sum of positive pixels in vertebral level v, along the estimated aortic wall line. h_v is the height of vertebra v in pixels, which in all instances was 64 pixels. This score approximates the technique used clinically (discussed in Section 2.2), with the relative total length of calcification in each vertebral section being scored 0-3 and summed to a maximum total of 24.

Multiclass Label Masks

During annotation, calcification in the DXA images was classified as anterior or posterior wall calcification. This distinction can be used to estimate the position of the aortic midline and score label masks. This alternative was investigated to attempt to improve the accuracy of scoring. With a midline defined by averaging PDM predictions, it was likely that the variable position of the aorta within the ROI was not well represented using the previous method.

The same TPS warping strategy was used to transform the ROI into the 256x64 pixel image space, but this time without including the aortic midline points. The 10 points defining the aortic ROI were used in the same manner, to eliminate much of the aortic curvature and standardise the vertebral heights. The approach was then to calculate a midline through the image that best separated anterior and posterior calcification. This was achieved by optimisation, finding the minimum of a cost function representing how well the classes were separated.

This problem was approached using the hinge loss, commonly used to train support vector machines (SVM). The hinge loss is a function of the distance to a point from a given decision boundary, and takes the form:

$$H(d_i) = \gamma max(\theta - c_i d_i, 0) \tag{4.2}$$

Where c_i is the class of point i, either 1 or -1. d_i is the signed distance from point i to the decision boundary, negative on one side and positive on the other. γ is a constant that decides how steeply the function should increase as the distance increases. θ is a constant that creates a margin in which the loss penalises correctly classified points too close to the decision boundary. This encourages a boundary which maximises distance between the two classes. The result was a loss function that punishes any point on the wrong side of the decision boundary as a linear product of it's distance from the

boundary, while discouraging all points from being too close to the boundary.

As these strategies for scoring will be used in future chapters with automated segmentation approaches, they had to also be resistant to noise. Calculating d_i using the raw distance from the boundary would have given heavy weighting to pixels large distances from the decision boundary, making it sensitive to noisy annotations where there could be clusters of posterior class pixels on the anterior side and vice versa. In an attempt to alleviate this problem, the hinge loss was modified to use the hyperbolic tangent, tanh, of the distance. As the decision boundary was not being used to classify future observations, like with an SVM, there was no need for a margin maximising the distance between the classes, θ could be set to 0. Additionally, to fit a vertical line the distance for any point could be simplified as the x-coordinate of the point minus that of the vertical line. This yielded the following loss function for all points and a given vertical line:

$$L(m) = \sum_{i=1}^{i_n} max(c_i \tanh(x_i - m), 0) \tag{4.3}$$

Where x_i is the x-coordinate of annotated pixel i. m is the x-coordinate of a proposed midline. c_i remains the class for pixel i, 1 for anterior and -1 for posterior (with background pixels having an effective class of 0). Figure 4.8 demonstrates the shapes of the hinge losses as a function of the distance between a single pixel with class 1 and the decision boundary. This shows the decreased impact of misclassified pixels at large distances, and that the new hinge loss remains monotonic. The monotonicity, along with the simplified calculation of distance allowed this loss function to be evaluated efficiently without the need for an SVM.

For each label mask, the positive pixels of each class were summed across columns, creating a histogram with 64 bins. With a candidate midline, the overall loss was quickly calculated by weighting the loss function at each x-coordinate by the number

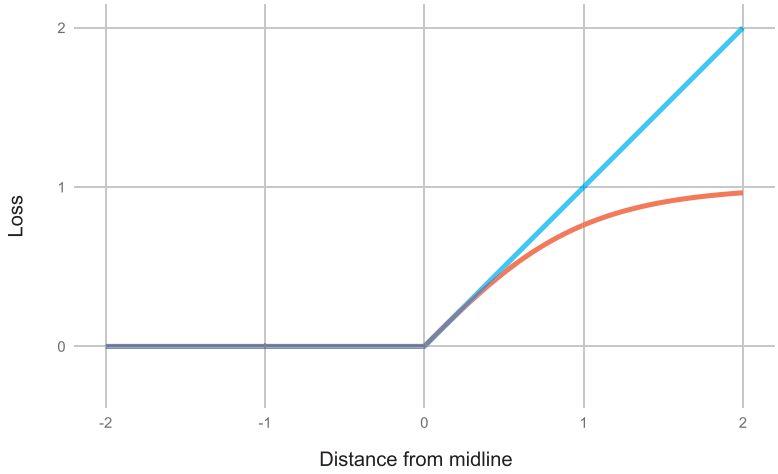

Figure 4.8: Shape of the hinge losses for a single annotated pixel and candidate midline. A pixel contributes 0 to the loss if it is correctly classified, and an increasing amount if it is misclassified, based on its distance. The blue line indicates the standard hinge loss, with the red line the modified tanh hinge loss.

of pixels of that class in that bin. The loss was minimised by proposing candidate midlines at $m = 31$ and $m = 32$, then iterating in the positive or negative x direction to find the minimum, as shown in Algorithm 1.

Once a midline had been found for an image, the AAC-24 score could be calculated in the same manner, using Equation 4.1. The only modification is that during this calculation any anterior class pixels on the posterior side can be ignored, and vice versa.

Algorithm 1 Algorithm for selecting the midline separating anterior and posterior class pixels for an annotated label mask.

Evaluate the loss function L with $m = 31$ and $m = 32$
if $L(32) < L(31)$ **then**
    $x = 32$, $n = 64$, $min = L(32)$
else
    $x = 31$, $n = 0$, $min = L(31)$
for $i := x$ **to** n **do**
    if $L(i) \leq min$ **then**
        $min = L(i)$
    else
        Minimum found. **return** i and **end**

4.3 Results and Discussion

Two main challenges were approached in this chapter. The first was selecting a region of interest containing the abdominal aorta from DXA VFA images. The second was automating the process of analysing label masks to produce AAC-24 scores. Both of these approaches are used in future chapters, to select ROIs to input into segmentation algorithms and to score the output segmentation maps. In this section the results of these approaches are presented, and the impact this will have on the work that follows is discussed.

4.3.1 ROI Prediction

100 CT sagittal projection images were used to build a point distribution model to predict the location of the aorta. The data was split into 20 folds, with 95 images used to train a model to test on the remaining 5. The absolute distance between each model predicted point and the labelled point was used as the metric of error.

The mean error was calculated across all points in all images for each number of modes of shape variation. The total number of modes required to explain 95% of the shape variation was 15. Reducing the dimensionality of the 30 $2D$ point model to 15 modes of shape variation shows that the positions of points within the shape are informative. Table 4.1 shows the mean error on the predicted points for the number of modes used in the model. The error is standardised as a proportion of a reference distance, in this case the width of L5, to allow comparison to any sized image.

Using only the first 4 modes yielded the lowest error on the aortic points, 0.1111, the equivalent of 3.58mm on the mean image. The modes are in order of the proportion of shape variation they encoded, and so the error rate dropped quickly with the addition of the first 3 modes. The fourth mode had little impact and then the error increased gradually with each additional mode, indicating that the remaining modes

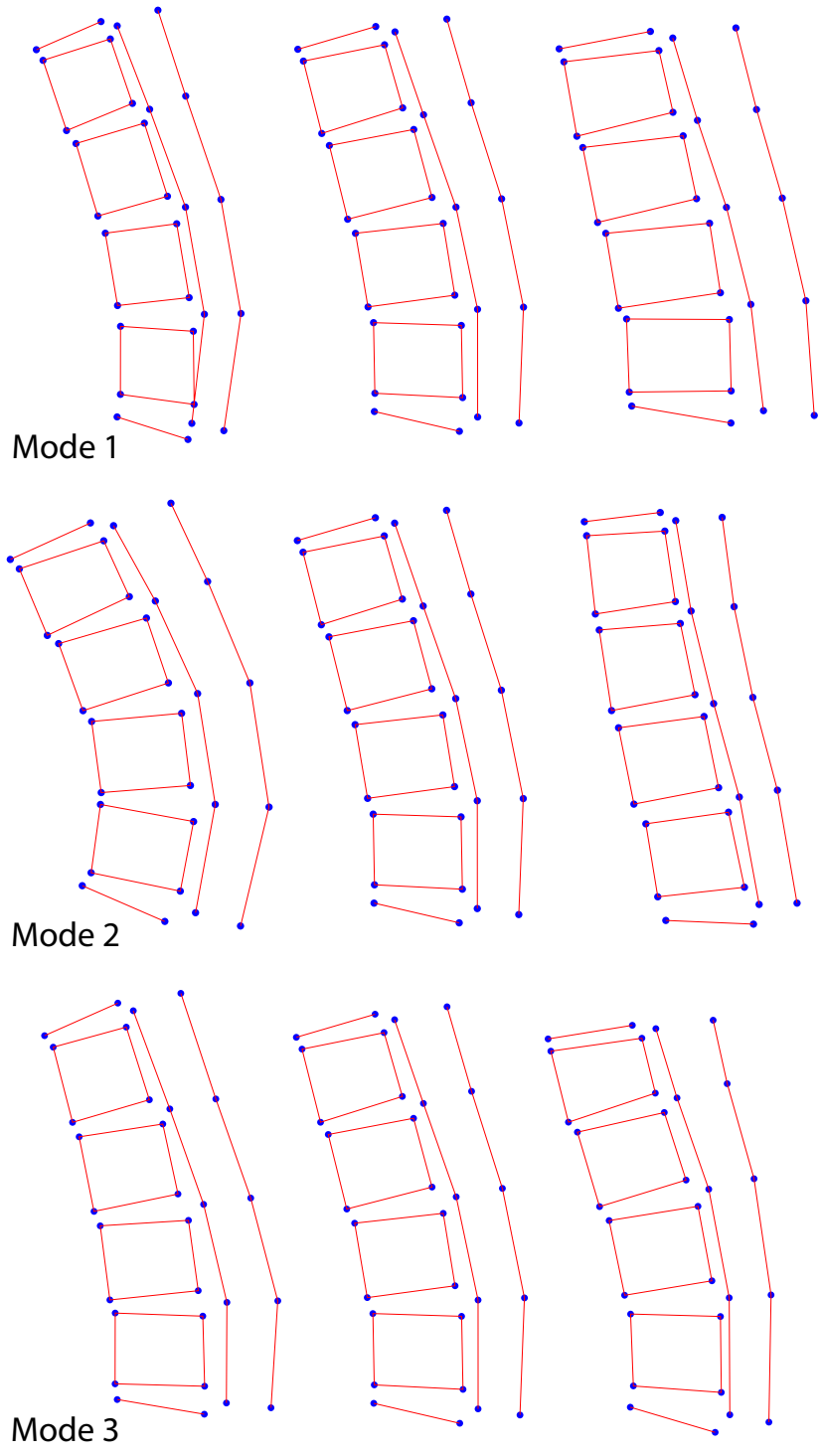

Figure 4.9: Examples of the main modes of the shape model built using CT annotations. The centre image of each row is the mean shape, with the left and right images showing the addition and subtraction of 3 standard deviations along that principal component respectively.

Table 4.1: Mean absolute distance from annotated aortic points to PDM predicted points using different numbers of modes of shape variation. The error is given as a proportion of the width of the L5 vertebrae. The L5 distance in the training dataset had a mean of 35.4mm, an estimate of the mm equivalent of the error on a mean vertebrae is also shown.

No. of Modes	Mean Error	Mean mm
1	0.150	5.32
2	0.120	4.23
3	0.112	3.97
4	0.111	3.93
5	0.119	4.21
6	0.126	4.45
7	0.127	4.49
8	0.129	4.57

No. of Modes	Mean Error	Mean mm
9	0.133	4.72
10	0.131	4.65
11	0.133	4.70
12	0.138	4.90
13	0.142	5.01
14	0.144	5.10
15	0.143	5.04

described variation within the vertebrae that did not affect the aorta or were too specific to examples in the training set. Figure 4.9 demonstrates these first three modes of shape variation from the mean shape. The quick drop-off in error with these modes demonstrated that the vast proportion of the shape information shared between aorta and vertebrae was in the larger more obvious movements. The first mode appears to account for the majority of the concavity and convexity in the spine and aorta. The second and third encompass the width of the two structures and the relative lengths of the vertebrae respectively.

To calculate the region of interest, a measure of the mean and variance of the prediction error was calculated. During the cross validation, the vector describing the translation between each model predicted point and the corresponding annotation was recorded for each number of modes. With 100 of these samples per point across the

Table 4.2: The mean and standard deviation of distance from predictions to the ground truth across all CT projection images using predicted points from a 4-mode PDM, presented as a percentage of the L5 vertebral width.

	Mean Error (Standard Deviation)	
Section	**Posterior**	**Anterior**
T12/L1	0.00051% (10.3%)	−0.0048% (14.3%)
L1/2	0.0047% (9.26%)	−0.0097% (13.04%)
L2/3	−0.017% (9.33%)	0.0017% (14.75%)
L3/4	0.020 (10.42%)	0.0063% (16.47%)
L4/5	0.013% (15.64%)	0.045% (25.74%)

data set, the mean and standard deviation were calculated for 4 modes. The results for all 10 points are shown in Table 4.2.

The mean distances on all points was close to 0, with the highest still under 0.5% of the L5 vertebral width, less than a pixel in all images. The standard deviation could therefor give a good idea of the error on these predictions. It is clear that the anterior points, those furthest from the spine, had higher errors than their posterior equivalents. Higher errors in these points indicates a limitation in how much the position of the spine informed the diameter of the aorta. The posterior wall of the aorta is limited by the posterior wall of the abdomen, but the width of the aorta can be affected by pathology, such as calcification, and age-related changes. The nature of the image projections also meant that any image information outside of the slice thickness was not included, so tortuosity in the aorta that caused deviation in the z-axis will have appeared to change the thickness of the aorta.

The error values also show a higher error for the lowest points on the aorta, at L4-5. During annotation of the training data it became apparent that these were the most variable points. In many of the images it was clear that the aortic bifurcation occurred

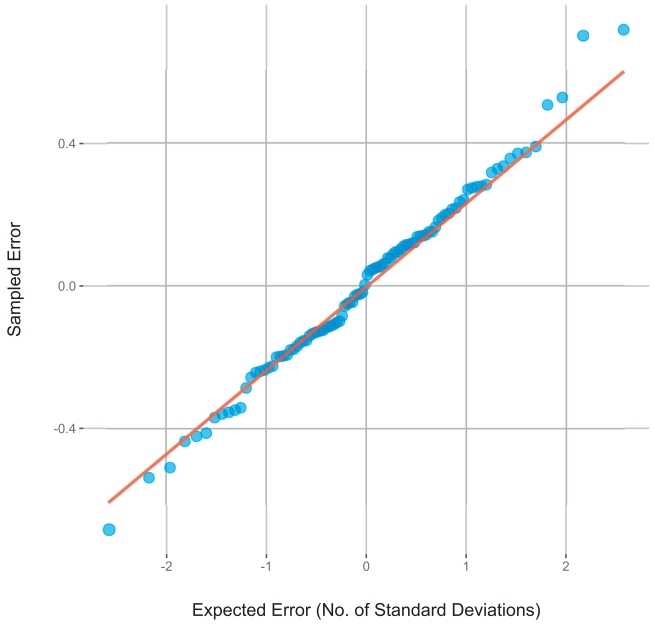

Figure 4.10: A normality plot of distance from predicted aortic points to the annotated aortic points for the L5 anterior point. The red line indicates the expected value for a Normal distribution with mean 0 and the same standard deviation.

above the L4-5 intervertebral space. The level of the bifurcation varies within the population and is a normal anatomical variant. Annotations were made on the most anterior and most posterior wall that was visible, but the appearance of the bifurcation varied between images. In some images both femoral arteries were visible, creating the appearance of a wide aorta, in others however the aorta thinned quickly as the femoral arteries exited the plane. This effect and the varied tortuosity affecting the anterior points were both the result of the limited slice thickness in the projection. In future work it may be worth investigating if these limitations are overcome with an increase in slice thickness. Provided soft tissue does not disguise the appearance of the aorta more accurate annotations for PDM training should be possible with a greater slice thickness by reducing the impact of z-axis movement.

Figure 4.10 visualises the distribution of predicted points around the L5 anterior annotated point. This was the most poorly predicted point but similar distributions were

seen at each other point. The error is the magnitude of the vector from predicted points to the annotated point along the intervertebral line, with negative values for predictions anterior to the true value. The normality plot compares the ordered errors against the expected error for a normal distribution with the same standard deviation and a mean of 0. The more closely the two lines match, the more normal the distribution of the data. It appears that the distribution of predicted points around the annotated point was approximately normal, with some more extreme outliers. This is reassuring evidence that the major informative modes of variation were included, there were no obvious and consistent misalignments that needed to be included in the model.

Though the normality of the data was not perfect, the standard deviation should include roughly the same proportion of samples, meaning it should be rare for the ROI not to include all the aortic points. Reassuringly all 100 training images had their aortic annotations within the predicted ROI during their test fold. After achieving sensible results with the projection images, the point by point errors were used to translate predicted points on DXA VFA images. All 100 training images were used to build the final PDM and tested on vertebral annotation points on the DXA VFA images, to predict aortic points.

The standard deviations of the errors from the cross-validation were then used as translations for the points of each image. All anterior points were extended by 3 standard deviations away from the spine along the intervertebral line, and all posterior points toward the spine. This produced a region of interest with a very low probability of excluding any of the aorta between L1 and L4.

In the majority of images the ROI was applied without issue, but 34 of the 350 images had the ROI extend beyond the image edge. This most commonly occurred with the lower anterior points, which was to be expected as they have the greatest variance. The impact of this problem on the performance of segmentation approaches will have to be assessed in the coming chapters. Overall, the results of the ROI prediction were

encouraging. The aim to select an area of the image containing the abdominal aorta has been achieved.

4.3.2 AAC-24 Scoring

Having produced an ROI prediction for each VFA image, the next stage was to automatically calculate an AAC-24 score for each image, using the label mask annotations. A thin plate spline was calculated to transform each ROI into a common 256x64 pixel image format. This new image eliminated differences in heights of wall sections and allowed easier calculation of AAC-24 scores. This section contains the results of the approaches to this scoring, and discusses how they compare to inter-rater and intra-rater reliability for AAC-24 scoring.

Manual Scoring

In order to contextualise the performance of any machine learning approach to AAC scoring and segmentation, a measure of human performance is required. The DXA VFA datasets contain expert annotated AAC-24 scores for each image. Each of the images in the data set were scored by the author to compare with expert annotations. Image level scoring was performed twice, with a substantial delay between repeats, to allow evaluation of the intra-rater reliability.

Figure 4.11 shows a plot of the repeat scoring by the author, to give a sense of the intraclass correlation. An intraclass correlation coefficient (ICC) of 0.939 was calculated using a two-way mixed-effects absolute-agreement model between the two observations of the 977 images. The 95% confidence interval placed the true ICC between 0.926 and 0.951, inferring an excellent level of reliability. This is an encouraging degree of consistency, matching scoring in the literature of 0.93 [48]. Without the context of the agreement between these scores and the expert scoring however, these values are

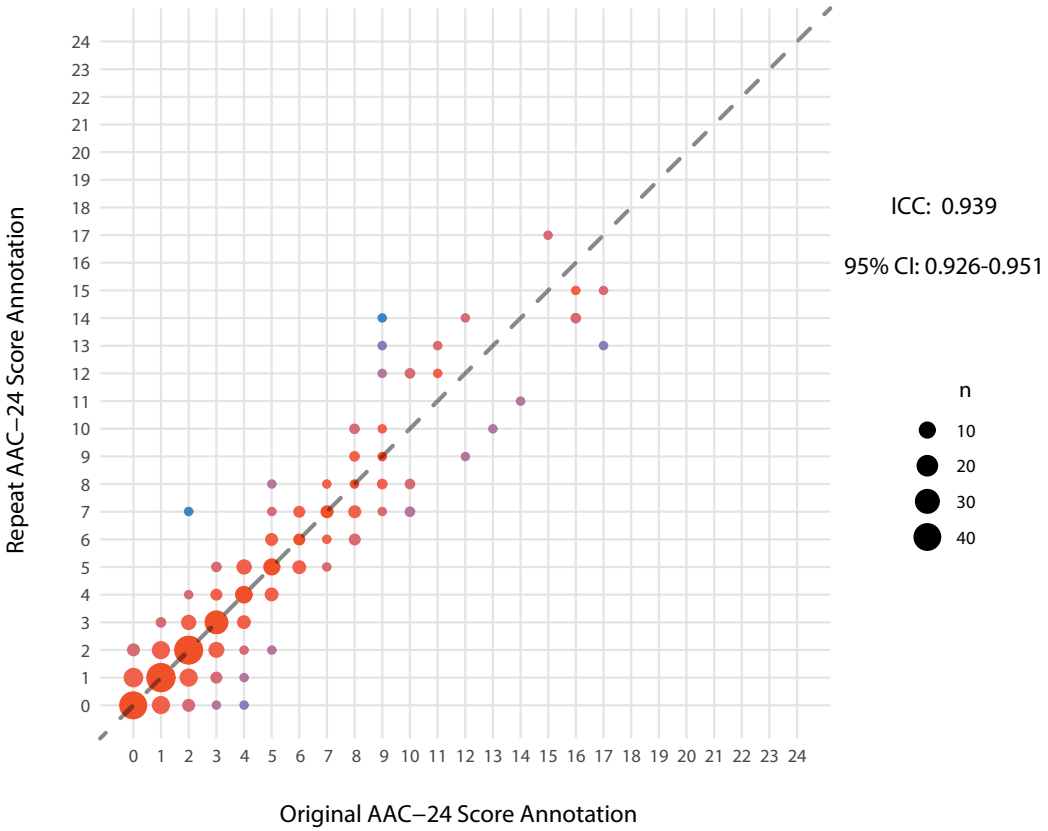

Figure 4.11: The distribution of AAC-24 scores between repeated annotations of the same 350 images by the same reader.

meaningless. Even a stopped clock is right twice a day.

Figure 4.12 demonstrates the agreement between the expert annotation from the data set, and the scoring in this work. An ICC of 0.920 (95%: 0.902 - 0.934), was calculated using a two-way mixed-effects absolute agreement model between the author and the expert annotation on the 977 images. This shows an excellent level of agreement, inter-rater scores in the literature are as high as 0.89 between expert annotators [31]. With 350 observations, this is a larger comparison than previous literature. As ICC can be very sensitive to variability in the data set, the high agreement may be in part due to decreased variance. In this case, a large amount of the disagreement came from high scores, which may have a larger impact in smaller data sets. The ICC is still reassuring and the correlation shown in Figure 4.12 indicates that annotations in this

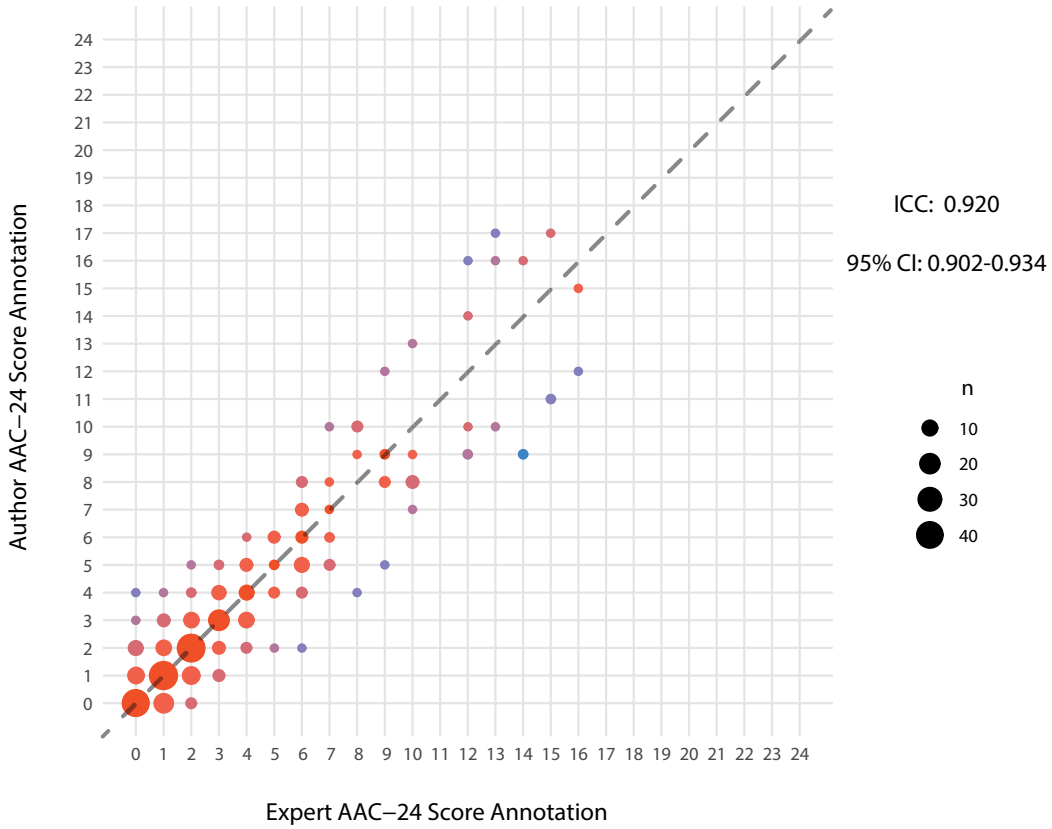

Figure 4.12: The distribution of AAC-24 scores between annotations of the same 350 images by a single reader compared to domain expert annotation.

work can be used as a suitable proxy for expert annotation to demonstrate the value of an automated system.

These metrics for inter-rater and intra-rater reliability form the context for the expected performance for any machine learning approach to automating scoring using this data set. The ICC of 0.939 for reliability between repeat scoring provides an estimate of the maximum agreement that could be expected from an exceptionally performing automated scoring method. As a measure of error on scoring AAC, this is the metric to which the automated AAC-24 methods will be compared. The inter-rater reliability ICC of 0.920 conveys the potential for improvement in any automated method, by using data annotated by domain experts, overcoming the most substantial limitation of this work.

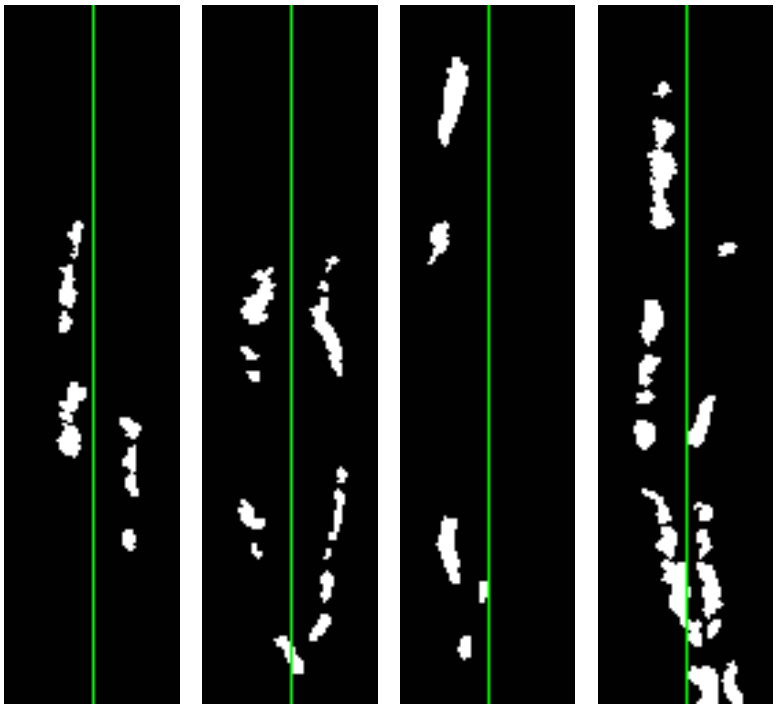

Figure 4.13: Examples of TPS warped annotated masks for scoring along the shape model midline prediction. The midline used for separating the anterior and posterior calcifications is shown in green.

Label Mask Scoring

350 images received pixel annotation of areas containing AAC to assess two approaches to automatic AAC-24 scoring. Figure 4.13 shows several examples of the transformation on label masks for scoring, using midline estimates from the aortic predictions. The majority of curvature is removed from the aorta, though in some instances a considerable amount is still present. These examples also highlight that this estimation of the midline is susceptible to the variable position of the aorta within the ROI, causing both walls of the aorta to fall on the same side of the midline.

The AAC-24 scores generated by the warped midline points were compared to annotations from the author and expert readers. The ICC for the automated score compared to the annotation of the author was 0.901 (95% CI: 0.872-0.922). This remains a

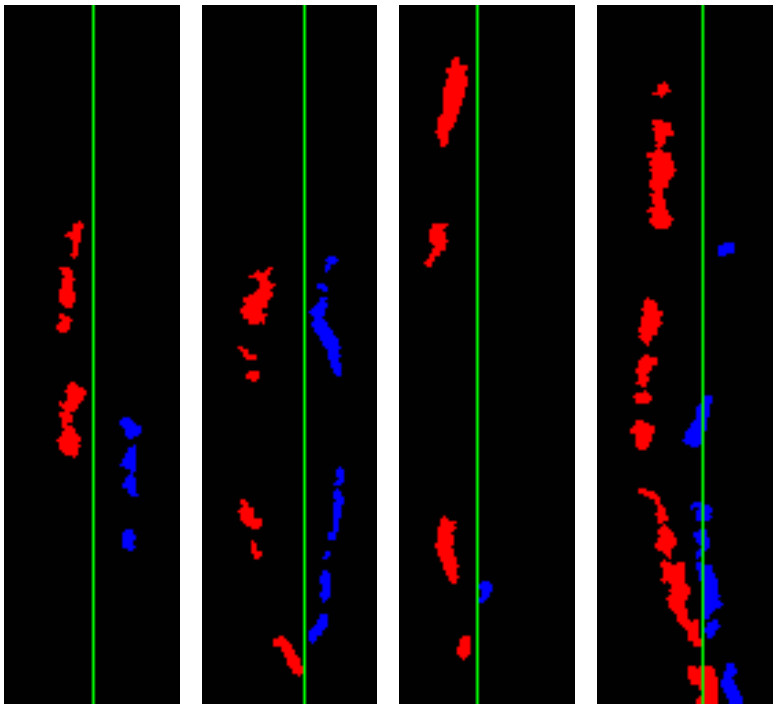

Figure 4.14: Examples of TPS warped annotated masks for scoring along the tanh midline prediction. The midline used for separating the anterior and posterior calcifications is shown in green, with posterior and anterior calcifications shown in red and blue respectively.

good to excellent agreement, but falls short of the 0.939 that would indicate it was acting as another observation by the author. Comparison between the expert annotation and the midline gave an ICC of 0.893 (95% CI: 0.860-0.917). While the agreement between these scores was high, there is room for improvement with the addition of midline optimisation.

Figure 4.14 demonstrates the calculated aortic midlines, an improved fit on the same masks as Figure 4.13. Additionally, an example of masks for which the fit is still not ideal are included. Judging these qualitatively, there were relatively few images where this mismatch was extreme, but the majority occurred in images with high scores. This likely indicates that the increased tortuosity of the aorta that accompanies severe calcification was the main remaining source of curvature in the aorta. With

these examples being few, the optimisation of the midline position, using a tanh loss, improved the agreement between the automated scores and the human annotations.

The ICC calculated between author AAC-24 annotations and the automated scores increased to 0.930 (95% CI: $0.914 - 0.943$) using the midline optimisation, an improvement on the fixed midline method. With a value very close to the 0.939 intra-rater coefficient, this is promising evidence that the automated method is capturing the scoring technique of the reader and accurately reproducing it. A comparison was also made between the expert annotations and this automated method. Figure 4.15 shows the distribution of the scores between the 350 images. This indicates a tendency for the automated method to underestimate AAC-24 scores, with more extreme errors in this direction and a reduction in outliers that overestimate AAC-24 scores. The correlation for midline optimisation was also improved, achieving an ICC of 0.916 (95% CI: $0.897 - 0.932$). This matched the performance seen for inter-rater performance, 0.920, and provides a benchmark for a fully automated system.

The use of the tanh function should also ensure that the scoring is more robust to noisy annotation in the masks. This will be assessed in future chapters, where these techniques will be used to automatically score masks produced by segmentation algorithms.

4.4 Conclusions

This chapter has presented the methods and results for automatically selecting a region of interest containing the abdominal aorta in VFA images, and automated conversion of manual label masks to clinical AAC-24 scores. These two strategies are used throughout this work in order to produce training data for and assess the performance of automated segmentation techniques.

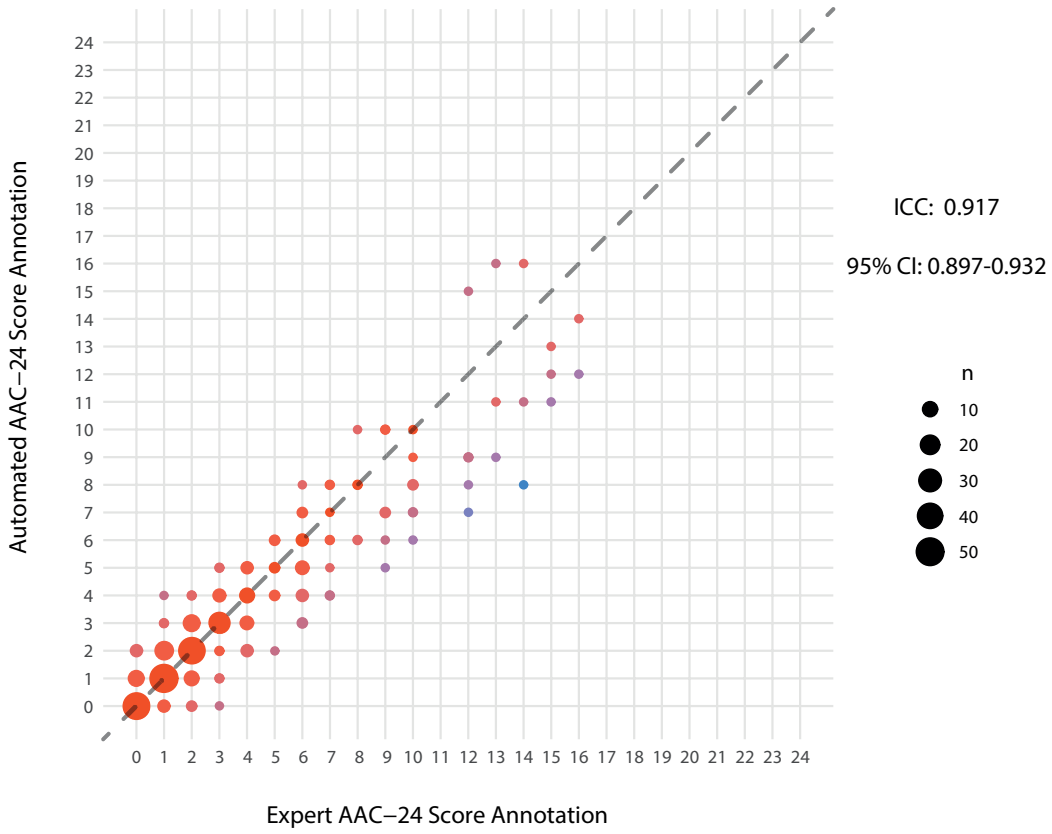

Figure 4.15: The AAC-24 scores produced by an automated mask scoring system with midline estimation on manual pixel-annotation of AAC. Automated scoring of manual annotations are compared to expert scoring of the same 350 images.

Region selection has been very successful, producing ROIs that include all annotated calcification across the 350 images that will be used in future chapters. Automated scoring showed impressive agreement with manual scoring and will be used in future chapters to evaluate performance, where the strategies for automated scoring will be assessed and compared further.

Chapter 5

Random Forest Approach to Segmentation of Calcification

With a predicted region of interest (ROI) containing the abdominal aorta, a segmentation algorithm can be developed to identify abdominal aortic calcification (AAC) within this region. This chapter explores the use of random decision forests on this task. Additional detail on random forests can be found in Section 3.2 and will be referenced throughout the chapter. The methodology to establish hyperparameter choices, evaluate cross validation performance, and assess image segmentation accuracy of the classifiers is presented. The results of these experiments are then discussed and compared to the performance of previous attempts at AAC segmentation in the literature.

5.1 Data and Resources

The random forest was trained and tested using a dataset of 350 DXA VFA images, acquired in single energy mode, along with corresponding annotated masks. Additional details on this data can be found in Section 4.1. 230 images were used from the MRC

NSHD dataset and 120 images from the CAIFOS dataset, along with their corresponding pixel-wise mask annotations of anterior and posterior aortic wall calcifications. Of these images, 297 contained AAC. 53 images without AAC were included to assess performance with masks containing no positive examples. The coordinates of the 10 landmark region of interest were included for each image, defined by the point distribution model in Chapter 4. Expert AAC-24 scores for the 350 images were also used to compare the scores generated from the segmentation masks.

20% of images were held out as a test dataset. Initial experiments to optimise the random forest and choice of image patches were performed on the remaining 280 images. This enabled experimentation with threshold choices and post-processing of segmentation masks to improve segmentation accuracy, while being able to assess the generalisability of these optimisations. Stratified sampling was used to create each image set, taking a proportional number of images from each class of AAC severity: none, AAC-24 score of 0; mild, AAC-24 score 1-2; moderate, AAC-24 score of 3-5; and severe, AAC-24 score of 6-24. Though the absolute number of pixels annotated as containing AAC will vary between images, depending on the relative size, number and position of the calcification. Stratification using these classes encourages a proportional mix of AAC severity in all training and testing splits.

5.2 Methods

The methodology of this chapter focuses on a patch based approach to training a random forest classifier to identify AAC. Random forest is an ensemble learning method, averaging the outputs of multiple decision trees to improve prediction accuracy. This method has been used successfully in a range of applications in medical imaging, from segmentation of the spine or proximal femur in x-ray modalities [68, 70], to tumor and stroke lesions in MRI images of the brain [90, 187]. The first task was to sample

patches from VFA images, train a random forest classifier, and validate the accuracy of classification. Image level predictions were then be made by testing patches sampled from all pixels in an image, enabling automated scoring.

5.2.1 Random Forest Optimisation

The random forest was trained using patches sampled from the VFA images. A patch based approach was selected as, without a consistent shape or location for individual calcifications, borders could not be defined for landmark based approaches. Using features extracted from image patches allowed analysis of local texture to differentiate calcification from background. Pixel-wise mask annotations were used to define the positive and negative classes. Both anterior and posterior calcifications were counted in the positive class.

Each sample patch consisted of a 21x21 region surrounding the target pixel. 1000 positive pixels were targeted for sampling of these patches, with an equal number of negative. Additionally, random image augmentation was applied to create multiple patches per sampling point. A random scaling factor was selected between 0.5 and 2.0, and a random rotation between -0.5rads and 0.5rads. The patch was sampled from the image without augmentation, and then 4 additional patches were sampled with random augmentations, yielding 10,000 samples. Each patch was normalised to have a mean intensity of 0 and standard deviation of 1 before feature extraction.

Patches were sampled around positive and negative target pixels within the ROI without replacement, as demonstrated in Figure 5.1. This approach constrained negative samples to those which would be most informative, from the same region but without the presence of calcification, encouraging feature selection specific to calcification. Positive examples were sampled from each image in the dataset proportional to the total number of positive pixels in each image. Negative examples were sampled

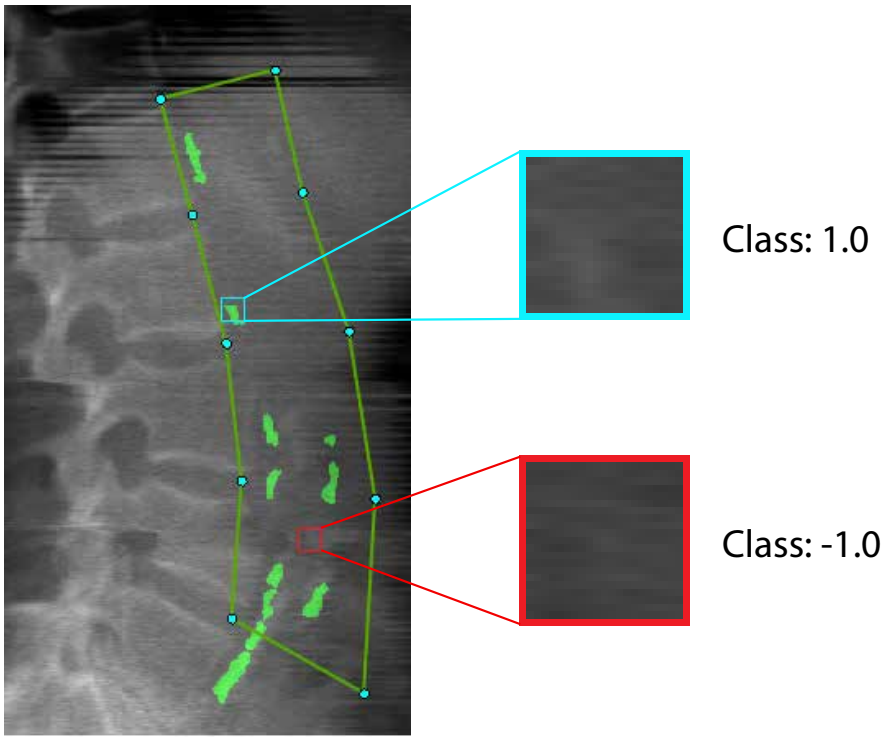

Figure 5.1: An example of a positive and negative patch sampled from within the region of interest. 21 by 21 pixel patches centered on positively and negatively labelled pixels were sampled from the image without replacement. The left-most image is the binary calcification mask overlaid on the VFA image. The patches on the right are normalised and used to produce Haar-like features, which are fed to the random forest with the associated label.

with an equal number of samples from each image, including those with no positive examples. With 20% of images held out for testing, samples were taken across the remaining 280 images.

These parameters governing the nature and number of patches were chosen to allow training and validation to take place in a reasonable time frame. These parameters were later refined based on the performance of image level predictions of calcification. The focus of early experimentation was to optimise the parameters of the random forest itself, to improve classification accuracy. A 4-fold cross-validation experiment was used on the set of image patches to evaluate the performance of a given set of

parameters. 1500 of the image patches and their augmentations were used to train a random forest in each of the 4 cross-validation experiments, predicting classes for the remaining 500 patches. After each fold had been used for testing, the accuracy of classification across all patches was assessed. 4-fold cross-validation was then repeated 5 times, with random perturbations of the splitting of folds. This gave a classification accuracy for each, with a measure of confidence.

Features and Parameters

Random decision forests were built and tested using 4-fold cross-validation. Each forest consisted of t regression trees. Regression trees were used alongside the patch labels 1.0 (calcification) and -1.0 (background), to give a continuous probabilistic value for the prediction, allowing the calculation of an optimal classification threshold. With continuous predictions, receiver operating characteristic (ROC) curves can be calculated and model performance compared using the area under the curve (AUC).

Each tree was trained using Haar-like features (discussed in Section 3.2.2) calculated from the image patches. Each 21x21 patch was used to calculate Haar-like features, with comparison of intensity across the patch at all positions and scales, a large number of features. Decision trees were trained on a subset of these features until they reached a maximum depth or a minimum leaf node size. The subset of features was chosen by taking random steps through the vector of features with a predetermined maximum step size. The number of trees, the maximum node depth for trees, and the maximum size of random steps to select the feature subset were all experimental parameters to maximise performance.

The cross-validation experiment allowed the comparison of the AUC metric between different parameter choices. All trees were built using a bootstrapped sample of the 7500 training patches. Split nodes were created from a random subset of features, using the feature which best splits the positive and negative samples, by minimising

the Gini impurity. Before training a forest, the number of trees to train, the depth of those trees and the proportion of features used in feature bagging were chosen. Each of these parameters was adjusted independently to assess how each impacted the cross-validation performance of the forest, and the training time, to find a suitable combination. Table 5.1 contains the settings tested.

Table 5.1: The tuning parameters of the random forest classifier. The number of trees and mean number of features were adjusted in factors of 2. Tree depth was increased in steps of 5. Each of the parameters was evaluated while keeping the others constant.

Parameter	**Value Range**
No. Trees	$2^4 - 2^8$
Max. Tree Depth	5 – 20
Mean No. Features	$2^{-2}\sqrt{n} - 2^2\sqrt{n}$

The optimum number of trees used to build the forests was first established. Increasing the number of trees in a random forest will typically increase the performance. This improvement is not proportional to the increase in trees and would become increasingly expensive for decreasing improvements to performance [80]. As a result, the number of trees was adjusted by factors of 2, from 16 to 256 to find a practical compromise. For these experiments, the mean number of features was kept at $\sqrt{n}$, where n is the total number of Haar-like features for a 21x21 patch, and maximum split node depth was set at 10.

The impact of maximum node depth will have a less straightforward relationship with performance. Limiting the maximum number of features that a tree can use to inform its decision will reduce overfitting, as it reduces the number of minimally informative features used which could be specific to examples from the training set. However, a tree which is too shallow cannot make use of sufficient features to make an

accurate prediction. Additionally, the lower the maximum depth the shorter the training time. To assess the suitability of this parameter, a range of values from 5-20 were evaluated. The number of trees used was 16, with a mean feature size of $4\sqrt{n}$.

The number of available features has a similar impact to tree depth. Fewer available features reduces overfitting, as each tree would be less likely to share features in the event that a small number of features account for a large proportion of variance. A larger number of features will increase the chances that features of sufficient value are found. To explore this relationship, a wide range of maxima for the random step size were chosen, to give the mean number of features used at each split node. The mean number of features was adjusted by factors of 2, from $\sqrt{n}/4$ to $4\sqrt{n}$. Increasing the number of features available to each split node also has a large impact on the time it takes to train the random forest, as each split node must calculate the optimum feature from a larger vector. To establish the performance of the number of features, the number of trees was fixed at 16, and the maximum tree depth at 10.

Though this was not an exhaustive search of the parameter space, an estimate of the optimum performance on individual patch classification could be established from these experiments. The choice of how patches would be extracted from the training data was then explored using predictions on whole images and comparison with ground truth annotations.

5.2.2 Patch Sampling Optimisation

While the cross-validation experiment was used to determine how the random forest model performed on random subsets of pixels in the image, to evaluate how the classifier would perform in the intended application, whole images were used for prediction. With an optimised random forest established on individual patches in the cross-validation, patches centred on every pixel across the whole image could be fed

into the random forest to predict classes.

280 images were available for training and validating performance. A 4-fold cross-validation experiment was used to determine the accuracy of predictions across the whole image set. With each run, 210 image-mask pairs were used to sample training patches to train a random forest, with the remaining 70 images used to test. Each random forest was built in the same manor as in Section 5.2.1, based on the optimal performance achieved in a reasonable time frame by these experiments. Throughout the cross-validation patch classification experiments, the same patch sampling method was used to optimise the performance of the random forest. 1000 samples were acquired around target pixels of each class, creating 21x21 pixel patches. In addition, random image augmentation was applied, creating 4 additional patches around each pixel.

Table 5.2: The number of Haar-like features that can be calculated for each patch size used for image level prediction.

Patch Size	**Features**
11x11	9036
21x21	119,460
31x31	564,640
41x41	1,722,546

Patch size was adjusted to assess its impact on image level prediction. Increasing patch size allowed Haar-like features to be calculated across a larger area, giving an increasing amount of context from the image around the target pixel. However, the nature of Haar-like feature calculation causes a rapid increase in the number of features as the patch size increases. Table 5.2 lists the 4 patch sizes that were tested, ranging from 11 to 41 in each axis, along with the number of Haar-like features which can

be calculated. The number of patches sampled remained at 1000 of each class, with 4 additional augmentations during these experiments. A scaling factor was randomly selected between 0.5 and 2.0, and a random rotation between -0.5rads and 0.5rads for each augmentation.

The performance of each patch size was assessed using 4-fold cross-validation. With a prediction generated for each of the 280 images, the optimum patch size can be identified with comparison to the ground truth annotation masks. ROC curves were calculated for classification at different thresholds to allow comparison of the trade off between sensitivity and specificity. Additionally, the Dice-Sørensen coefficient (DSC) was calculated to evaluate segmentation performance. These metrics were all calculated comparing only pixels within the ROI.

As well as the size of patches which are sampled, the number of training examples and augmentation of images would also impact the classification performance. These two factors are linked, as the total number of patches used for training is a product of both. Generally, an increasing number of training examples will improve classification performance, so fair comparison required changing the number of augmented images while keeping the total fixed. The nature of Haar-like features makes them more sensitive to features which are horizontal, vertical or diagonal. As a result, small amounts of rotation should improve generalisability, though this effect could be dominated by additional information contributed by additional training samples. Table 5.3 lists the patch sampling strategies which were tested, to assess the most appropriate to improve segmentation.

Augmentation consisted of applying random scaling and rotation to the original images before sampling patches. A maximum scaling factor was chosen s, with scaling randomly chosen for each patch between $1/s$ and s. A maximum rotation r was also chosen for each experiment, randomly sampling from $-r$ to r. To assess the impact of image augmentation on classifier performance, the magnitude of r and s were adjusted.

Table 5.3: The patch sampling strategies which were tested for random forest classification. The total number of patches is the product of the number of augmentations (Augs.) and the number of unique sampling points.

Total Patches (Augs.)	Aug. Magnitude
10,000 (1)	None
10,000 (5)	Low
10,000 (5)	High
10,000 (10)	Low
10,000 (10)	High
20,000 (1)	None
20,000 (5)	Low
20,000 (5)	High
20,000 (10)	Low
20,000 (10)	High
40,000 (5)	Low
40,000 (5)	High
40,000 (10)	Low
40,000 (10)	High

For experiments with low augmentation r was set to 0.5rads and s to 1.5. For high augmentation, an r of 1.0 and s of 2.5 were used.

With optimised patch sampling methods and random forest parameters, the cross-validation experiments predicted segmentation maps for each image in the training set. The prediction masks consisted of the mean confidence of the decision trees classification of each pixel. Before predictions on the test set, the training set was used to determine the optimum parameters for post-processing and choice of threshold which maximised the prediction accuracy.

Morphology and Thresholding

Morphology techniques can be used to improve segmentation performance. They fundamentally consist of erosion and dilation operations, shrinking or growing the areas in the positive class. A kernel is passed over an image, setting the value of each pixel in the new image equal to the lowest (erosion) or highest (dilation) value within the kernel. Dilation is typically used to combine broken up regions of a segmented object. Erosion is used to remove diffuse noise, any small areas are removed while larger areas remain. These operations can also be combined, such as the use of an erosion operation and then dilation, termed opening, which can be used to eliminate noise with less impact on the larger areas.

With the noise present in DXA images and the nature of using local patch context for prediction, it was assumed that there would be small areas of high response throughout the predicted masks, and areas of low response within calcifications. Experiments were run to find a combination of morphological operations to minimise these effects and improve segmentation accuracy. This was assessed on the 280 predicted masks from the training set, produced by the 4-fold cross-validation.

The kernel used was a square with height and width x. A combination of different erosion and dilation kernel sizes, x_e and x_d were used, with the segmentation accuracy assessed with each. The kernel size and the order of these operations was chosen based on the features of the produced masks, with values for x of 7, 5, 3 or 1 (no change). The performance of these erosion-dilation pairs were compared using the AUC and DSC of the resulting masks. The classification threshold which gave the highest DSC metric in the training set was also identified to inform segmentation of the test set.

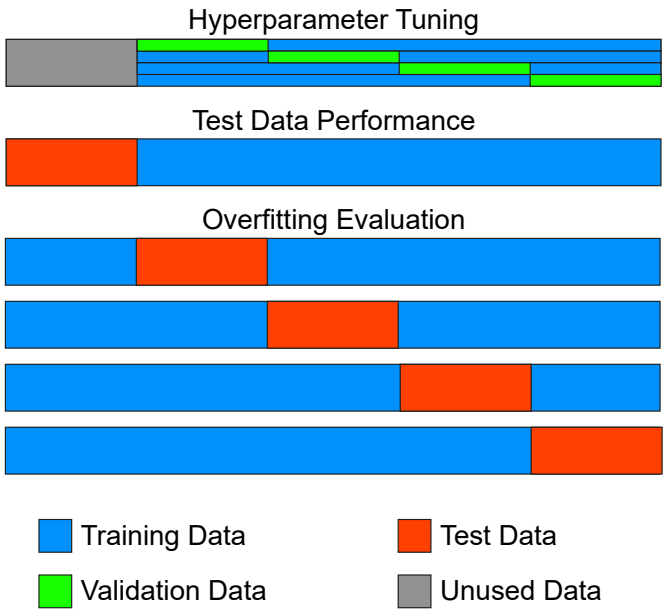

Figure 5.2: 5-fold cross-validation used to assess overfitting of the random forest. Each horizontal bar represents the full 350 image dataset, in the same class stratified order. Comparison of the held-out test data performance compared to other folds gives insight into the degree of overfitting produced by model selection.

5.2.3 Test Segmentation and Scoring

Having determined optimal parameters for random forest classifiers, training patch sampling methods, morphological operations and thresholding on the training data set, the true performance of the random forest approach to AAC segmentation could be assessed on the test set. The 280 image training/validation set was used to train a random forest according to the best performing method established in previous experiments. The random regression forest was trained on these patches and used to predict on all pixels from the 70 test images. Ground truth annotations for the test set were used to assess classification performance with DSC overlap metric to compare to previous work.

The performance of the classifier on the test set was used to assess the degree of generalisability of the approach. To assess the generalisation error of the model selection, additional experiments were used to assess the model performance on the

training/validation set. With a testing/training split of 20%/80% an additional 5-fold cross-validation could be used to build identical random forests and test on folds of the original training/validation split, as demonstrated in Figure 5.2. The extent of any overfitting on the original training set during the parameter optimisation was assessed by comparing the overlap metrics of the held-out test set against those of the training folds.

The predicted segmentation masks were then used to generate AAC-24 scores. These scores were produced with the estimated midline of the aorta from the point distribution model used to predict the ROI. The proportion of positive pixels within the ROI, on either side of the midline and adjacent to each lumbar vertebrae were used to calculate scores. The detailed methods for producing these AAC-24 scores is described in Section 4.2.2. With a score generated for each of the images, the correlation between these automated scores and expert scoring can be evaluated, providing the efficacy of the random forest segmentation approach.

5.3 Results and Discussion

This chapter addresses the training and testing of a random regression forest approach to image segmentation. 350 DXA VFA images were used to train and test the segmentation of AAC using a patch based approach. Cross-validation was used to optimise the random forest parameters. Overlap scores on image level predictions were then used to optimise patch selection, and final performance of the model was then assessed with automated scoring. In this section the results of this model are presented, and compared to previous work in this area.

5.3.1 Random Forest Optimisation

A 4-fold cross-validation experiment was undertaken to tune the parameters of the random regression forest used to classify patches sampled from the DXA VFA images. 2000 patches were sampled across the 280 images used for training, with 1000 of each class: background and calcification. These sampled patches were transformed with random augmentations to generate a total of 10,000 patches. Haar-like features were calculated for each patch of 21 by 21 pixels, generating 119,460 features per patch. Each fold consisted of 7500 examples used for training and 2500 for testing, with 5 random perturbations of these splits to produce a mean performance.

The first parameter of the random forest that was optimised was the number of decision trees to average across in the forest. The performance of these parameters was compared using the area under the ROC curve. Figure 5.3 shows the ROC curves for the repeat experiments with 16 decision trees. These ROC curves were averaged to produce a mean ROC curve, with the AUC used to compare the performance between forests. The consistency of the AUC and ROC curves indicates that the impact of training and testing splits is minimal, that there is sufficient training data to assess parameter choices.

The mean ROC curves of the cross-validation for each choice of forest size are shown in Figure 5.4, with Table 5.4 quantifying the differences in performance. The maximum depth of each tree was limited to 10 nodes for these experiments and feature bagging was implemented by taking random steps through the features with a maximum step size of 690 at each node. This gives a mean sample size equal to the square root of the number of features available. The best AUC performance achieved was 0.880, by both 128 and 256-tree forests. The relatively high AUC across these experiments is a good indication that the random forest is able to identify meaningful Haar-like features from the image patches to allow classification. The smooth and

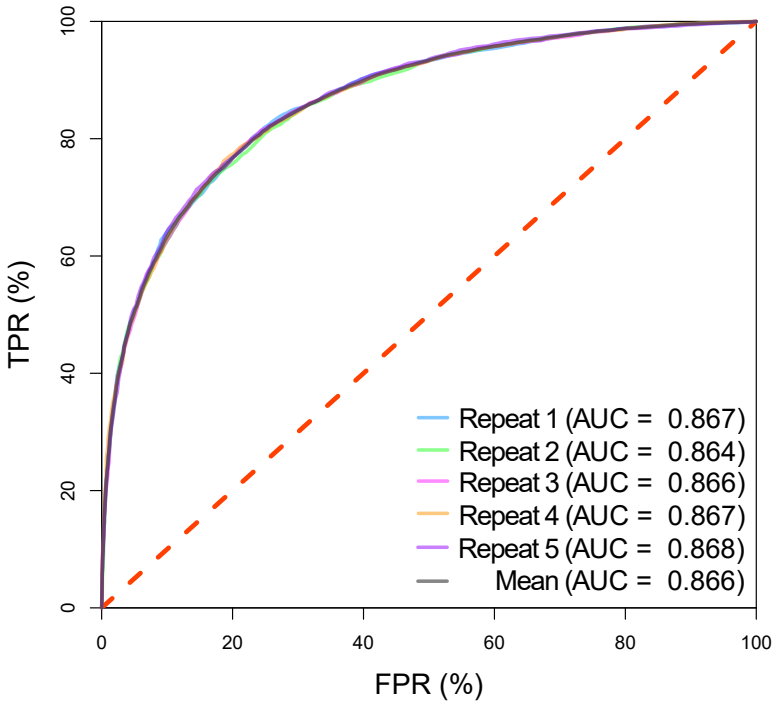

Figure 5.3: The ROC curves for the random forest built with 16 trees. Each repeat of the cross-validation experiment is shown, along with the average curve. These ROC curves demonstrate the TPR and FPR of the cross-validation classification of patches with varying thresholds from -1 to 1 for assigning the positive class.

symmetrical ROC curves do not indicate a clear optimal trade-off between sensitivity and specificity, a classification threshold would have to be determined from image level prediction experiments.

Based on the AUC of each configuration it was clear that increases in performance from larger forests rapidly shrink. The time to train a forest was approximately linear with the number of trees. With random forests achieving a statistically similar AUC score with both 128 and 256 decision trees, while doubling the training time. Indeed, the probability that the 64-tree and 128-tree forests would give these results if there were no difference in their performance is in excess of 20% on a two-sample t-test. A 64-tree forest gave a significant (p-value = 0.01) improvement over 32-tree and seemed a reasonable compromise of training time and performance.

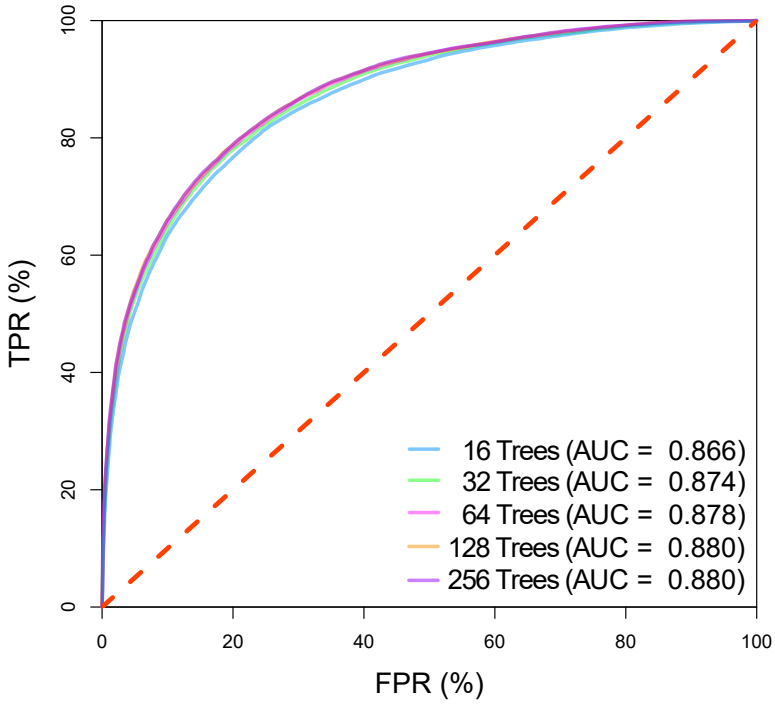

Figure 5.4: Mean ROC curves for random forest classification with different forest sizes.

Table 5.4: The performance of the cross-validation experiments for random forests of different sizes. The mean AUC, across repeat resampling of training and testing splits, is shown as a percentage along with the standard deviation. TPR and FPR percentages are also shown for a classification threshold halfway between classes. The training time is shown for 4 folds and 5 repeats.

No. Trees	AUC (σ)	TPR (σ)	FPR (σ)	Time (min.)
16	**0.866** (0.1)	**75.9** (0.4)	**19.3** (0.4)	**133**
32	**0.874** (0.2)	**76.5** (0.3)	**18.6** (0.3)	**245**
64	**0.878** (0.2)	**76.9** (0.3)	**18.5** (0.3)	**451**
128	**0.880** (0.2)	**77.0** (0.1)	**18.2** (0.5)	**949**
256	**0.880** (0.2)	**77.0** (0.3)	**18.2** (0.2)	**2033**

Table 5.5: Performance of random forests with varying decision tree depths.

Tree Depth	AUC (σ)	TPR (σ)	FPR (σ)	Time (min.)
5	**0.845** (0.1)	**72.0** (0.5)	**20.0** (0.4)	**266**
10	**0.868** (0.1)	**77.0** (0.2)	**19.6** (0.5)	**549**
15	**0.872** (0.3)	**77.7** (0.4)	**20.0** (0.7)	**821**
20	**0.874** (0.2)	**78.4** (0.3)	**20.1** (0.3)	**1146**

The next optimisation was for the maximum depth of the decision trees in the forest. Table 5.5 contains the mean AUC for the 4-fold cross-validation experiments with each maximum tree depth, and the time taken to train them. These experiments were trained with 16 trees in each forest and a maximum random step size of 172 through the features, giving a mean subset of features at each split node of 1382.5 ($4\sqrt{n}$). The maximum performance achieved was an AUC of 0.874. The gains to performance quickly slowed as the tree depth increased, which is expected as the early split nodes use the features which account for the largest proportion of the variance. Doubling the tree depth to 20 allowed this 16-tree forest to achieve comparable performance to the previous 32-tree forest. However, this forest had 4 times as many features to use, and took a considerable amount more time to train. To truly assess the efficacy of increasing tree depth, the impact of feature sample size was examined.

The number of features subsampled at each split node was adjusted to assess its impact on classifier performance. Table 5.6 shows the AUC for each parameter setting, with a maximum performance of 0.869. There was very little change in performance within this range of features, with significant gains in training speed for smaller subsets. The $2\sqrt{n}$ and $4\sqrt{n}$ experiments had almost indistinguishable results. These experiments had a consistent forest size of 16 and maximum depth of 10.

Across these optimisation experiments, the biggest impact in performance was the

Table 5.6: Performance of random forests with split nodes trained on different sizes subsets of Haar-like features. Features were selected by taking a random walk through the feature vector with a maximum step size equal to twice the desired mean features.

Mean Features	AUC (σ)	TPR (σ)	FPR (σ)	Time (min.)
86.5 ($\sqrt{n}/4$)	**0.862** (0.1)	**75.1** (0.5)	**19.3** (0.3)	**53**
173 ($\sqrt{n}/2$)	**0.864** (0.1)	**75.5** (0.4)	**19.3** (0.4)	**79**
345.5 ($\sqrt{n}$)	**0.866** (0.1)	**75.9** (0.4)	**19.3** (0.4)	**133**
691 ($2\sqrt{n}$)	**0.869** (0.1)	**76.6** (0.3)	**19.7** (0.4)	**265**
1382.5 ($4\sqrt{n}$)	**0.868** (0.1)	**77.0** (0.2)	**19.6** (0.5)	**549**

number of trees. Provided there was sufficient depth to each, it appeared that lowering the number of features available to each split node had minimal impact on accuracy while reducing training time substantially. While every combination of these parameters was not searched exhaustively, the patterns established in each parameter individually had reasonable consistency. This allowed the selection of parameters for the forest without concern that there were significant improvements possible in reasonable time scales. All subsequent experiments were performed using a forest size of 64, a tree depth of 10, and a mean feature subset of $\sqrt{n}/2$.

While relatively high TPR was achieved with low FPR, the extreme imbalance in classes in a full image meant that there would be a substantial number of false positives across the image. The sampling of the patches at random from the ROI meant that negative examples were unlikely to be proximate to the positive examples. The negative pixels closest to areas of calcification would be the most uncertain. Given that these were unlikely to be tested, the performance of the classifier would be overestimated in this cross-validation. To assess the performance of the chosen random forest configuration without these biases and to optimise patch selection, image level prediction was

investigated.

5.3.2 Patch Sampling Optimisation

With optimised random forest parameters, this configuration was used for image level prediction to further assess classification accuracy and to assess the method for patch sampling. 4-fold cross-validation was used to train the random forest on random patches from the training folds and predicting on whole images in the test fold. The strategy for patch sampling was adjusted with each cross-validation experiment to compare performance, starting with patch size.

Figure 5.5 shows the segmentation masks produced by the random forest for each patch size. While every pixel in the VFA image is classified in these masks for illustration, only the pixels within the ROI are used to quantify performance. The value of each pixel relates to the probability that it belongs to the calcification class, as estimated by the random forest. For most of these images, the response in areas containing calcification was high, which reassured that the patch based approach has captured sufficient context to classify. However, there is a high response for many pixels in the same area. While noisy responses could be filtered out, the masks contained many clusters of high response pixels that take on the appearance of calcification.

These pseudo-calcifications appeared to be less numerous with increasing patch size, indicating that the widened view provided context important for eliminating patterns in soft tissue which had similar texture to calcification. A side-effect of the widening patch size, was the introduction of a border of uncertainty at the image edge. This region has a response of 0, halfway between classes, where the Haar-like features are so different from the training data that they could not be classified. The width of this region was equal to half of the patch size, where the patch extends beyond the image. As DXA VFA images are taken primarily to investigate the vertebrae, it is common

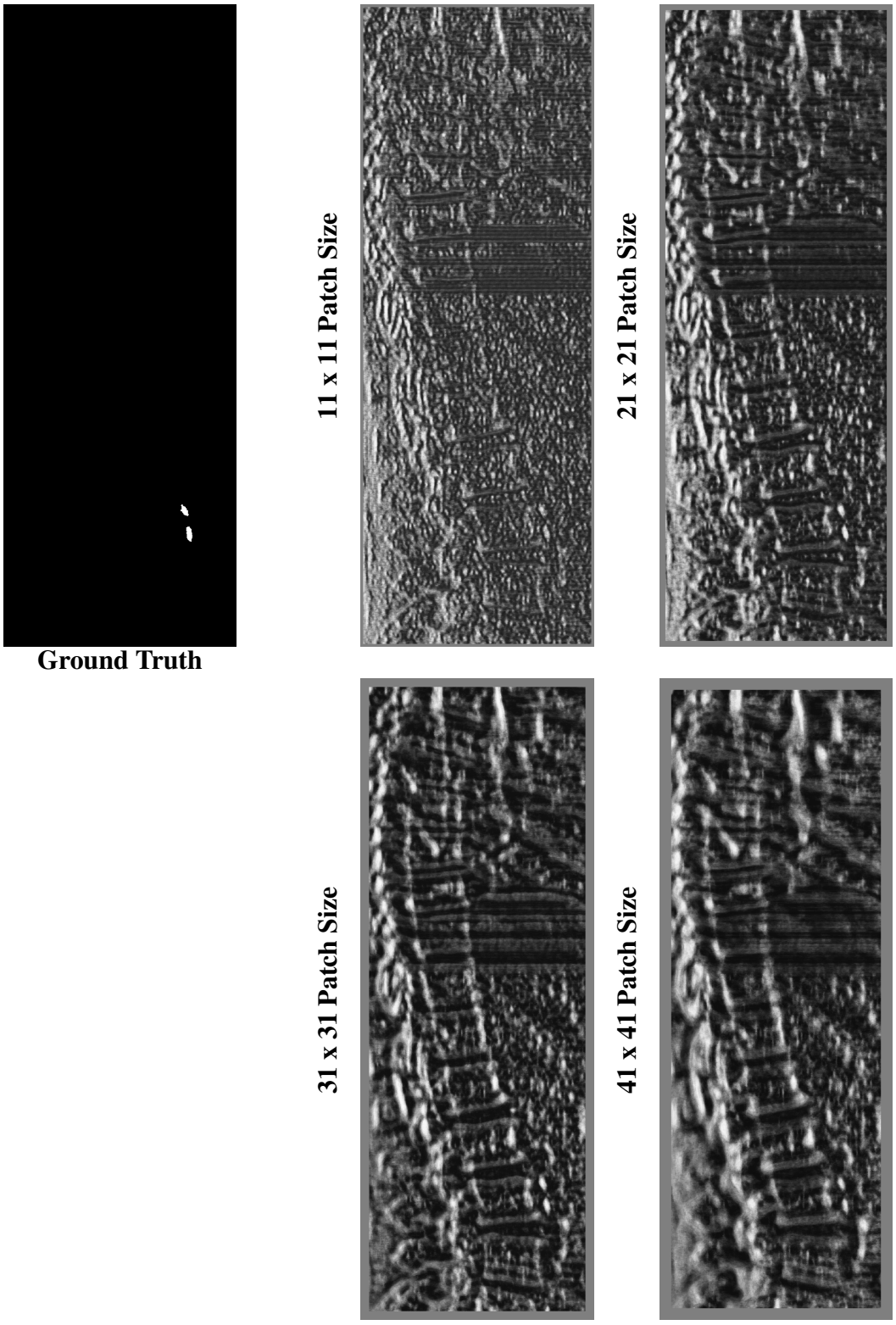

Figure 5.5: Segmentation masks for patch based random forest classification, varying patch sizes. The intensity of each pixel is proportional to the probability from the classifier that it belongs to the calcification class.

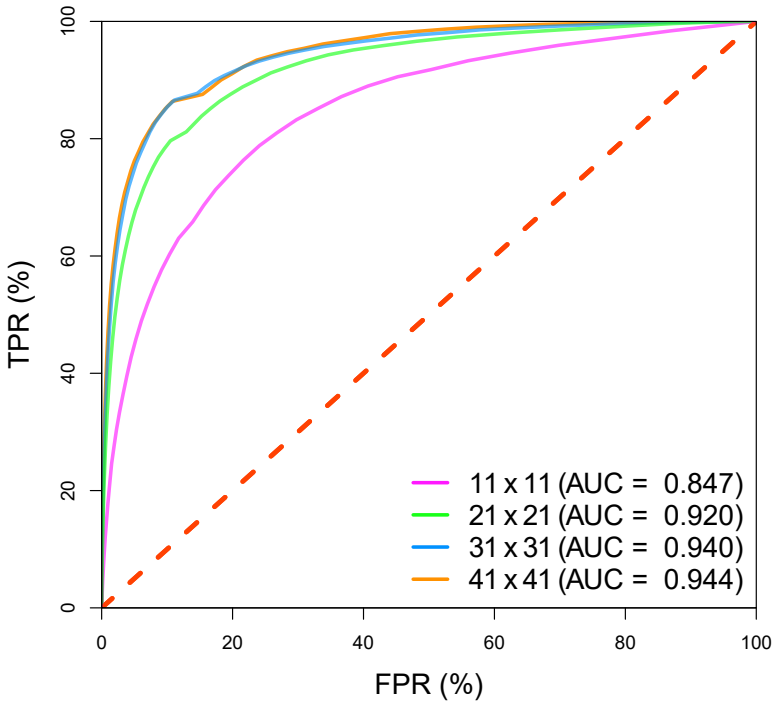

Figure 5.6: Comparison of ROC curves for random forest classifiers trained on increasing patch sizes

for many images to exclude some of the aorta, or for it to be on the very edge of the image. This growing border caused calcifications in some images to be missed.

By adjusting the threshold for classification between calcification and background classes and comparing to the ground truth annotations, an ROC curve could be produced for all pixels across the image set. Figure 5.6 shows the ROC curve for each patch size, along with the AUC. The ROC curves showed improved performance for increasing patch size. Of particular note was the increasing steepness of the slope for small FPR values. This corroborated the qualitative results of the segmentation masks, fewer false positives with widening patch size. The high number of the background class means that accuracy can give a poor representation of classification performance, as giving all pixels a class of -1 would still have given a deceptively high accuracy. A DSC was calculated, which ignores the true negative examples, across the pixels in all images, making it a better metric for comparison of classification.

Table 5.7: Segmentation metrics for random forests trained with increasing sample patch size. Area under the ROC curve and Dice-Sørensen Coefficient are calculated across all pixels in the image set. Training time is for all 4 folds in minutes.

Patch Size	AUC	DSC	Time (min.)
11 x 11	0.847	0.139	29
21 x 21	0.920	0.256	122
31 x 31	0.940	0.306	297
41 x 41	0.944	0.329	701

Table 5.7 demonstrates the performance gains of increasing patch size in terms of AUC and DSC. The best performance was achieved by the 41x41 patch, with an AUC of 0.944 and DSC of 0.329. This DSC is the best value achievable while adjusting the classification threshold. The small 11x11 patches, while rapid to calculate, had unacceptably poor performance compared to larger patches. As the AUC and DSC metrics were calculated across all images in the data set, only one estimate of each was sampled. This excluded true statistical comparison, though the relatively large improvements to DSC with increasing patch size were encouraging. Combined with the improved appearance of the segmentation masks, seen in Figure 5.5, the improving metrics for overlap seemed a worthwhile benefit for the cost of training time. Additionally, in potential clinical application, the training time for these models is likely not to heavily hinder usability, and a much greater value is placed on accuracy.

The number of patches sampled and the degree of image augmentation were also adjusted to optimise performance. Table 5.8 shows the variations of these parameters which were tested, and the AUC and DSC metrics. The highest DSC achieved was 0.342, with 8,000 patches sampled with 5 augmentations each. The results of these experiments indicated that the most impactful parameter on performance is the number of patches, with increased sampling improving performance more than increased

Table 5.8: The performance of the patch sampling strategies tested for random forest classification. The total number of patches is the product of the number of augmentations (Augs.) and the number of unique sampling points. The DSC is the highest achievable when adjusting the classification threshold. The time taken to train the 4 random forests is shown in minutes.

Total Patches (Augs.)	**Aug. Magnitude**	**AUC**	**DSC**	**Time (min.)**
10,000 (1)	None	0.941	0.327	346
10,000 (5)	Low	0.935	0.313	352
10,000 (5)	High	0.916	0.233	350
10,000 (10)	Low	0.927	0.280	354
10,000 (10)	High	0.914	0.233	352
20,000 (1)	None	0.950	0.340	686
20,000 (5)	Low	0.948	0.337	696
20,000 (5)	High	0.923	0.263	693
20,000 (10)	Low	0.938	0.321	692
20,000 (10)	High	0.921	0.266	692
40,000 (5)	Low	0.951	0.342	1411
40,000 (5)	High	0.928	0.284	1405
40,000 (10)	Low	0.940	0.327	1414
40,000 (10)	High	0.923	0.269	1417

augmentation to generate more patches. The degree of augmentation was varied during these experiments, with low augmentation outperforming high.

Morphology and Thresholding

With an optimised policy for patch selection and random forest training, post-processing parameters were established to improve predictions. The most impactful shortcoming of the methods to this point was the high number of false positive responses in the segmentation masks. This high rate of false positives agrees with similar findings on radiographs by Petersen et al. [178]. Morphology and thresholding methods were implemented to minimise the impact of these false positives. Given the large areas of high confidence around true calcifications in the ground truth annotations, the main aim of morphological operations was to erode smaller areas of high response and dilate to restore true calcifications. This method of erosion and subsequent dilation is termed closing.

Table 5.9 shows the segmentation performance for predicted segmentation masks after the application of closing operations with varying kernel sizes. The 4-fold cross-validation random forests were trained using 20,000 41x41 patches without augmentation. Comparing the performance metrics before and after morphology, small kernels appeared to improve overlap measures. Both 3x3 and 5x5 kernels achieved an AUC of 0.956, with 5x5 achieving the best DSC measure with 0.367. Further increase to the kernel size resulted in a considerable drop in performance. While these differences are fairly minor, these methods were implemented with negligible compute time, making them worthwhile.

DSC metrics for the morphology performance were calculated at each threshold value, with the highest score presented in Table 5.9. This threshold was 0.725 for the 5x5 kernel size. This is a high threshold, only classifying pixels with high confidence in the positive class. Figure 5.7 shows the same example segmentation mask after

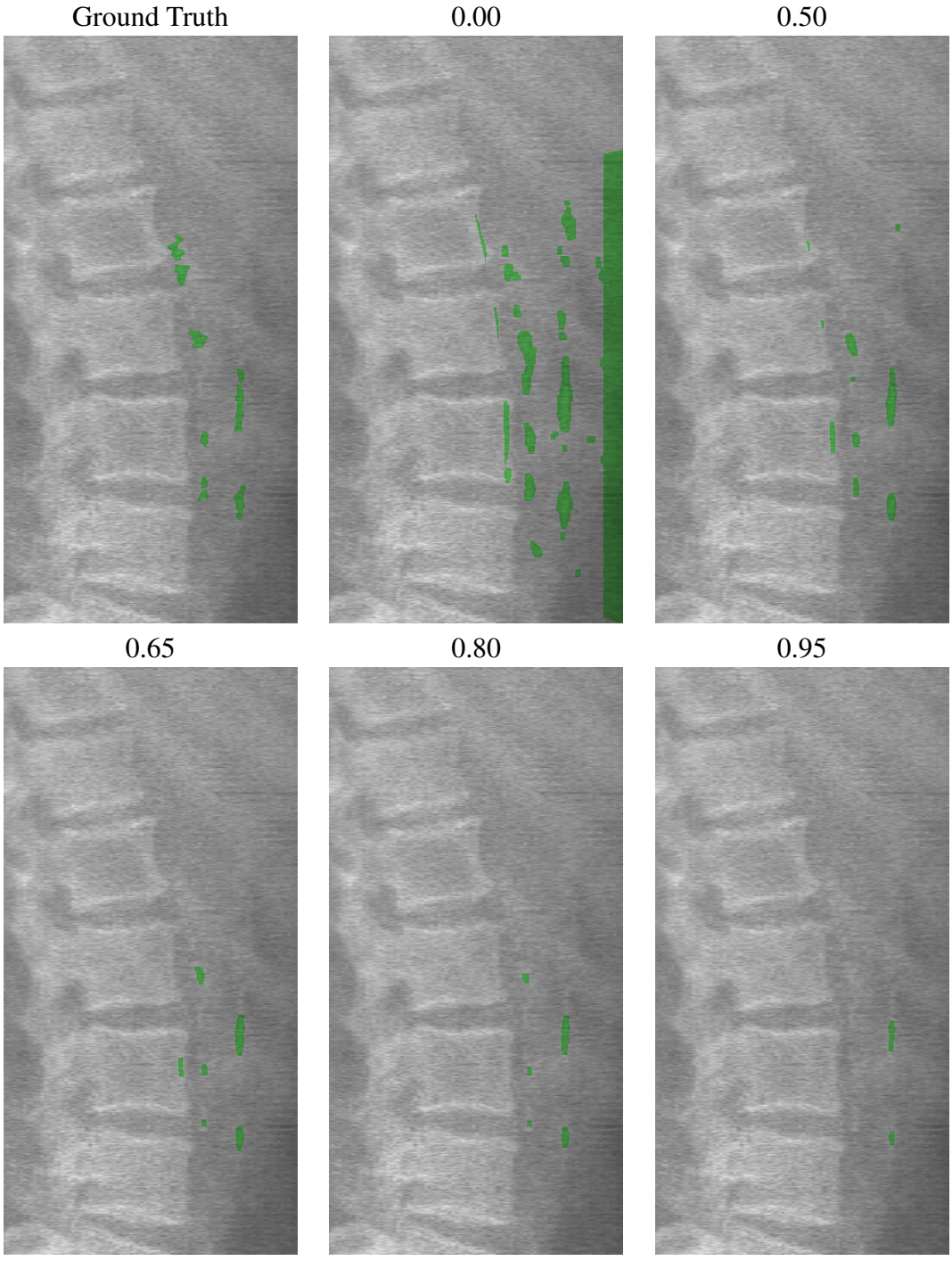

Figure 5.7: A comparison of the same segmentation mask with increasing classification thresholds, overlaid on the VFA image. The ground truth annotation is included for comparison.

Table 5.9: Segmentation performance of predicted masks after application of a morphological closing operation with varying kernel size.

Kernel Size	AUC	DSC
None	0.954	0.344
3x3	0.956	0.349
5x5	0.956	0.367
7x7	0.945	0.333

differing classification thresholds. This demonstrates the need for a high threshold to eliminate areas of false positive pixels. However, this was achieved with the cost of removing areas of true positives. This highlights the importance of the DSC metric, which is sensitive to changes in true positives.

These segmentations also highlight a source of pseudo-calcifications, the vertebral border. Due to the landmark definition of the ROI, some areas of the vertebral body were included in images. This was especially pronounced in images which included spinal pathology such as osteophytes. While thresholding can be used to minimise their impact, the inclusion of the vertebrae in some images is always likely when expanding the predicted ROI to ensure inclusion of the aorta.

With the optimal parameters of the random forest, patch selection and post-processing determined experimentally, the final experiment was to predict on the test split of the data set and calculate AAC-24 scores. The chosen post-processing steps were; closing with a kernel size of 5x5 pixels, and classification thresholding at 0.725.

5.3.3 Test Segmentation and Scoring

All 280 images used for training and validation of model parameters were available to train the random regression forest for segmentation, with 70 images held out for

testing. Using the established random forest parameters, patch sampling methods and post-processing steps, a segmentation mask was produced for each of the 70 images. These segmentations were compared to the ground truth annotations of the data for qualitative and quantitative evaluation. Figure 5.8 demonstrates a number of example images from the test set. Each image is overlaid with both the ground truth and automated segmentations for comparison of areas of AAC.

These examples were chosen to demonstrate some of the shortcomings of the system. There are still a number of images containing false positive areas along the vertebral borders. These are the result of slight misalignment of the vertebral landmark points defining the ROI, which caused aortic points extending to the vertebrae to include a small amount of the vertebral body. Due to the random sampling of patches, it is unlikely that many of these vertebral pixels were sampled, prohibiting the random forest from learning features to discount these pixels. Increasing the complexity of the vertebral shape model, to include annotations along the borders of the vertebral body may be able to minimise this problem. This would be at the risk of excluding some calcifications very close to the vertebrae, as in patients with severe spinal pathology such as scoliosis, the abdominal aorta may be partially obscured by vertebral bodies.

With comparison to the ground truth annotations, the performance of the random forest segmentation was quantified. Table 5.10 shows the segmentation metrics calculated for all images in the 70 image test set, achieving a DSC score of 0.365. With a measure of the model performance on novel test images, additional experiments were performed to assess the generalisation error produced from the model and hyperparameter selection methods. 4 additional folds of the training/test split were evaluated to give this insight. A 5-fold cross-validation was also performed, with the original test set used as the test set for fold 1. This enabled prediction of segmentation masks for each of the 350 images in the dataset. There is some variation in the AUC and DSC metrics between folds, likely a result of the variation in the numbers of positive

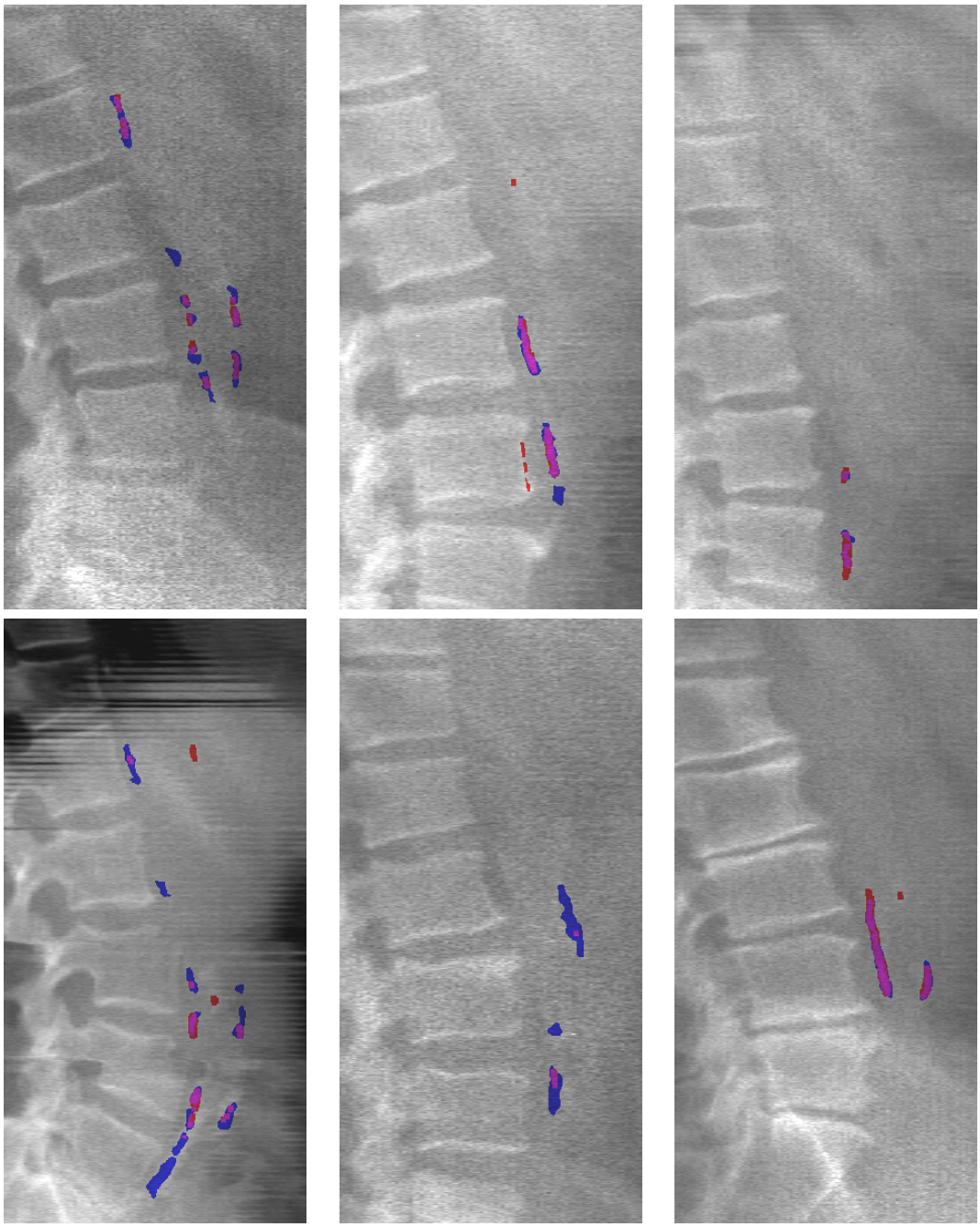

Figure 5.8: Example images from the test set with ground truth and random forest generated masks overlaid. Areas in blue represent false negatives, where the random forest has failed to identify calcification, and red areas indicating false positives, where the random forest has misidentified background as calcification. Magenta indicates true positive pixels, where there is agreement.

pixels in the ground truth annotations for each split. It does not seem that the performance of fold 1 is substantially affected by the optimisation process. As this was the only fold for which none of the test data had been previously used for training, a lower performance could indicate overfitting to the training data. This is reassuring that the morphology and threshold choices were not specific to the training set.

Table 5.10: Overlap metrics comparing the random forest segmentation to ground truth annotations. DSC, IoU and accuracy are all included for comparison to previous work. Fold 1 represents the truest evaluation of performance, as the test data was not part of the training data during experimental optimisation.

Test Data	**AUC**	**DSC**	**IoU**	**Acc.**
Held-Out Test Set	0.952	0.365	0.223	0.991
Training Fold 1	0.944	0.340	0.204	0.992
Training Fold 2	0.968	0.369	0.226	0.991
Training Fold 3	0.963	0.397	0.248	0.992
Training Fold 4	0.958	0.367	0.225	0.992
All Training	0.958	0.368	0.226	0.992
All Images	0.957	0.367	0.225	0.991

The aim is to produce a classifier that can be used to match human identification of AAC. A measure of the performance of human annotation was needed in order to compare how the random forest segmentation performs. As described in Chapter 4, repeat annotations of all images was performed, with significant time between and blind to the first annotations or scores. By comparing the first and second annotations a metric

for intra-rater reliability could be calculated. The DSC achieved in this comparison was 0.784, giving an impression of the maximum performance that could be achievable by a perfect automated system trained on these annotations. The segmentations produced by the random forest approach were considerably less accurate than repeat annotations by the same annotator, indicating poor performance. Petersen et al. [178] found that the IoU between trained radiologists annotating the same images was 0.51, the equivalent of a 0.675 DSC. While this inter-rater reliability measure appears more favourable, it is still much more reliable than the 0.367 achieved by the random forest.

The IoU was included to compare to previous work, with this approach falling short of the 0.28 achieved by Petersen et al. in their automated approach. Their work similarly involved the use of statistical shape models to guide random forest segmentation. A Bayesian framework was used to incorporate spatial and location prior information along with texture analysis to estimate the location of the aorta and segment calcification within. Additionally, an IoU of 0.42 was achieved by Lauze and de Bruijne [176]. An active shape modelling approach was used to estimate the position of the aorta, with an in-painting segmentation technique. Both of these approaches were assessed on radiograph images, which typically have a better resolution and less noise, so achieving the same performance would be more difficult. However, with an IoU of 0.225, there was a substantial gap in efficacy.

The impact of small shifts in the alignment of similar sized segmentations, like those seen in Figure 5.8, have a large impact on overlap measures such as DSC. It is important that segmentation accuracy is high, to allow confidence in interpretability of AAC-24 scores generated from the data. However, these differences in overlapping areas may be the result of ambiguous pixels that would have no effect on the overall severity measure of the calcification. To assess the impact these segmentation inaccuracies have on the measures of AAC severity, the automatically segmented masks were scored. An AAC-24 score was generated for each of the 350 segmentation masks,

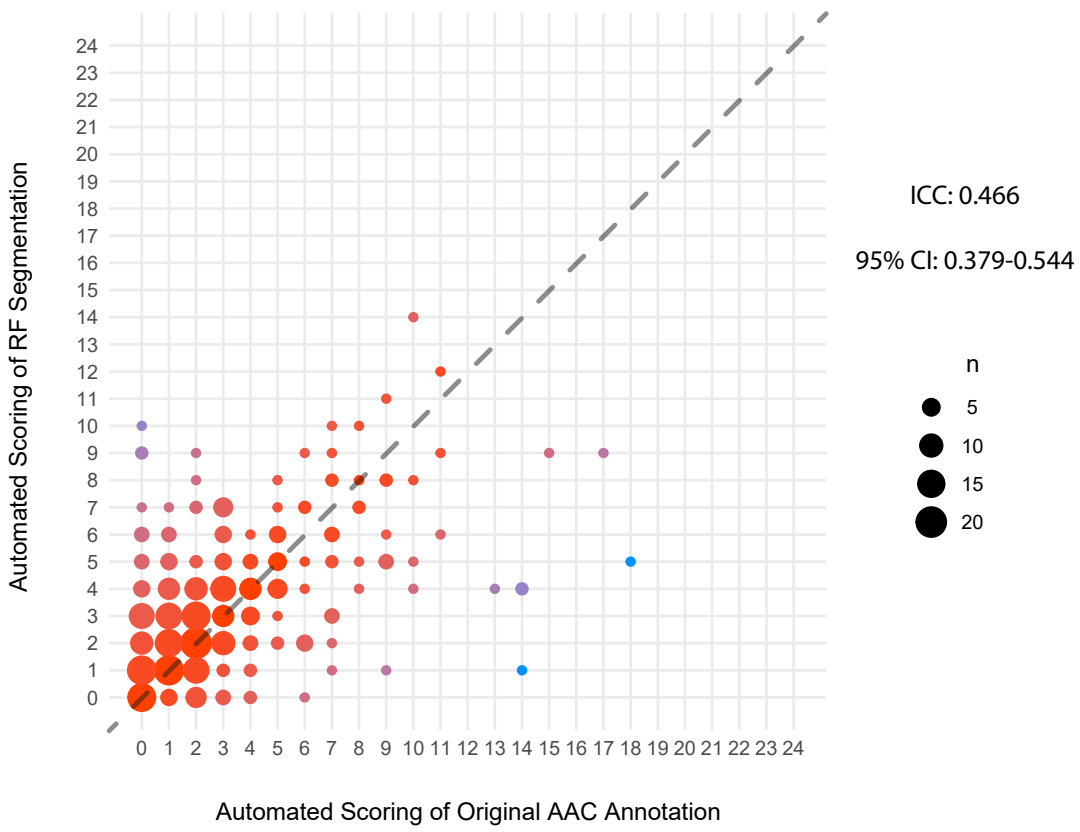

Figure 5.9: Intraclass correlation between AAC-24 scores generated from manual annotations and segmentation masks produced by the random forest.

using the estimated aortic midline methods described in Chapter 4. This was also performed on the ground truth annotations, and an intraclass correlation coefficient (ICC) was calculated between the two methods. The results of this comparison are presented in Figure 5.9.

The overall ICC between AAC-24 scores generated using ground truth annotations and random forest segmentations was 0.466. This was calculated as a two-way mixed-effects absolute-agreement model. This was a low value, indicating unreliable agreement between the two scores. This confirmed that the low overlap metrics do indeed indicated a fundamental failure of the random forest to capture the features in the data which indicate the presence of calcification. The comparison of AAC-24 scores

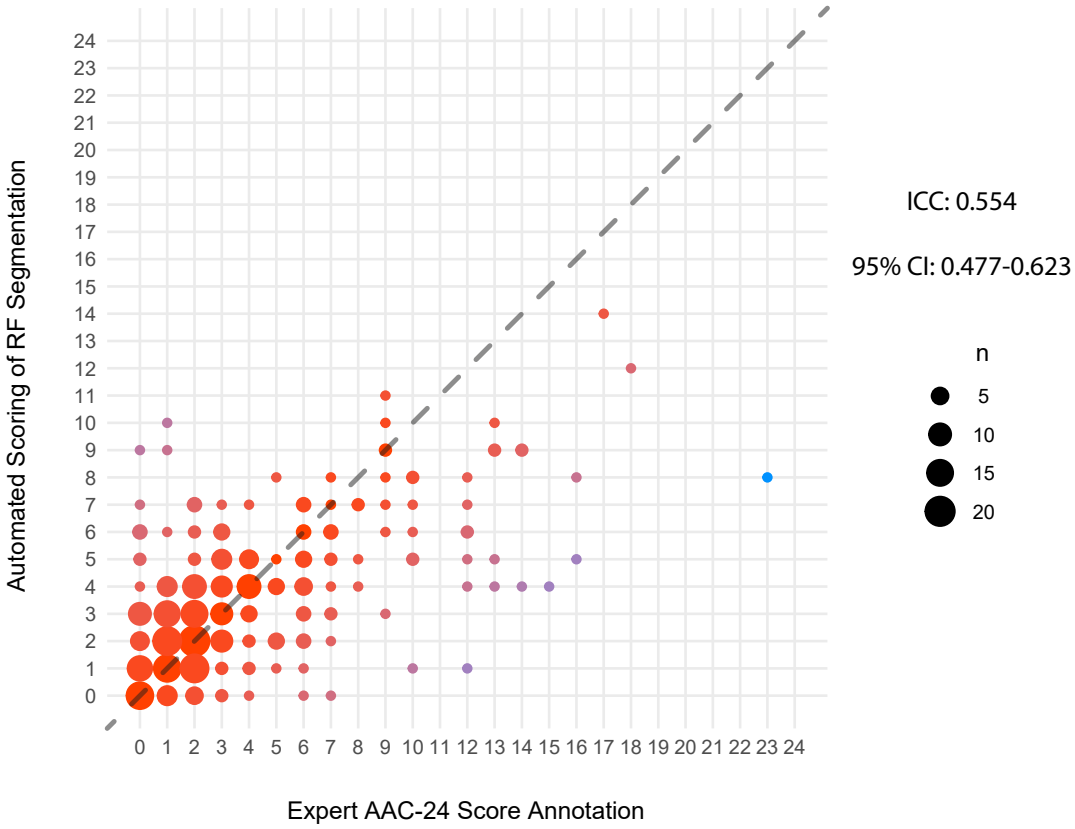

Figure 5.10: Intraclass correlation between expert AAC-24 scoring and those generated by the random forest.

indicated that the majority of errors in the random forest were in underestimating larger scores, and overestimating images containing no calcification. This was likely the result of pseudo-calcifications. To maximise the DSC a high threshold was set, this created a compromise which heavily reduced the true responses in high scoring images, but still left areas of calcification in images which should have been eliminated.

There was also the compounding factor of the inaccuracy of the midline estimates in the shape model. For a true impression of the random forest segmentation scoring performance, the AAC-24 scores were compared to expert annotation of the images. The intraclass correlation of these scores is shown in Figure 5.10. An ICC of 0.554 was achieved between the two sets of scores. This value indicates some weak correlation,

though with a lower bound on the 95% confidence interval of 0.477, this is not strong evidence of correlation. The increase in correlation compared to the scores generated using the same method on manual pixel annotations may indicate that performance is affected by the shortcomings of the scoring method. However, this impact would not be sufficient to explain the low correlation if the random forest is identifying calcification reliably. This is further evidence that improvements to the segmentation approach are required to allow reliable automated scoring of these images.

5.4 Conclusions

This chapter has presented the methods and results for automatically segmenting abdominal aortic calcification in vertebral fracture assessment images using a random regression forest. Achieving a DSC overlap metric of 0.367, the overall agreement of automated segmentation with manual annotation is relatively low, below performance achieved in previous literature in radiograph images. Automated scoring of the images achieved an ICC of 0.554 with expert annotation, leaving a substantial shortcoming in the utility of this approach for replacing manual scoring. The two largest obstacles to this scoring were relying on the aortic midline predictions of the shape model and the considerable number of false positive clusters. Redesigning the random forest to perform multiclass classification, where anterior and posterior calcifications were classified separately, would allow the more accurate weighted midline approach to scoring (discussed in Section 4.2.2) to be used. This would give more accurate scoring from the segmentation, provided a similar degree of accuracy could be achieved. Pseudo-calcifications, clusters of false positive pixels, would still have a heavy impact.

It appears that there is room for improvement in the random forest classification performance. Changes to the structure and training of the random forest are likely to give only small improvements in accuracy. In particular these experiments indicated

that substantial decreases in false positive rate are not easily achievable with optimisation of this methodology. Though more elaborate patch sampling methods are the most promising area for the furthering of this approach. A strategy to improve the selection of negative patch samples could yield improvement. By pairing negative patches from those in the vicinity, choosing a random distance weighted to encourage proximity, the negative samples would enable better distinction of features which represent calcification and not location.

Similarly, there may be benefit in weighting the selection of negative samples to favour regions in the ROI which share similar intensities to positive samples. This might alleviate some of the pseudo-calcifications generated along the included vertebral edges and soft tissue noise such as bowel gas borders, encouraging selection of features which distinguish these from calcifications. Additionally, it may be valuable include spatial features for selection by the random forest. The absolute x and y coordinates of the input patches are unlikely to be informative, as the position of the lumbar vertebrae in the image is variable. However, the distance of the target pixel from each of the landmark points which constitute the ROI could be valuable. These features would help distinguish vertebral signal from calcification and would be very informative in a multiclass approach.

Some limitations could be addressed with modifications to the ROI prediction model. Larger patches and sample sizes had the largest impacts on performance. The downside of larger patch sizes was exclusion of a border of pixels around the image, which is problematic for calcifications near the extremes of the imaging window. The ROI prediction in Chapter 4 would often predict a ROI that extended beyond the image, indicating a risk for part of the aorta to be excluded. A partial solution for this problem could be quantification of this risk, contributing a measure of uncertainty around an automated score, where imaging data is not available. This could be extended to also quantify the likelihood that the random forest exclusionary zone has influenced

the score.

While there are a number of avenues for improvement of this approach, exploration of new approaches was favoured over further optimisation of the random forest. Many of the best performing segmentation techniques have concentrated on deep learning, in particular convolutional neural networks. These techniques, applied to the segmentation of abdominal aortic calcification, are the focus of Chapter 6.

Chapter 6

U-Net Approach to Segmentation of Calcification

This chapter presents a U-Net based deep learning approach to segmentation of abdominal aortic calcification (AAC) in dual-energy x-ray absorptiometry vertebral fracture assessment (VFA) images. The U-Net, developed by Ronneberger et al. [112], has become a popular deep learning network for segmentation in biomedical images [144, 142, 149, 148, 141] and beyond [135, 136, 137, 138, 139].

This chapter is intended to briefly introduce and then demonstrate the performance of the U-Net and its variations, without demanding familiarity with the literature. Additional detail and justification for experimental design can be found in Section 3.4 and will be referenced throughout the chapter. Performance of the original architecture is first established, architectural variations are then explored to optimise for application to this particular problem. The results are then discussed and compared to the performance of random forest segmentation and previous attempts at AAC segmentation in the literature.

6.1 Data and Resources

The U-Net model was trained and tested using the same dataset of 350 DXA VFA images used in experiments in Chapter 5, and described in Section 4.1. A two class mask of AAC annotation, anterior and posterior calcifications, for each image is used to train and test the segmentation. Expert AAC-24 scores were included for each image for comparison of automated scores. Landmark points predicted using the same shape model discussed in Chapter 4, were used to define the region of interest, including both vertebral and aortic landmarks.

While the original U-Net paper used tiles selected from large images, the area of the VFA images that was of relevance to AAC-24 scoring was well defined and relatively small [112]. As the images in this application were small enough to easily fit in memory, the use of whole images over tiles avoided the redundancy intrinsic to a tile sampling strategy. The region of interest (ROI) defined by vertebral annotations and the point distribution model in Chapter 4 was used to create the training data. Thin plate spline (TPS) warping was used in the same way to transform the ROI of the ground truth mask annotations into a consistent size using nearest neighbour interpolation. The corresponding images from the dataset were warped using the same TPS, except with bilinear interpolation. The lumbar vertebrae were included in the region of interest as this allowed the network to disregard high attenuation areas of the spine that could have been included in the region of interest and caused false positives. This problem was demonstrated by the random forest approach in Chapter 5.

Figure 6.1 demonstrates the resulting processed data that was used for training and evaluation of the U-Net model. Images and masks were a consistent size of 256x128 pixels. The nature of the ROI prediction and warping caused some processed images to include areas outside the original image, these areas were given a value of 0.

A hallmark of the success of the U-Net is the use of significant data augmentation

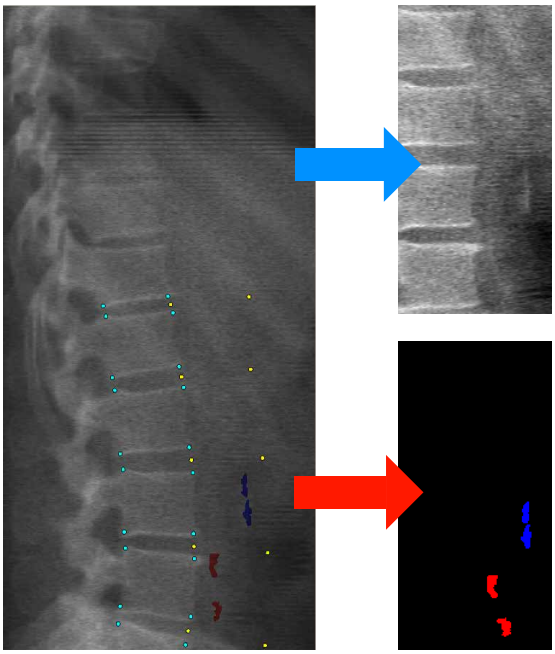

Figure 6.1: Example of thin plate spline warping to produce pairs of training images and masks.

in the input images. This allows multiple training examples to be produced from each mask annotated image in the dataset. The original paper used elastic deformations and linear transformations, including vertical and horizontal flipping to produce augmentations. The consistent window and patient position used when acquiring these images and the AAC scoring technique mean that horizontal and vertical flipping of the images do not create sensible training examples. However, small affine transformations can be used to augment training data, producing further training examples from a single image and reducing overfitting.

Additional augmentation of the image data was produced using TPS. The addition of small amounts of noise to the vertebral point annotations before shape model prediction of the aortic points was used to create non-rigid transformations in the image data. Figure 6.2 shows an example of this process on the original ROI and the resulting transformation with an increasing magnitude of random noise on vertebral annotations.

These newly defined points were then used as the source points for a TPS, with

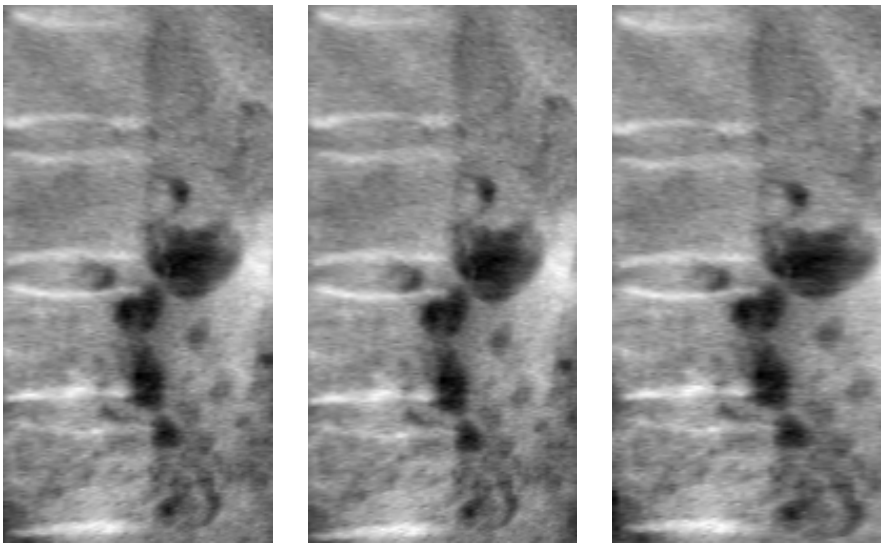

Figure 6.2: Examples of noisy vertebral annotations to generate non-rigid augmentations. The left image has no noise added, the centre has low noise, and the right high noise.

target points at the extremes of the uniform 256x128 pixel image. TPS warps defined by these noisy predictions could also undergo affine transformations, to add additional augmentations. Small amounts of translation, scaling and rotation were added as an online process to the target points, sampling the magnitude of these transformations from a Gaussian distribution.

With a total of 350 images, these were separated into training, validation and test sets. 20% were kept as a test set (70 images), this allowed a true representation of the performance of the final tuned model. The remaining 280 images were then subdivided into training and validation sets using a 4-fold cross-validation, allowing 4 observations of validation performance on $1/4$ of the images, training on the remaining images. Stratified sampling was used to create each image set, taking a proportional number of images from each class of AAC severity: none, AAC-24 score of 0; mild, AAC-24 score 1-2; moderate, AAC-24 score of 3-5; and severe, AAC-24 score of 6-24. This cross validation did not provide independent observations of the model performance,

as there is significant overlap in the training data, but it did give an indication of the variance involved in predictions. Additionally, this ensured that all images were used for validation, reducing the impact of coincidental large values specific to only one fold.

Each image in the dataset also had an expert annotation of AAC-24 score. This allowed comparison of scores produced from automatically generated segmentation masks with those of experts. An additional 628 images from the CAIFOS dataset were also available with expert scores. Though these images did not receive pixel level annotation, they were included to allow scoring.

Implementation

All models were built in python using Tensorflow [188]. Training and testing of models was performed on the University of Manchester's Computational Shared Facility, using their high performance computing nodes with an Nvidia V100 16GB GPU.

6.2 Methods

Since its inception by Ronneburger et al. [112], the U-Net model has been used in a myriad of segmentation tasks. Through skip connections and feature rich upsampling between symmetrical encoder and decoder pathways, the U-Net architecture incorporates spatial information from a range of scales. The intention was that this could overcome difficulties the random forest approach had with eliminating candidate patches based on high level context. The main advantage of a random forest over neural network approach is a much lesser demand for examples to successfully train. Through use of data augmentation, the U-Net attempts to compensate for this shortcoming.

The overall methodology of this chapter was to train the U-Net to recognise and

segment abdominal aortic calcification in the dataset and assess its performance on validation data. Once the performance of the base U-Net had been established, variations of the network from the literature were tested to evaluate their segmentation accuracy in this application. With the best performing network established, a quantitative assessment of segmentation and AAC-24 scoring could be undertaken and compared to human performance on test data. This section explores the details of the methodology and how this comparison was made.

6.2.1 Hyperparameter Optimisation

The first task was to assess the efficacy of the original U-Net architecture on the AAC segmentation problem. Ronneburger et al. developed the U-Net and trained it using stochastic gradient descent with momentum and a batch size of 1, and an original cross-entropy loss function which included weighting for the borders between cells. As this application did not require the same strict borders between segmented objects, the standard cross-entropy loss was used for optimisation. Due to the large mismatch between positive and negative examples, weighting was applied in the calculation of the cross-entropy. This weighting was equal to the ratio between negative and positive pixels in the training data, around 60.

For this application the U-Net was trained on whole images, rather than image patches, as the global image context contains valuable information and the relatively low resolution of the images allowed them to fit within memory. Additionally, during convolution operations, zero padding was applied for any convolution operations that extend outside of the feature matrix, avoiding constantly shrinking feature matrices. Outside of these modifications, the training process remained unchanged, the network architecture is shown in Figure 6.3.

To establish a baseline performance of the U-Net model, a search was performed

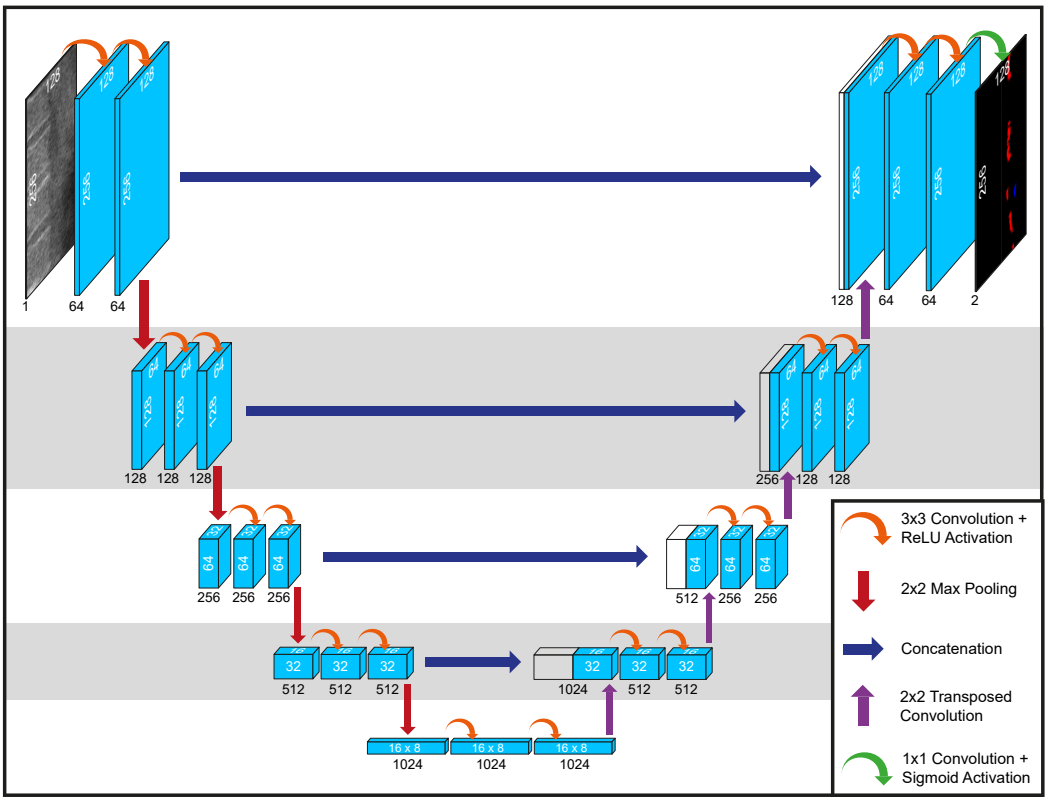

Figure 6.3: The architecture of the original U-Net. Each blue box represents a multi-channel feature matrix, with the feature depth below the box. Each box has the x and y dimensions of the feature matrix, zero padding is used to avoid loss of image size. Adapted from Ronneberger et al. [112].

to find the optimum hyperparameters. The stochastic gradient descent (SGD) with momentum required learning rate, η, and momentum decay constant, α, hyperparameters. Additionally, dropout connections were used in the network, which require a probability hyperparameter unspecified in the paper. Table 6.1 summarises the range of possible values for each hyperparameter. η is the most impactful hyperparameter and is very changeable between applications and models, so a wide range of values was chosen on a logarithmic scale. An α value must be between 0 and 1, but a high value, such as 0.9 is typical for use in deep learning. The range was therefore restricted between 0.5 and 0.999 to allow some exploration while keeping values high. Dropout rate is a probability, with typical values in the literature being 0.2 and 0.5 on

the high end. As dropout was primarily used as a regularisation strategy, the relationship between the loss on training and validation data gave an indication of whether more regularisation was required. Dropout was therefore included in hyperparameter optimisation with a wide range of 0.0 to 0.5, with the possibility for manual adjustment in later experiments.

Table 6.1: Range of possible values for hyperparameter optimisation of the U-Net model

Hyperparameter	**Value Range**
Learning Rate	10^{-6}-10^{-1}
Momentum	0.5-0.999
Dropout Rate	0.0-0.5

The hyperparameter search was performed using Bayesian optimisation to explore the parameter space and identify an optimum configuration. Hyperparameter search strategies, including Bayesian optimisation, are further discussed in Section 3.3.4. Bayesian optimisation was chosen due to limited computational resources. Genetic algorithms have demonstrated superior results for hyperparameter optimisation, with a substantial increase in the number of networks that need to be trained. As they are far more parallelisable than Bayesian optimisation, they can give these results in a practical time-frame with sufficient resources. With the capacity for only a very limited number of simultaneous training networks, Bayesian optimisation was chosen for time efficiency.

Bayesian optimisation was implemented using Gaussian processes with a radial basis function kernel as the probabilistic model. The first three sampling points in the hyperparameter space were chosen at random to establish a starting point for sampling, this allowed some parallelisation as these samples do not depend on the others.

After this initial random sampling, further candidate sampling points were identified by an expected improvement acquisition function, with a total of 20 samples overall. The performance of each hyperparameter combination is compared using the Dice-Sørensen coefficient (DSC) between the validation data and ground truth annotations, which the Bayesian optimisation worked to maximise.

Each sample of the hyperparameter space was evaluated using k-folds cross-validation. With 20% of data held out as test data, the choice of 20% validation data can provide 4 folds for cross-validation. The maximum DSC achieved during training is subject to some noise, especially in the case of a high learning rate, the performance on one epoch could coincidentally overfit to the validation data and give a high metric that is unrepresentative of general performance. By taking the mean DSC across the 4 folds, the impact of this noise was reduced. Each fold was trained on 3 folds and validated on the remaining 1, rotating for 4 runs at each point in the hyperparameter space.

Training was performed with a mini-batch size of 1, updating the network weights after each forward pass, matching the original application. The order of training images and paired ground truth annotated masks was randomised before each epoch, and each pair was subject to random augmentation when sampled. While later experiments were used to calibrate the optimum magnitude of this data augmentation, a base performance of the U-Net was established using a small amount of augmentation. Random affine transformations were applied to images, consisting of a rotation by factor r in radians, a scaling by factor $1+s$, and translation with factors $v*t_x$ and $v*t_y$, where v is the width of the L5 vertebrae. r, s, t_x and t_y are all randomly sampled from a Gaussian distribution with a mean of 0 and a standard deviation of 0.1. Using the L5 width reduces the impact of variations in subject size and image resolution. No shearing or reflection was applied.

Non-rigid augmentation was achieved by applying small perturbations to the vertebral point annotations used to predict the aortic region of interest. Random x and

y translations were sampled and applied to each of the 20 vertebral points, from a Gaussian distribution with $\mu = 0$ and a σ of $0.02v$ of the L5 vertebral width.

At the end of each epoch, the loss function and performance metric were calculated for the validation data. The DSC was calculated over the entire validation set with a threshold of 0.5 indicating a positive prediction for a given pixel. The DSC was calculated with all pixel predictions across all validation images. This gives a measure of performance which avoids the ill-defined case of images with no true positive pixels. Accuracy and IoU metrics were also calculated in a similar manner, to compare with previous work.

The relative trend of how training and validation loss change over time gives valuable insight into how successful the training process has been. When training loss continues to decrease, while validation loss plateaus or decreases, the model is overfitting to the training data. In order to reduce this overfitting, and to reduce the training time, early stopping was used. As there were a large number of experiments to run, with cross-validation, over a wide range of hyperparameters, it was important to save time by abandoning models which failed to converge or which diverged. The validation loss was monitored to keep a record of the best performing epoch. Training was halted on a model once 10 epochs had passed with worse performance than the current best epoch.

To prevent overfitting of the model to the training data, U-Net makes use of data augmentation and dropout layers. These methods of regularisation, in particular data augmentation, allow the U-Net to perform well on unseen data with relatively few training examples. In initial experiments a small amount of data augmentation was used, with affine transformations and non-rigid transformations using TPS warping. Once a baseline performance of the U-Net was established the optimum degree of data augmentation was assessed, to improve prediction performance.

A grid search was performed to test the degree of affine and non-rigid transformations. Each transformation type was divided into three levels: none, low, moderate and high. Factors r, s, t_x and t_y were again randomly sampled from a Gaussian distribution with a mean of 0 and a standard deviation of 0.1 for low, 0.25 for moderate and 0.5 for high. Random translations were also sampled for vertebral annotations, to add non-rigid warping of the images. These translations were sampled for each point from a Gaussian distribution with mean 0 and σ $0.02v$, $0.05v$, and $0.10v$, for low, moderate and high respectively. Performance of the U-Net was tested with combinations of all four levels of each augmentation to ascertain the impact it had on validation segmentation.

6.2.2 Optimisation Algorithms

With performance of the U-Net established, modifications to the training scheme were evaluated. A range of optimisation algorithms have been developed to improve the convergence of deep neural networks, this is further explored in Section 3.3.3. With the original paper making use of SGD with momentum, three additional optimisation algorithms were implemented to train the network: SGD with Nesterov momentum, Adadelta, and adaptive moment optimisation (Adam). These algorithms have all been shown to improve the reliability and rate of convergence, and can be implemented without modification to the network.

Hyperparameter optimisation with Bayesian optimisation was undertaken to train and validate a U-Net model with each optimisation algorithm. The U-Net architecture remained identical to the initial hyperparameter optimisation for each. Each optimisation algorithm had its own hyperparameters requiring adjustment. SGD with Nesterov momentum required no additional parameters, using the same learning rate and momentum ranges. Adadelta required learning rate and a moving average decay factor,

which was optimised within 0.8 and 0.999. The Adam optimiser, in addition to learning rate, required decay factors for both the first and second moment estimates, which were both chosen from a range between 0.8 and 0.999. These ranges were used as hyperparameters to optimise during training. DSC and time to train were used to find a best performing optimisation algorithm, which was then used in subsequent experiments.

6.2.3 Architectural Variations

With an optimised learning strategy, and a validation segmentation performance for the original U-Net architecture, a baseline had been set against which newer U-Net variants could be tested. There have been a number of structural changes made to the U-Net since its inception, with Section 3.4.1 covering the literature on architectural variations of U-Net in more detail.

The advantages of the U-Net architecture lie in its ability to acquire features at various scales, gaining spatial context and using these features to upscale to segmentation predictions. Figure 6.4 lays out a generalised form of the original U-Net depicted in Figure 6.3. Each component of the architecture has seen variation in the literature, a selection of these methods were built and trained to compare performance to the original U-Net.

A fundamental parameter of the U-Net architecture is the number of levels in the encoder and decoder, this heavily influences the spatial features available for training. In addition, the number of filters used at each convolution layer will heavily impact both the size of the model and the performance. A range of values for these parameters were tested in order to assess the effect on performance.

U-Net models were built and trained using a grid search over a range of values for number of levels, n, and initial filter depth f. The convolutional blocks of the first level

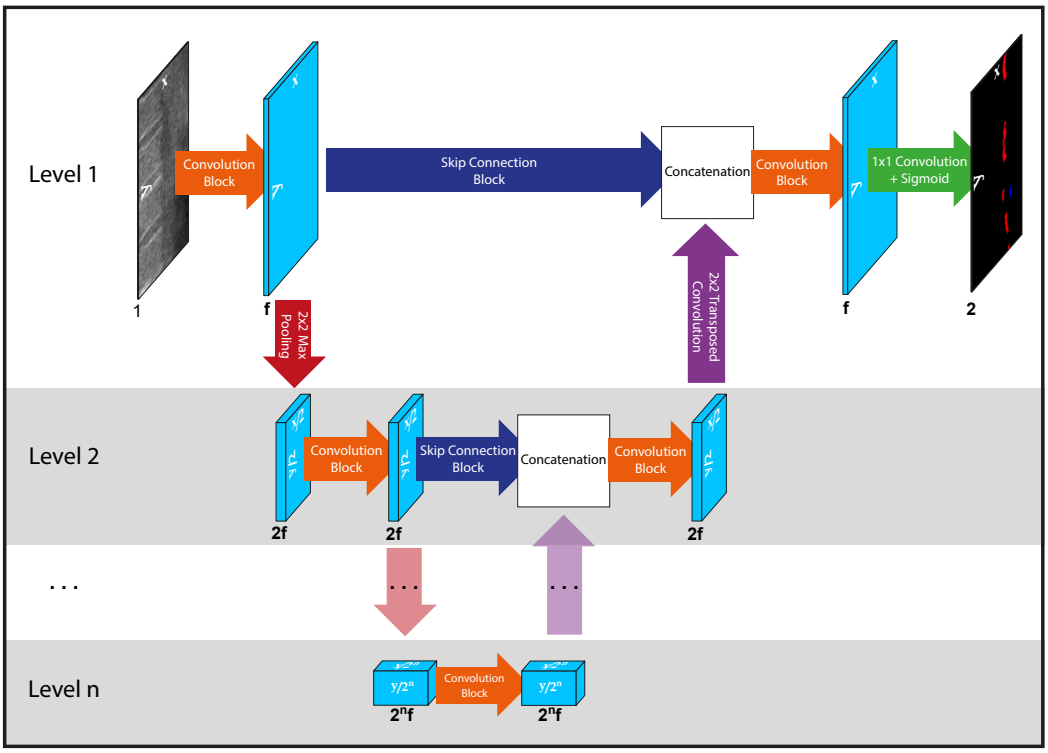

Figure 6.4: The generalised structure of the U-Net architecture. The number of levels and filters can be varied, along with the operations involved in convolution and skip connection operations.

learned a number of filters equal to the initial filter depth, the number of filters in each convolutional block then doubling at each level of the U-Net. All combinations of f and n were tested, with tested values for the two parameters chosen from: $n = 4, 5, 6$ and $f = 2^5, 2^6, 2^7$.

Once well performing values had been established for level and filter depth, additional modifications concentrated on the convolution blocks. Batch normalisation is a technique used in many modern deep learning models to improve training by scaling inputs to layers to have a mean of 0 and a standard deviation of 1. While contributing a small regularisation effect, often eliminating the need for dropout, batch normalisation is more commonly used for its benefits to training speed. This strategy was originally

designed to counter the training challenge of internal covariate shift [189]. It now appears that this is a small part of the reason for the benefits to training, dwarfed by its ability to smooth the optimisation space [190]. Batch normalisation layers were added after each activation function, replacing the use of bias terms in the network, as scaling replaces any constant added to all inputs. Dropout connections were also, dropped. A batch size of 16 was used for these experiments.

The addition of residual connections within the convolution blocks was also tested. Discussed in Section 3.4.1 and illustrated in Figure 6.5, residual blocks allow less impeded propagation of information between layers within the network. Pre-activated residual blocks were implemented, rearranging the order of operations to have the activation function before convolution. A skip connection was added between the result of the convolutions and the input layer, involving a 1x1 convolution to map between the different filter depths and an addition of features to create the output of the block. The residual block has the advantage of adding little to the size of the model while enabling faster learning in deep networks.

Additional connections between feature maps in the convolution block can produce densely connected blocks. This modification, shown in Figure 6.5 was also tested alongside residual blocks. Dense connections also require the use of additional batch normalisation and 1x1 convolution operations before pooling and upsampling operations, in order to manage the large concatenated feature maps between levels. The advantages of dense and residual blocks are in there ability to better propagate information between layers in deep networks. This enables the training of deeper networks, faster. Additional convolution operations were added to each convolution block, to test if more densely connected architectures would have an advantage in these deeper networks. The Residual U-Net and Dense U-Net were tested and compared to previous validation performance, with both 2 and 3 convolution repetitions in each of the convolution blocks. The best performing model, in terms of DSC metric, was selected

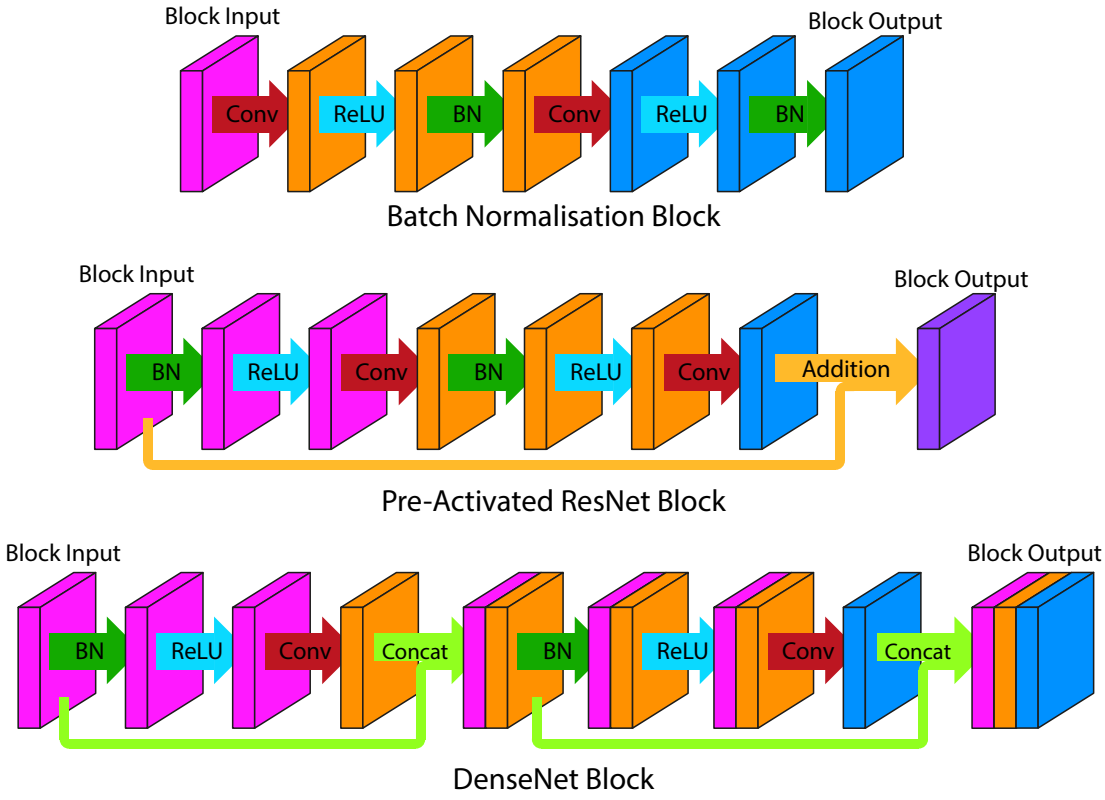

Figure 6.5: The structure of convolution blocks with batch normalisation, residual connections and dense connections. The dense block is shown with three repeated convolution operations.

for evaluation on the test dataset and to produce AAC-24 scores.

6.2.4 U-Net Test Performance

Having optimised the U-Net architecture and training strategy on the validation data, to find the best performing model, the next step was to assess the performance of this model on the held-out test data set. This gave a measure of how much the model optimisation process had overfit it to the validation data, and a true representation of how accurately the network can segment new images.

The model with the best performing 4-fold cross validation on the validation examples was selected from previous experiments. This model was built and trained once

again with randomly sampled training (80%) and validation (20%) datasets from the 280 images in the training-validation dataset. Training was performed using the best performing optimisation algorithm and batch size found in prior experiments. Early stopping was used, ending training if validation loss stopped decreasing for more than 10 consecutive epochs, to avoid overfitting. The weights were then loaded for the epoch with the highest DSC on the validation data, and prediction was performed on the test dataset.

Ground truth manual annotations of the test dataset were used to compare segmentation performance. The metrics produced by this comparison gave the truest impression of the generalisability of the U-Net approach. To assess the generalisation error of the model selection, a nested cross-validation approach was used, demonstrated in Figure 6.6. Experiments up to this point had used a 4-fold cross-validation to establish best performing hyperparameters on the validation set. These results form one fold of the outer loop, with a complete inner loop of the nested cross-validation. With a testing/validation/training split of 20%/20%/60%, an outer 5-fold cross-validation can be used to evaluate predictions across the whole dataset. For each of the 5 70 image test sets, an additional inner cross-validation is performed to evaluate the best performing hyperparameters, while maintaining the same architecture and optimisation algorithm. The extent of any overfitting on the original training set during the parameter optimisation was assessed by comparing the overlap metrics of the held-out test set against those of the other outer folds, with an increase in performance indicating that the choice of architecture is biased to favour the examples in the original training-validation split of the data.

The predicted segmentation masks were then used to generate AAC-24 scores. This was achieved using the multiclass tanh midline estimator described in Section 4.2.2. A midline was estimated using a tanh weighting function to best split the anterior and posterior classes. The relative lengths of each vertebral section of each wall were then

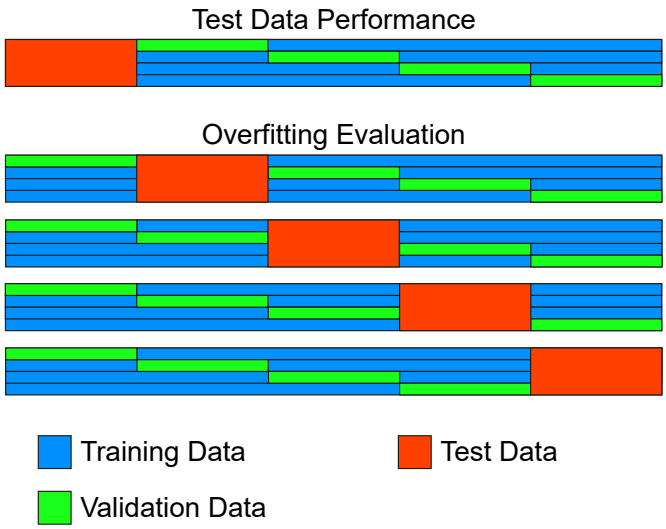

Figure 6.6: 5-fold cross-validation used to assess overfitting of the U-Net architecture. Each horizontal bar represents the full 350 image dataset, in the same class stratified order. The degree of overfitting from model selection can be assessed by comparing the performance of the model on the test set and folds of the training/validation data.

measured and used to automatically calculate AAC-24 scores. With a score generated for each of the 70 images, the correlation between these automated scores and expert scoring was evaluated. Expert level annotations were also available for an additional 627 VFA images which did not have pixel level annotations. These images were used to further evaluate the AAC-24 score prediction for novel images, comparing the generated AAC-24 scores to expert annotation.

6.3 Results and Discussion

The primary aim of the work in this chapter was to establish the performance of the U-Net, a fully convolutional deep neural network designed for semantic segmentation, on the task of segmenting abdominal aortic calcification in DXA VFA images. using regions of interest and data augmentation provided by the methods in Chapter 4, the U-Net was trained and evaluated using validation data. Variations in architecture and

training algorithms were then tested to establish the most promising methodology on the validation data. The best performing technique was then chosen to assess performance on the whole dataset, gaining an understanding of the generalised performance for segmentation and semi-quantitative scoring. In this section the results of these approaches are presented, and a comparison to previous work and the random forest approach from Chapter 5 are discussed.

6.3.1 Hyperparameter Optimisation

350 DXA VFA images, and corresponding annotated masks, were used to train, validate and test U-Net based neural networks. With a 60/20/20 split of data, 280 examples were split into a 4-fold cross-validation for training and validation. The first task was to train and validate the U-Net model in its original form. A hyperparameter tuning process was undertaken using Bayesian optimisation. The first three hyperparameter combinations were chosen at random from the preselected parameter bounds, with the mean Dice-Sørensen coefficient (DSC) over the 4 folds used as the performance metric. Additional sample points were then selected by the Gaussian process probabilistic model and expected improvement acquisition function.

Table 6.2 shows the progression of the hyperparameter tuning process. It is difficult to know how smooth the underlying objective function is, as it is possible that any small perturbation in a hyperparameter could have a large effect. There does appear to be good coverage of the hyperparameter space however, with reasonable consistency between points close to each other. An absence of excessive disagreement for samples close together and the best performing hyperparameters indicates a reasonable smoothness to the function.

The final configuration yielded the highest mean DSC, with 0.510. This is equivalent to a mean IoU score of 0.342, and the calculated accuracy was 0.9962. This was

Table 6.2: Results of the tuning process while training the U-Net network, with the hyperparameters of each tested configuration. The mean Dice-Sørensen coefficient over 4-fold cross-validation is used to compare configurations, with the standard deviation also shown. The number of epochs taken to achieve the maximum DSC was recorded for each fold, the range of these values is shown for each configuration.

Config.	**lr** (10^{-x})	**Momentum**	**Dropout**	**Mean DSC** $(\pm\sigma)$	**Epochs**
1	1.47	0.932	0.488	**0.000** ($\pm$0.0000)	1
2	1.25	0.606	0.334	**0.466** ($\pm$0.0288)	32-36
3	2.68	0.980	0.347	**0.477** ($\pm$0.0242)	66-75
4	4.23	0.500	0.000	**0.006** ($\pm$0.0023)	2-14
5	1.00	0.500	0.000	**0.402** ($\pm$0.0684)	28-33
6	6.00	0.999	0.500	**0.009** ($\pm$0.0020)	3-9
7	2.48	0.500	0.000	**0.388** ($\pm$0.1807)	41-49
8	3.01	0.500	0.500	**0.057** ($\pm$0.0234)	2-28
9	2.89	0.999	0.000	**0.505** ($\pm$0.0266)	53-67
10	1.00	0.500	0.500	**0.408** ($\pm$0.0404)	25-33
11	1.62	0.500	0.000	**0.501** ($\pm$0.0261)	26-37
12	2.46	0.969	0.009	**0.498** ($\pm$0.0143)	43-46
13	1.33	0.500	0.235	**0.498** ($\pm$0.0344)	17-27
14	2.66	0.993	0.102	**0.505** ($\pm$0.0221)	44-59
15	3.51	0.999	0.000	**0.507** ($\pm$0.0174)	64-82
16	3.87	0.999	0.500	**0.492** ($\pm$0.0131)	58-77
17	3.42	0.999	0.386	**0.504** ($\pm$0.0129)	61-76
18	4.61	0.999	0.500	**0.008** ($\pm$0.0017)	1-9
19	3.68	0.999	0.284	**0.486** ($\pm$0.0396)	42-68
20	3.24	0.999	0.000	**0.510** ($\pm$0.0154)	65-78

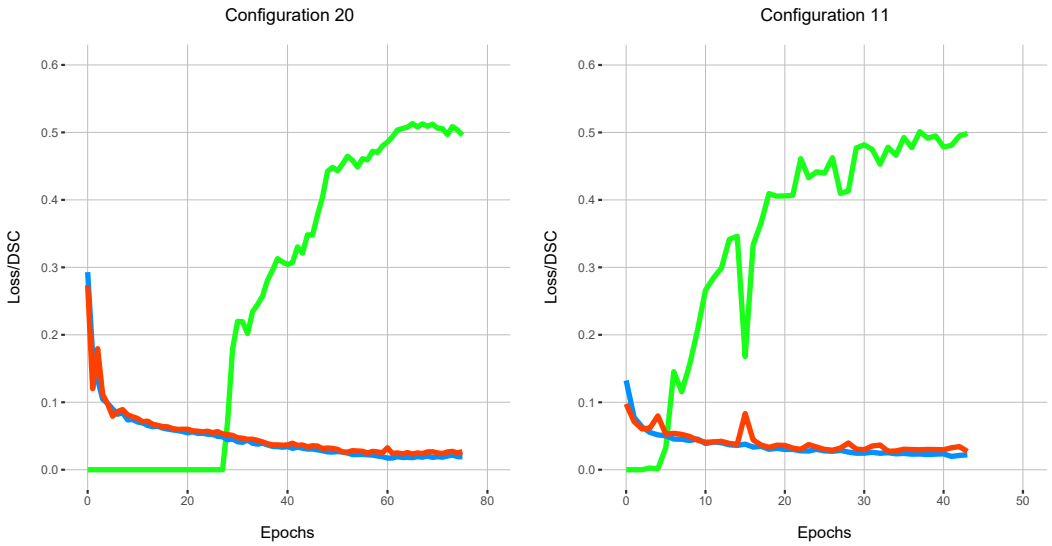

Figure 6.7: Graphs of the loss and performance metrics over training epochs for hyperparameter configurations 11 and 20. Pixel-wise cross-entropy was used to calculate training loss (blue) and validation loss (red), alongside the DSC performance metric (green) calculated across all validation examples at the end of each epoch.

an encouragingly high DSC compared to the 0.383 achieved by the random forest approach in Chapter 5. The actual performance on the test dataset was likely to be lower than the performance here, as the hyperparameter optimisation was liable to overfit to the validation data. The extent of this effect was explored once the most promising architecture and training methods had been identified.

The loss and DSC at each epoch for the first fold of configuration 20 are shown in Figure 6.7. Only the first fold is shown for clarity, but this is representative of the shape of the other folds. There is a notable delay before any change is seen in the DSC, with the loss having to decrease to a certain level before DSC increases. The best performance being achieved by the final hyperparameter configuration indicates that additional performance gains are likely with further optimisation. The overall trends of the hyperparameter tuning gave valuable information for further experimentation.

Given the variance and proximity of scores, it is likely that the actual performance of the best scoring hyperparameter configurations is likely to be very similar. From

Table 6.2 it appears many of the best performing hyperparameters involved moderate learning rates with very high momentum. This can be seen in configurations 9, 14, 15, 17 and 20; with poor performance from configurations which strayed into lower learning rates, regardless of momentum. This is apparent in configurations 4, 6 and 18, though is far from having completely explored this area of the hyperspace.

The other group of hyperparameters that appear to perform well are those with very high learning rates and low momentum, such as 2, 11 and 13. These configurations achieved comparable DSC scores, and did so in fewer epochs. The length of each epoch was very consistent across all configurations, between 340-374 seconds with a mean of 353. The difference in the number of epochs lead to training times in the region of 2.5 hours per fold compared to 6 hours for the lower learning rate configurations. The increased variance exhibited by these configurations indicates that the high learning rate leads to instability in the convergence. The high learning rate is likely to result in oscillations around minima in the loss function, which could allow noisy sampling of the DSC function, leading to the high but inconsistent scores. Figure 6.7 explores the loss and metric curves for the training process for configuration 11, indicating that this was indeed the case.

Based on the performance of the configurations in Table 6.2 it appeared that the influence of the dropout probability on network performance was minimal. Extremes of this parameter were used extensively during tuning, with some exploration of moderate values. While it was possible that there were regions of the hyperparameter space where dropout had a large impact, there appears to be little correlation with the other hyperparameters; with paired examples such as configurations 5 and 10. As a regularisation technique, it is possible that dropout is not necessary to improve performance, or may be dominated by the regularising effect of early stopping.

Figure 6.8 shows examples of how the segmentation predictions on validation data

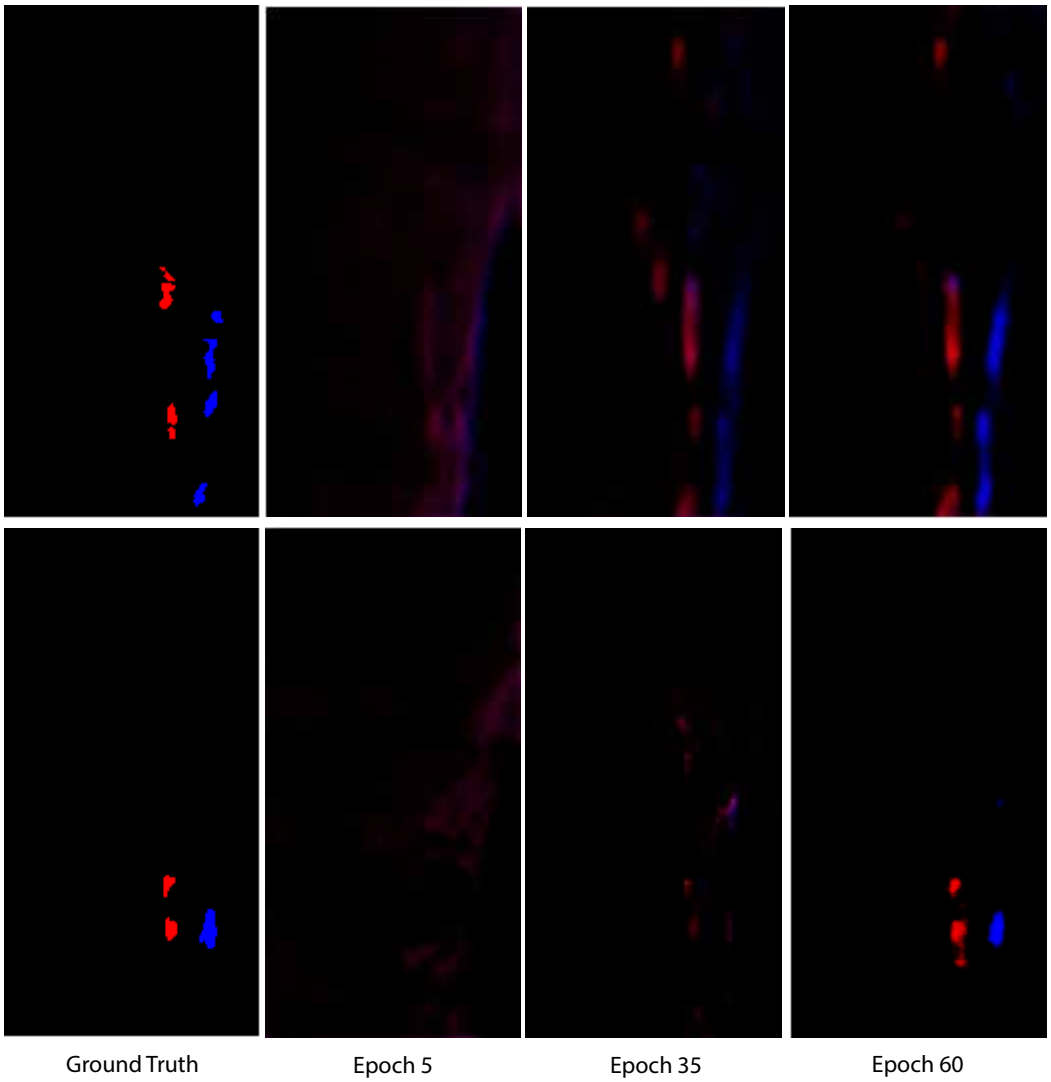

Figure 6.8: Examples of two validation predictions at different stages of the training process for configuration 20.

change over the training process. It appears from these predictions that the cross-entropy loss function encourages low response over the entire image until the optimisation finds informative features for the calcification, at which point the DSC sharply increases. Qualitatively, these images demonstrated that sensible segmentations were being generated, and the network was able to identify informative features. Given the slightly increased performance of configuration 20, and the more stable DSC curve, this was chosen as the hyperparameter configuration for subsequent experiments. A DSC metric of 0.510 was established as the baseline performance of the unmodified U-Net.

For these experiments, image-mask pairs were sampled with small non-rigid and affine transformations randomly applied. The next experiments concentrated on optimising the amount of data augmentation that would best allow the U-Net to learn from the limited training data while generalising sufficiently to the validation data. A grid search was performed with combinations of none, low, moderate and high amounts of augmentation, training the U-Net with the optimisation parameters defined in the hyperparameter search.

Table 6.3: Segmentation performance for U-Net with varying intensity of image augmentation. The mean DSC is given across the 4-fold cross-validation, with the standard deviation in brackets.

		Non-rigid Warping			
		None	**Low**	**Mod.**	**High**
Affine	**None**	**0.367** (0.0466)	**0.507** (0.0192)	**0.510** (0.0166)	**0.491** (0.0111)
	Low	**0.373** (0.0529)	**0.510** (0.0154)	**0.513** (0.0203)	**0.488** (0.0254)
	Mod.	**0.333** (0.0368)	**0.503** (0.0256)	**0.505** (0.0253)	**0.310** (0.0389)
	High	**0.226** (0.0454)	**0.240** (0.0422)	**0.217** (0.0533)	**0.188** (0.0829)

Table 6.3 demonstrates the mean DSC achieved across 4-fold cross-validation with each combination of data augmentation. Due to the noisy nature of the observations, it was hard to ascertain any strong evidence of an optimum degree of augmentation. It appeared that the extent of affine transformations had only a small impact on performance outside of the high setting. Moderate non-rigid image warping did seem to perform slightly ahead of low, but this effect is minor and is likely due to noisy observations of the performance. These results did however, demonstrate a clear improvement to segmentation accuracy with the use of noisy landmark generated TPS warps for image augmentation. With no augmentation, the training loss decreased far quicker than validation loss, indicating overfitting.

The use of noisy annotations of bony landmarks in order to generate non-rigid transformations of the images is a novel aspect of this work, and appears to provide improvements to segmentation. Generated images are convincing in appearance, and provided the extent of these adjustments is not excessive, can be used in a similar manner to the original U-Net application [112] to predict segmentations with a relatively small dataset. With little justification for increasing the magnitude of data augmentation, and avoiding excessive movement of the region of interest which could exclude areas of calcification, further experiments were performed with low affine and non-rigid warping.

6.3.2 Optimisation Algorithms

With a successfully trained U-Net model, the same architecture was used to establish the performance of alternative optimisation algorithms. Hyperparameter optimisation was performed using Bayesian optimisation with three optimisation algorithms: SGD with Nesterov momentum, Adadelta and Adam. Table 6.4 shows the best performing hyperparameter combination for each optimisation algorithm, with the previous

experimental performance included for comparison.

Table 6.4: Segmentation performance of U-Net hyperparameter tuning using different optimisation algorithms. The mean of a 4-fold cross-validation is used for maximum DSC and number of epochs to reach the maximum.

Optimisation Parameters	Mean DSC ($\pm\sigma$)	Mean Epochs
SGD + Momentum ($lr = 10^{-3.24}$, $momentum = 0.999$)	**0.510** (0.0154)	71.5
SGD + Nesterov ($lr = 10^{-3.63}$, $momentum = 0.999$)	**0.508** (0.0171)	74.25
Adadelta ($lr = 10^{-2.31}$, $decay = 0.8$)	**0.477** (0.0281)	92.5
Adam ($lr = 10^{-3.65}$, $\beta_1 = 0.856$, $\beta_2 = 0.999$)	**0.516** (0.0192)	64.5

The best DSC performance achieved on the validation set was 0.516 using the Adam optimiser. This was achieved using a learning rate of 0.00022, and first and second moment decay constants of 0.856 and 0.999. Changing to Nesterov momentum appeared to have little impact on the performance of the optimiser, as the achieved segmentation performance and time taken to train were very similar. The Adadelta algorithm segmentation performance was consistently below that of the other algorithms. The choice of learning rate and decay parameters did not have a large impact, with many acceptable configurations achieving a similar DSC. The total epochs to reach maximum performance, and the number of epochs before DSC started to increase were both much higher than previous experiments, this may indicate that early update gradients were large and the decaying gradient sum kept the learning rates too low to converge to a competitive segmentation.

The U-Net trained using the Adam optimisation algorithm appeared to have outperformed the original SGD with momentum. Similar to Adadelta, Adam demonstrated increased flexibility to choice of hyperparameters, with many choices for learning rate and β which yielded DSC results in the region of 0.51. It seemed likely that additional hyperparameter tuning with SGD + Momentum could have also achieved this level of performance, as the final hyperparameter configuration was the most successful. With the increased stability of Adam to the choice of initial hyperparameters, it was likely that the improved performance was due to the optimisation spending fewer samplings testing poor performing configurations. Adam had a strong effect on the time taken to train the network, reducing the mean number of epochs to reach maximum performance by 7, saving an average of 24 minutes per fold. For these reasons Adam optimisation with these parameters was implemented for the experiments on U-Net architectural variations.

6.3.3 Architecture Variations

To evaluate the feature depth of the U-Net architecture and its ability to capture the complexity of the training data, experiments with various U-Net models were run. The number of levels and the filter depth of the convolution operations were varied and tested using the DSC performance metric, to establish performance. Table 6.5 demonstrates the results of these variations.

Table 6.5: Segmentation performance for U-Net models with varied level and feature depths. The mean DSC and standard deviation are shown for a 4-fold cross validation.

		Initial Feature Depth		
		32	**64**	**128**
No. Levels	**4**	**0.477** (0.0128)	**0.513** (0.0211)	**0.511** (0.0305)
	5	**0.494** (0.0133)	**0.516** (0.0192)	**0.518** (0.0206)
	6	**0.493** (0.0151)	**0.512** (0.0213)	**0.509** (0.0251)

The best performance was achieved by a 5-level network with an initial feature depth of 128, at a DSC of 0.518. However, it should be noted that the performance of many of the models was very similar provided the initial feature depth was at least 64. Deeper networks can achieve better theoretical performance, but may train slowly or fail to converge due to the vanishing gradient problem. With a relatively small number of convolutional operations, it is unlikely that additional levels were failing to improve performance due to vanishing gradients. It is likely that the 4-level network was sufficiently complex to encode the features of the training data, and that any differences in DSC were due to noisy observations of the same performance.

Importantly, the mean time to train the 4-level 64-depth network was 244 minutes, compared to the 374 minutes of the 6-level 64-depth network. With very little evidence of a benefit to segmentation performance, it appears that a shallower network is a suitable trade-off for training time. Subsequent experiments were performed using a 4-level U-Net architecture with an initial feature depth of 64. The smaller memory footprint of the model is also an advantage, allowing more flexibility on the hardware required to test the model on new images, an important advantage for any strategy to implement the model clinically. Smaller models also allow the use of larger batch sizes during training while still fitting in GPU memory.

U-Net configurations involving changes to the convolution blocks were also tested. The addition of batch normalisation, as well as additional skip connections within the U-Net architecture were compared for segmentation performance and training time, this is shown in Table 6.6. All of these variations of the convolution block were implemented within a 4-level U-Net with the Adam optimiser and a mini-batch size of 16.

Table 6.6: Segmentation performance of U-Net with variations on the convolutional block. Batch normalisation is added, with both 2 and 3 convolutions in each block, as well as residually and densely connected variations. The mean of a 4-fold cross-validation is used for maximum DSC and number of epochs to reach the maximum.

Block Type	No. Conv.	Mean DSC ($\pm\sigma$)	Mean Epochs
Batch Normalisation	2	**0.512** (0.0161)	68.0
Batch Normalisation	3	**0.508** (0.0219)	75.5
Residual	2	**0.527** (0.0163)	66.25
Residual	3	**0.530** (0.0194)	71.5
Dense	2	**0.538** (0.0221)	69.0
Dense	3	**0.542** (0.0217)	74.25

The best performing convolutional block was a densely connected concatenated block with 3 convolutional operations to each. With a mean DSC of 0.542 on the validation dataset, this is a significant jump in performance over models trained without residual connections. This increase in performance was seen in all residual and dense connected networks, with the differences between each of these models being much smaller. Evidence of an advantage to adding an extra convolutional layer in each block is weak, but it is clear that additional connections within the convolutional block allow the U-Net to better identify and optimise features which improve prediction.

The addition of batch normalisation appeared to reduce training time, reducing the number of epochs to reach the best performing. The use of mini-batches reduced the training time of each epoch to a mean of 167 seconds, so overall training time for the best performing model was 3.5 hours compared to the 6 hours required to train the best performing unmodified U-Net.

Having evaluated a range of model variations and tuned the parameters of the U-Net, the final architecture of the U-Net was decided for assessing the model on the test dataset. With encouraging evidence that densely connected networks improved model performance, and that the substitution of dropout for batch normalisation improved the rate of convergence while still providing sufficient regularisation, these modifications were made to the final U-Net model. Based on the performance of the residual and dense U-Nets, it appeared that there was not strong evidence for the addition of a third convolutional layer to the convolutional blocks, alongside the cost of additional training time and memory demands, it was decided to keep the number of convolutions at 2. The same Adam optimisation algorithm and data augmentation strategies were implemented, with a mini-batch size of 16. This final network configuration was tested on the holdout set.

6.3.4 U-Net Test Performance

The experimentally validated U-Net model was trained on a stratified randomly selected 75% of the training data, and validated on the remaining 25%. This U-Net had 4 levels, an initial feature depth of 64, 2 convolutional operations in each block, and dense connections concatenating feature maps between convolution operations. Training was performed using the Adam optimiser with a learning rate of 0.00022, and β_1 and β_2 constants of 0.856 and 0.999 respectively. The highest DSC achieved on the validation data was 0.546. The weights of the epoch which achieved this score were

loaded, and testing was performed on the 70 image held-out test set.

Figure 6.9 shows example segmentation masks from the U-Net model. These examples show a well performing segmentation along with two predictions containing errors. Some of the predictions still identified borders around bowel gas as areas of calcification, an example of this is given in the middle row. This was the main source of false positive responses, and contributed to overestimation of scores. The bottom row demonstrates an example of a segmentation which identified the correct calcification, but misidentified the wall to which it belongs.

The segmentation performance of these predictions was assessed quantitatively using overlap measures, by comparing the segmentation masks to those of ground truth annotations. A DSC calculated across the pixels of all predictions gave a score of 0.532. This is equivalent to an IoU score of 0.362, and an accuracy of 0.997. Compared to many segmentation challenges, this was not a particularly high level of agreement between prediction and manual annotation. However, for a task subject to a great degree of noise, image variation and rater subjectivity, this represents definite progress and improvement. Overall, the U-Net segmentation underestimated areas of calcification, leading to false negative pixels on the borders of areas of high response, though the impact of this on AAC scoring would be later examined. This effect, alongside the misidentification of bowel gas and image noise as calcification, lead to the relatively low DSC scores.

Petersen et al. [178] performed automated segmentation of AAC in lateral radiograph images of the spine, where they achieved an IoU of 0.28. In terms of overlap of segmentations, the U-Net outperformed this random forest and shape prior approach, which was especially notable in an image format which is lower resolution and subject to increased noise. Additionally, the predictions in this work included images with no calcification, which could substantially decrease overlap metrics with additional false positive predictions. This score also represented a significant improvement over the

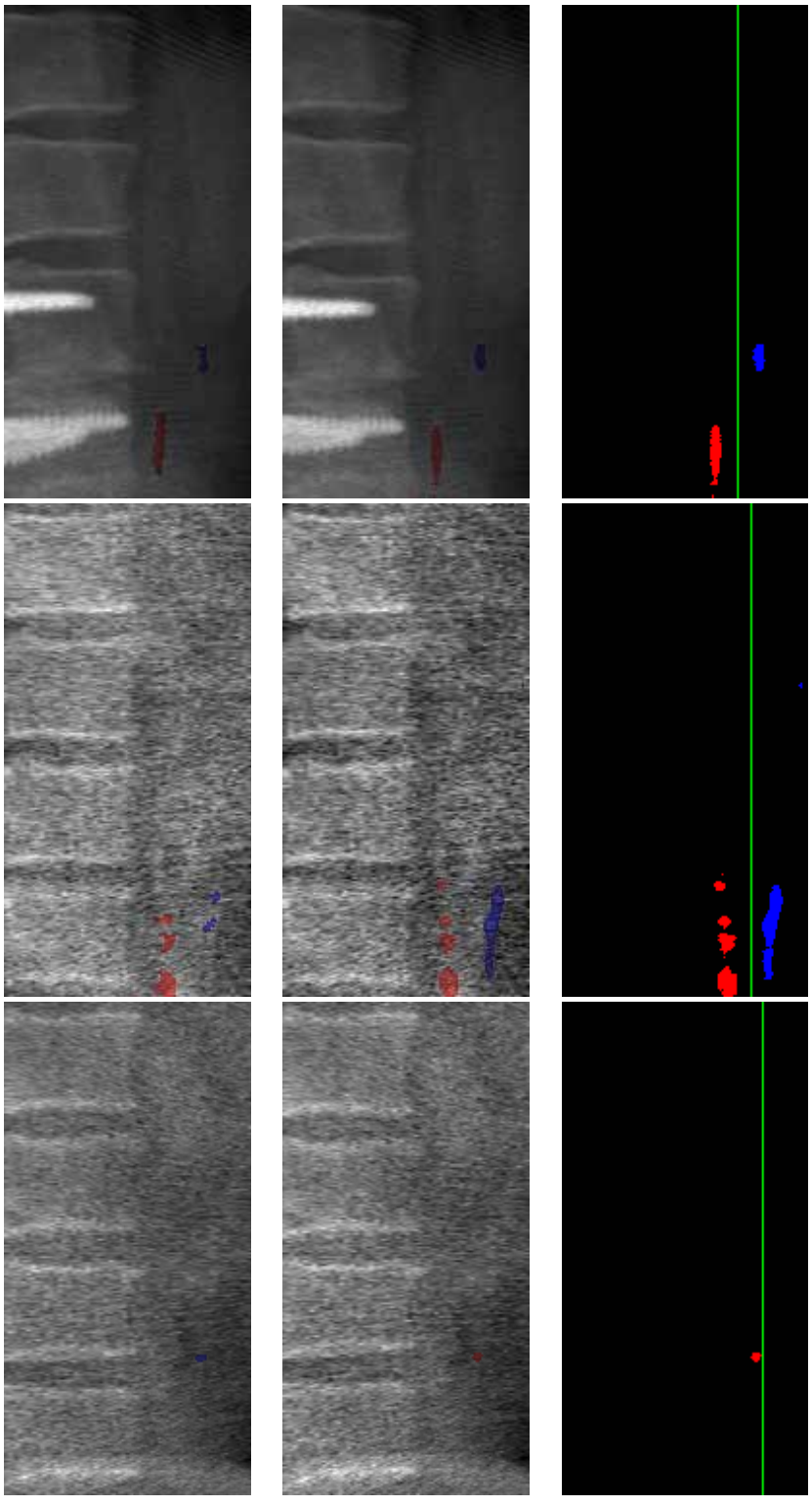

Figure 6.9: Examples of the final U-Net test performance. The ground truth annotations and predicted segmentations are shown overlaid on the warped images. The predicted midline for scoring is also shown.

random forest approach in Chapter 5. With far fewer false positive regions, the U-Net achieved a result that improved on both qualitative and quantitative grounds. Petersen et al. also measured the inter-rater performance for pixel-wise annotation, achieving an IoU of 0.51 (DSC 0.68) [178]. Repeat annotation of the data set in this work gave an intra-rater reliability with a DSC of 0.784. The segmentation performance of the U-Net still falls short of these measures, indicating that there is still information in the images that can be used by human interpreters but is not captured by the model.

Comparing the DSC of 0.546 achieved during validation and the 0.532 during testing, it appears that the process of hyperparameter optimisation and model selection has caused overfitting in the non-test data. With so many samples of the validation performance taken during hyperparameter optimisation, it was inevitable that there would be some selection of models and hyperparameters that caused improvement specific to the training and validation set. To truly assess this discrepancy, the holdout test set results were compared to predictions on cross-validated experiments on the training and validation data. U-Net models with the same configuration were trained and tested on different folds of the data. The results for this comparison are shown in Table 6.7.

Table 6.7: Overlap metrics comparing the U-Net segmentation to ground truth annotations. DSC and IoU are presented for the test data, which gives the true performance of the segmentation. The performance of using the training and validation data for predictions is also presented to examine overfitting.

Prediction Data	**DSC**	**IoU**
Test Set	0.532	0.362
Train/Val. Set	0.545	0.375
All Images	0.542	0.371

The results of these experiments made it clear that there was a substantial amount

of overfitting produced by the model selection process. Using the segmentations produced on the training-validation data to perform scoring of AAC would not give a realistic impression of the performance of the U-Net to predict AAC-24 scores. Tanh midline prediction, developed in Chapter 4 was used on the test set predictions to score calcification using the AAC-24 score. This prevents direct comparison of the relationship between random forest and U-Net approaches to scoring on the same images, but their overall correlation to expert annotation can be used to compare these methods.

The DSC is heavily influenced by small differences in segmentation. Examples such as the bottom row of Figure 6.9, in which calcifications were identified as the wrong class, have large impacts on DSC, scoring 0 for this image. This impact does not affect the score generated however, as it still successfully identifies the image as having an AAC-24 score of 1, with the estimated midline simply on the other side of the calcification. To assess the correlation between expert predictions and scores generated automatically from images, the intra-class correlation coefficient (ICC) was calculated. Figure 6.10 visualises this correlation for the test set.

An ICC of 0.869 was achieved, with a lower bound of 0.798, indicating good correlation with expert annotations. This is a relatively small sample, 70 images, but the correlation is encouraging. Most of the images in this sample, and indeed the whole dataset, have only mild AAC, which the model appears to classify well. From the few high scoring images, it seems that there is a tendency to underestimate the scores, in some cases quite substantially. Additionally, the number of expert annotations containing no calcification, which was correctly identified as such was very low. This is problematic, as this represents an informative difference in cardiovascular risk. The number of these examples is low however, the sampling of the test set could have a large impact on this comparison. To further evaluate the correlation, 627 additional images which did not have pixel annotation were used for scoring, and compared to expert scores. The results of this comparison are shown in Figure 6.11.

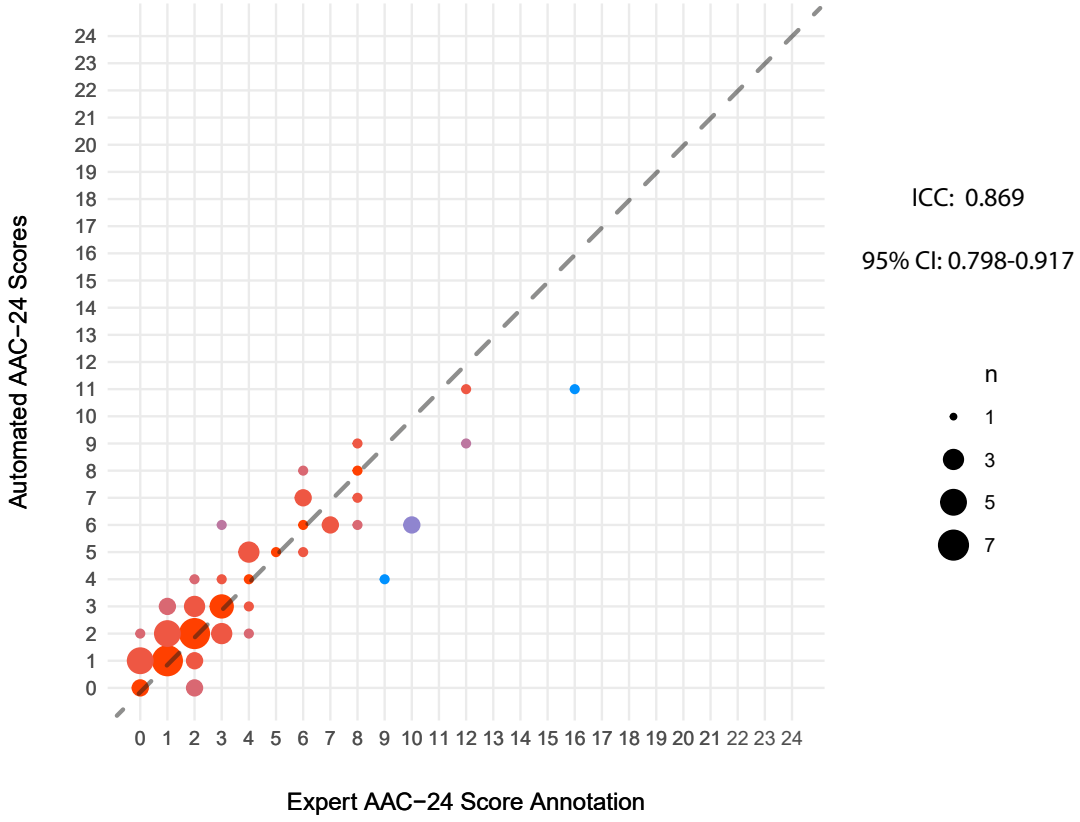

Figure 6.10: Intraclass correlation between U-Net derived AAC-24 scores and those from expert annotation for 70 images held out for testing the final U-Net configuration.

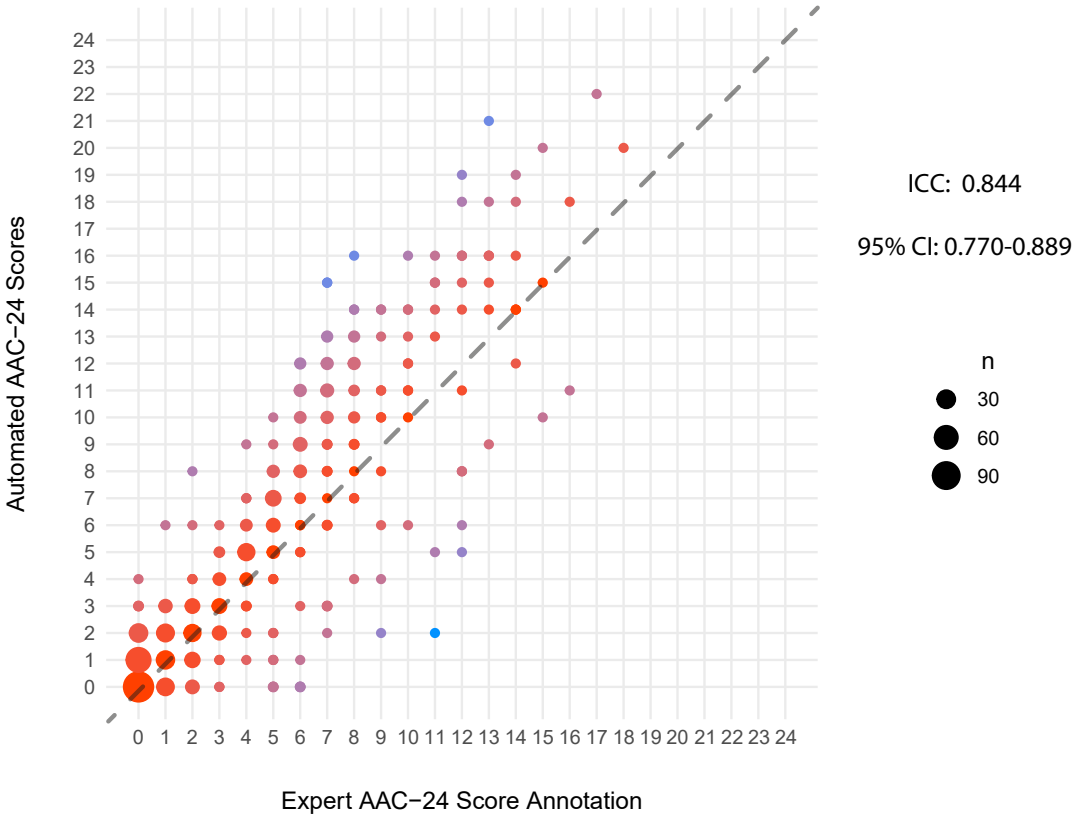

Figure 6.11: Intraclass correlation between U-Net derived AAC-24 scores and those from expert annotation for 697 images which were not included in the training and validation sets.

This comparison gave a much better impression of how well the U-Net predictions correlated with expert scoring. With a much larger sample size, the ICC fell slightly to 0.844, with a lower bound of 0.770. This still indicates good correlation, but indicates that the 70 image sample was favourably biased. This chart indicates that there isn't a heavy bias toward under or overestimation of scores, though there are still a number of large outliers for high scoring images, even some severe images misclassified as having no calcification. With a larger dataset it is clear that there are a substantial number of images with no calcification, and a more encouraging proportion which have been identified correctly. Though given the importance of this distinction it is still a troubling error rate.

This correlation still has a large discrepancy with the excellent correlation seen between scores generated by repeat manual annotation of calcification by the author, which achieved ICC 0.939, and those manual annotations against expert scoring, which achieved ICC 0.920. This is further evidence that the majority of the error in scoring is due to the segmentation performance of the U-Net, and that there is further improvement possible with a better segmentation approach. Though there is indication that reliably scoring severe images may require an improvement in midline prediction, as the increased tortuosity may require the use of fitting a curved line to the segmentation mask to better score these images.

Work by Elmasri et al. is the main work that has been done in this problem to date, classifying AAC severity in VFA images [179]. Their approach is detailed in Section 3.5, but uses active appearance models to classify images as mild, moderate and severe. Though it is not directly comparable, as in this work AAC-24 scores were calculated from segmentation masks, the calculated AAC-24 scores were then used to classify images into severity classes using the same criteria as their work: mild (0-4), moderate (5-12), severe (13-24). The results of this classification were class accuracies of 91.2%, 83.9% and 81.8% for mild, moderate and severe respectively. Compared to

the performance of 93.1%, 90.4% and 95.2% in the work of Elmasri et al., these results are lacking [179]. However, this comparison is between a classifier optimised for this exact purpose which did not use images without calcification. Combined with the ICC for AAC-24 scores, there was still substantial evidence of good correlation for the first attempt at direct AAC-24 scoring in VFA images.

Overall the scoring capabilities of the U-Net are impressive, with an improvement over random forest segmentations on this and other datasets.

6.4 Conclusions

This chapter has presented the methods and results for automatically segmenting abdominal aortic calcification in vertebral fracture assessment images using a U-Net based deep learning model. Achieving a DSC overlap metric of 0.532, the overall agreement of automated segmentation with manual annotation is moderate, exceeding performance achieved in previous literature in radiograph images. Automated scoring of the images achieved an ICC of 0.844 with expert annotation, showing good correlation, but with a sizeable gap in performance compared to manual scoring.

The multiclass predictions allowed the use of more flexible tanh predicted midlines, to improve scoring based compared to random forest. There were still a large number of false positives in the images, leading to low DSC scores. Similarly to the random forest segmentations, many of these were from areas of high intensity due to bowel gas. These borders between gas and soft tissue create convincing pseudo-calcifications which are sometimes difficult for readers to distinguish. The dataset appears to not have enough examples of these for the U-Net to learn, or the model is not appropriately complex to capture these features. It may be possible to improve this performance issue by changing the data annotation methods. By annotating sources of pseudo-calcification, such as bowel gas or skeletal regions, as separate classes it may

be possible to overtly indicate areas which are not AAC, and improve performance. Unlike the random forest, the inclusion of the vertebrae in the training images effectively eliminated false responses on the vertebral bodies.

Changes to the data made available to the model, and the amount of data are likely to be the largest sources of improvement. Even with substantial effort to optimise the model, the increases to DSC were modest. Hyperparameter optimisation with Bayesian optimisation appeared to yield sensible training parameters and allow the model to learn. In future work it may be sensible to use the variance of the k-fold performance in addition to the mean, in order to better predict hyperparameters. Continued work to optimise the model will likely only yield small increases. A potential source of future model improvement is in the more efficient acquisition of more data. Pre-trained models have been used to form the encoding path of U-Net models previously, such as the VGG-11 model [191]. Using models pre-trained on large datasets of general images, or medical images, would allow the U-Net fine-tuning to discover combinations of already identified features which are valuable in the problem. In a similar manner, the acquisition of informative features from the more easily available image level annotations could be incorporated into the segmentation algorithm, allowing collaborative learning of classification and segmentation to provide some semi-supervised rewards to the U-Net [192].

With the segmentation performance and automated scoring of the U-Net based model established, and shown to compare favourably to the literature, the final chapter of this work briefly summarises the main conclusions. Additionally, some avenues for future work in this area are presented.

Chapter 7

Conclusions and Future Work

The main contribution of this thesis work is the development of a system for automated AAC-24 scoring of abdominal aortic calcification in dual-energy x-ray absorptiometry vertebral fracture assessment images. Previous work in measuring AAC has concentrated on radiograph images. The opportunity to gain this valuable clinical information from bone density scans is an important problem, which could yield significant improvements for the diagnosis and management of cardiovascular disease. This work has presented the first automated scoring of AAC in DXA images, using the semi-quantitative AAC-24 clinical score.

A random forest regression algorithm was trained on patches sampled from the abdominal aortic region, to predict probabilities that the patch was centred on AAC. This approach did manage some level of discrimination between background and calcified pixels, with a DSC of 0.367, but did not manage to achieve the segmentation performance of previous approaches in radiograph images. Automated scoring of images with this method also showed poor correlation with expert annotation, achieving an ICC of 0.554.

A U-Net based deep learning segmentation algorithm was trained on annotations of AAC within pre-defined ROIs. This is the first attempt at a deep learning approach to

AAC segmentation in the literature, including radiograph images. Additionally, a novel noisy point annotation of bony landmarks was used to generate non-rigid image augmentation using a point distribution model and thin-plate spline warping. Performance exceeded the overlap metrics of previous attempts at this segmentation, achieving a DSC of 0.532 and highlighting another problem for which deep learning can yield improvements. With early work on random forest and U-Net segmentation presented to the medical image community [193], a full exploration of the automated AAC-24 scoring on clinical predictive models will follow.

AAC-24 scores produced by the predicted segmentations had good correlation with expert scores, achieving an ICC of 0.844, indicating the potential for automating this process clinically. Though the correlation was still not as high as manual annotation of calcification, or the inter-rater correlation for experts, the speed at which these predictions can be made, and the additional information that segmentation masks provide, could outweigh these shortcomings.

To calculate the AAC-24 scores, a simple weighted function to fit midline estimates of the aorta was combined with a statistical shape model which predicted the location of the abdominal aorta. The scores generated with this function showed excellent correlation with expert scores, and combined with an automated segmentation algorithm, have produced the first fully automated AAC-24 scores in DXA VFA images. The demonstrated robustness of the scoring system allows for flexibility in the segmentation algorithm of choice, so that future work to improve segmentation methods can be incorporated without adjusting the scoring system. Future improvement to this scoring algorithm appears possible, with the use of a similar loss function to optimise the position of a curve between classes, penalising extreme curvatures.

This work has concentrated on the reproduction of human performance in segmentation and scoring of AAC. However, the information which these methods could

provide is not subject to some of the same limitations. The semi-quantitative AAC-24 score was designed to enable simpler and faster assessment of AAC. Without any modification, the system developed in this work can output a continuous measure of calcification. A score generated which gave the total percentage of the aortic wall identified as AAC could allow future work to assess the exact risk associated with increasing calcification. Additionally, with a continuous measure, the changes in AAC over time are more easily inferred, allowing better assessment of the impact of treatment. Additionally, already validated scores of AAC such as the MACD index [50] could be built into this system, allowing them to be calculated without clinician input.

There is still substantial room for improvement of the segmentation algorithm. The main challenge to the generalisability of this work lies in relying on pixel-wise annotation from the author, instead of a domain expert. This relates to the overall problem of the high cost of acquiring expert training data in this and similar problems. This could be partially mitigated in future work with the use of approximate or 'messy' annotations. Work has been done to incorporate annotations in the form of bounding boxes and circling of areas of interest. As these can be acquired with far less work from a clinician, these features can be incorporated into a prediction system to improve performance. Another problem is that scoring of AAC-24 is still subject to a high degree of subjectivity on the part of human annotators. Datasets containing both abdominal CT and VFA images could allow direct annotation and comparison of AAC identified on VFA and the gold standard, CT.

With a sufficiently large dataset of images and patient CVD risk factors, it would also be possible to directly predict patient risk and validate it with patient outcomes in order to get a predictive clinical model based on the true ground truth for the problem. Dynamic models could be used to identify those individuals for which screening might be the most informative. Additionally, increasing AAC-24 score is associated with risk in other diseases. The automated analysis of AAC can inform treatment decisions

for osteoporosis and in chronic kidney disease. It is also notable that AAC is a risk factor for the rupture of abdominal aortic aneurysm [194]. There is substantial work in identifying these aneurysms automatically in images [195]. In the future, it may be beneficial to automate the detection of both aneurysms and calcification to better catch and stratify risk in these diseases.

The methods to automatically analyse abdominal aortic calcification in vertebral fracture assessment have yielded interesting results, and helped to identify several avenues for future research to use this work to improve clinical outcomes.

Bibliography

[1] World Health Organisation, *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016*. Geneva, 2018.

[2] L. A. Solberg and J. P. Strong, "Risk factors and atherosclerotic lesions. A review of autopsy studies.," *Arteriosclerosis: An Official Journal of the American Heart Association, Inc.* **3** no. 3, (May, 1983) 187–198.

[3] R. van Dijk, R. Virmani, J. von der Thüsen, A. Schaapherder, and J. Lindeman, "The natural history of aortic atherosclerosis: A systematic histopathological evaluation of the peri-renal region," *Atherosclerosis* **210** no. 1, (May, 2010) 100–106.

[4] K. Sakakura, M. Nakano, F. Otsuka, E. Ladich, F. D. Kolodgie, and R. Virmani, "Pathophysiology of Atherosclerosis Plaque Progression," *Heart, Lung and Circulation* **22** no. 6, (Jun, 2013) 399–411.

[5] J. Hsia, J. G. MacFadyen, J. Monyak, and P. M. Ridker, "Cardiovascular Event Reduction and Adverse Events Among Subjects Attaining Low-Density Lipoprotein Cholesterol <50 mg/dl With Rosuvastatin," *Journal of the American College of Cardiology* **57** no. 16, (Apr, 2011) 1666–1675.

[6] J. G. Robinson, S. Wang, and T. A. Jacobson, "Meta-Analysis of Comparison of Effectiveness of Lowering Apolipoprotein B Versus Low-Density

Lipoprotein Cholesterol and Nonhigh-Density Lipoprotein Cholesterol for Cardiovascular Risk Reduction in Randomized Trials," *The American Journal of Cardiology* **110** no. 10, (Nov, 2012) 1468–1476.

[7] G. L. Basatemur, H. F. Jørgensen, M. C. H. Clarke, M. R. Bennett, and Z. Mallat, "Vascular smooth muscle cells in atherosclerosis," *Nature Reviews Cardiology* **16** no. 12, (Dec, 2019) 727–744.

[8] E. Falk, "Pathogenesis of Atherosclerosis," *Journal of the American College of Cardiology* **47** no. 8 SUPPL., (2006) 0–5.

[9] D. Proudfoot and C. M. Shanahan, "Biology of Calcification in Vascular Cells: Intima versus Media," *Herz* **26** no. 4, (Jun, 2001) 245–251.

[10] J. J. Patel, L. E. Bourne, B. K. Davies, T. R. Arnett, V. E. MacRae, C. P. Wheeler-Jones, and I. R. Orriss, "Differing calcification processes in cultured vascular smooth muscle cells and osteoblasts," *Experimental Cell Research* **380** no. 1, (Jul, 2019) 100–113.

[11] A. Bellasi and P. Raggi, "Techniques and technologies to assess vascular calcification," *Seminars in Dialysis* **20** no. 2, (2007) 129–133.

[12] C. Karohl, L. D'Marco Gascón, and P. Raggi, "Noninvasive imaging for assessment of calcification in chronic kidney disease," *Nature Reviews Nephrology* **7** no. 10, (Oct, 2011) 567–577.

[13] C. D. Chue, N. A. Wall, N. J. Crabtree, D. Zehnder, W. E. Moody, N. C. Edwards, R. P. Steeds, J. N. Townend, and C. J. Ferro, "Aortic calcification and femoral bone density are independently associated with left ventricular mass in patients with chronic kidney disease," *PLoS ONE* **7** no. 6, (2012) .

[14] R. C. Johnson, J. A. Leopold, and J. Loscalzo, "Vascular Calcification," *Circulation Research* **99** no. 10, (Nov, 2006) 1044–1059.

[15] D. Yamamoto, S. Suzuki, H. Ishii, K. Hirayama, K. Harada, T. Aoki, Y. Shibata, Y. Negishi, Y. Tatami, T. Sumi, T. Ichii, K. Kawashima, A. Kunimura, T. Kawamiya, R. Morimoto, Y. Yasuda, and T. Murohara, "Predictors of abdominal aortic calcification progression in patients with chronic kidney disease without hemodialysis," *Atherosclerosis* **253** (Oct, 2016) 15–21.

[16] A. Bazzocchi, F. Ciccarese, D. Diano, P. Spinnato, U. Albisinni, C. Rossi, and G. Guglielmi, "Dual-Energy X-Ray Absorptiometry in the Evaluation of Abdominal Aortic Calcifications," *Journal of Clinical Densitometry* **15** no. 2, (2012) 198–204.

[17] H.-H. S. Oei, R. Vliegenthart, A. Hak, A. I. del Sol, A. Hofman, M. Oudkerk, and J. C. Witteman, "The association between coronary calcification assessed by electron beam computed tomography and measures of extracoronary atherosclerosis," *Journal of the American College of Cardiology* **39** no. 11, (Jun, 2002) 1745–1751.

[18] N. Westerhof, J. W. Lankhaar, and B. E. Westerhof, "The arterial windkessel," *Medical and Biological Engineering and Computing* **47** no. 2, (2009) 131–141.

[19] M. F. O'Rourke, "Arterial aging: pathophysiological principles.," *Vascular medicine (London, England)* **12** no. 4, (Nov, 2007) 329–41.

[20] S. S. Franklin, W. Gustin, N. D. Wong, M. G. Larson, M. A. Weber, W. B. Kannel, and D. Levy, "Hemodynamic Patterns of Age-Related Changes in Blood Pressure," *Circulation* **96** no. 1, (Jul, 1997) 308–315.

[21] C. M. McEniery, B. J. McDonnell, A. So, S. Aitken, C. E. Bolton, M. Munnery, S. S. Hickson, Yasmin, K. M. Maki-Petaja, J. R. Cockcroft, A. K. Dixon, and I. B. Wilkinson, "Aortic Calcification Is Associated With Aortic Stiffness and Isolated Systolic Hypertension in Healthy Individuals," *Hypertension* **53** no. 3, (Mar, 2009) 524–531.

[22] J. T. Schousboe, D. Claflin, and E. Barrett-Connor, "Association of Coronary Aortic Calcium With Abdominal Aortic Calcium Detected on Lateral Dual Energy X-Ray Absorptiometry Spine Images," *The American Journal of Cardiology* **104** no. 3, (Aug, 2009) 299–304, arXiv:NIHMS150003.

[23] A. S. Agatston, W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte, and R. Detrano, "Quantification of coronary artery calcium using ultrafast computed tomography," *Journal of the American College of Cardiology* **15** no. 4, (Mar, 1990) 827–832.

[24] Z. Qian, H. Anderson, I. Marvasty, K. Akram, G. Vazquez, S. Rinehart, and S. Voros, "Lesion- and vessel-specific coronary artery calcium scores are superior to whole-heart Agatston and volume scores in the diagnosis of obstructive coronary artery disease," *Journal of Cardiovascular Computed Tomography* **4** no. 6, (Nov, 2010) 391–399.

[25] E. Schulz, K. Arfai, X. Liu, J. Sayre, and V. Gilsanz, "Aortic Calcification and the Risk of Osteoporosis and Fractures," *The Journal of Clinical Endocrinology & Metabolism* **89** no. 9, (Sep, 2004) 4246–4253.

[26] J. J. Chan, L. A. Cupples, D. P. Kiel, C. J. O'Donnell, U. Hoffmann, and E. J. Samelson, "QCT Volumetric Bone Mineral Density and Vascular and Valvular Calcification: The Framingham Study," *Journal of Bone and Mineral Research* **30** no. 10, (Oct, 2015) 1767–1774.

[27] M. Cecelja, M. L. Frost, T. D. Spector, and P. Chowienczyk, "Abdominal aortic calcification detection using dual-energy X-ray absorptiometry: Validation study in healthy women compared to computed tomography," *Calcified Tissue International* **92** no. 6, (2013) 495–500.

[28] Y. S. Levitzky, L. A. Cupples, J. M. Murabito, W. B. Kannel, D. P. Kiel, P. W. Wilson, P. A. Wolf, and C. J. O'Donnell, "Prediction of Intermittent Claudication, Ischemic Stroke, and Other Cardiovascular Disease by Detection of Abdominal Aortic Calcific Deposits by Plain Lumbar Radiographs," *The American Journal of Cardiology* **101** no. 3, (Feb, 2008) 326–331.

[29] L. I. Kauppila, J. F. Polak, L. A. Cupples, M. T. Hannan, D. P. Kiel, and P. W. F. Wilson, "New indices to classify location, severity and progression of calcific lesions in the abdominal aorta: A 25-year follow-up study," *Atherosclerosis* **132** no. 2, (1997) 245–250.

[30] J. T. Schousboe, J. R. Lewis, and D. P. Kiel, "Abdominal aortic calcification on dual-energy X-ray absorptiometry: Methods of assessment and clinical significance," *Bone* **104** (Nov, 2017) 91–100.

[31] J. T. Schousboe, K. E. Wilson, and D. P. Kiel, "Detection of Abdominal Aortic Calcification With Lateral Spine Imaging Using DXA," *Journal of Clinical Densitometry* **9** no. 3, (2006) 302–308.

[32] E. Honkanen, L. Kauppila, B. Wikström, P. L. Rensma, J. M. Krzesinski, K. Aasarod, F. Verbeke, P. B. Jensen, P. Mattelaer, and B. Volck, "Abdominal aortic calcification in dialysis patients: Results of the CORD study," *Nephrology Dialysis Transplantation* **23** no. 12, (2008) 4009–4015.

[33] E. Pariente-Rodrigo, G. A. Sgaramella, P. García-Velasco, J. L. Hernández-Hernández, R. Landeras-Alvaro, and J. M. Olmos-Martínez,

"Reliability of radiologic evaluation of abdominal aortic calcification using the 24-point scale.," *Radiologia* **58** no. 1, (2016) 46–54.

[34] R. Setiawati, F. Di Chio, P. Rahardjo, M. Nasuto, F. J. Dimpudus, and G. Guglielmi, "Quantitative Assessment of Abdominal Aortic Calcifications Using Lateral Lumbar Radiograph, Dual-Energy X-ray Absorptiometry, and Quantitative Computed Tomography of the Spine," *Journal of Clinical Densitometry* **19** no. 2, (2016) 242–249.

[35] P. Szulc, "Abdominal aortic calcification: A reappraisal of epidemiological and pathophysiological data," *Bone* **84** (Mar, 2016) 25–37.

[36] M. A. Frye, L. J. Melton, S. C. Bryant, L. A. Fitzpatrick, H. W. Wahner, R. S. Schwartz, and B. L. Riggs, "Osteoporosis and calcification of the aorta," *Bone and Mineral* **19** no. 2, (1992) 185–194.

[37] M. A. Ikram, G. G. Brusselle, S. D. Murad, C. M. van Duijn, O. H. Franco, A. Goedegebure, C. C. Klaver, T. E. Nijsten, R. P. Peeters, B. H. Stricker, H. Tiemeier, A. G. Uitterlinden, M. W. Vernooij, and A. Hofman, "The Rotterdam Study: 2018 update on objectives, design and main results," *European Journal of Epidemiology* **32** no. 9, (2017) 807–850.

[38] M. Hollander, A. Hak, P. Koudstaal, M. Bots, D. Grobbee, A. Hofman, J. Witteman, and M. Breteler, "Comparison Between Measures of Atherosclerosis and Risk of Stroke," *Stroke* **34** no. 10, (Oct, 2003) 2367–2372.

[39] M. Ganz, M. de Bruijne, and M. Nielsen, "MACD: an imaging marker for cardiovascular disease," in *SPIE Medical Imaging*, N. Karssemeijer and R. M. Summers, eds., vol. 7624, p. 76240P. Mar, 2010.

[40] M. Nielsen, M. Ganz, F. Lauze, P. C. Pettersen, M. de Bruijne, T. B. Clarkson, E. B. Dam, C. Christiansen, and M. A. Karsdal, "Distribution, size, shape, growth potential and extent of abdominal aortic calcified deposits predict mortality in postmenopausal women," *BMC Cardiovascular Disorders* **10** no. 1, (Dec, 2010) 56.

[41] National Institute for Health and Care Excellence, "Osteoporosis - prevention of fragility fractures," 2016. `https://cks.nice.org.uk/osteoporosis-prevention-of-fragility-fractures`.

[42] J. P. W. van den Bergh, T. A. C. M. van Geel, W. F. Lems, and P. P. Geusens, "Assessment of Individual Fracture Risk: FRAX and Beyond," *Current Osteoporosis Reports* **8** no. 3, (Sep, 2010) 131–137.

[43] T. J. Aspray, "Fragility fracture: recent developments in risk assessment," *Therapeutic Advances in Musculoskeletal Disease* **7** no. 1, (Feb, 2015) 17–25.

[44] M. Naves, M. Rodríguez-García, J. B. Díaz-López, C. Gómez-Alonso, and J. B. Cannata-Andía, "Progression of vascular calcifications is associated with greater bone loss and increased bone fractures," *Osteoporosis International* **19** no. 8, (Aug, 2008) 1161–1166.

[45] P. Szulc, E. J. Samelson, E. Sornay-Rendu, R. Chapurlat, and D. P. Kiel, "Severity of aortic calcification is positively associated with vertebral fracture in older men - A densitometry study in the STRAMBO cohort," *Osteoporosis International* **24** no. 4, (2013) 1177–1184.

[46] J. Damilakis, J. E. Adams, G. Guglielmi, and T. M. Link, "Radiation exposure in X-ray-based imaging techniques used in osteoporosis," *European Radiology* **20** no. 11, (2010) 2707–2714.

[47] J. T. Schousboe, K. E. Wilson, and T. N. Hangartner, "Detection of aortic calcification during vertebral fracture assessment (VFA) compared to digital radiography," *PLoS ONE* **2** no. 8, (2007) 1–5.

[48] N. D. Toussaint, K. K. Lau, B. J. Strauss, K. R. Polkinghorne, and P. G. Kerr, "Determination and Validation of Aortic Calcification Measurement from Lateral Bone Densitometry in Dialysis Patients," *Clinical Journal of the American Society of Nephrology* **4** no. 1, (Jan, 2009) 119–127.

[49] N. D. Toussaint, K. K. Lau, B. J. Strauss, K. R. Polkinghorne, and P. G. Kerr, "Using vertebral bone densitometry to determine aortic calcification in patients with chronic kidney disease," *Nephrology (Carlton, Vic.)* **15** no. 5, (2010) 575.

[50] N. Barascuk, M. Ganz, M. Nielsen, T. C. Register, L. M. Rasmussen, M. a. Karsdal, and C. Christiansen, "Abdominal aortic calcification quantified by the Morphological Atherosclerotic Calcification Distribution (MACD) index is associated with features of the metabolic syndrome," *BMC Cardiovascular Disorders* **11** no. 1, (Dec, 2011) 75.

[51] A. L. Catapano, I. Graham, G. De Backer, O. Wiklund, M. J. Chapman, H. Drexel, A. W. Hoes, C. S. Jennings, U. Landmesser, T. R. Pedersen, Ž. Reiner, G. Riccardi, M.-R. Taskinen, L. Tokgozoglu, W. M. Verschuren, C. Vlachopoulos, D. A. Wood, and J. L. Zamorano, "2016 ESC/EAS Guidelines for the Management of Dyslipidaemias," *Atherosclerosis* **253** (Oct, 2016) 281–344.

[52] N. K. Wenger, "Coronary heart disease: the female heart is vulnerable," *Progress in Cardiovascular Diseases* **46** no. 3, (Nov, 2003) 199–229.

[53] H. C. McGill, C. A. McMahan, and S. S. Gidding, "Preventing Heart Disease in the 21st Century," *Circulation* **117** no. 9, (Mar, 2008) 1216–1227.

[54] National Institute for Health and Care Excellence, "CVD risk assessment and management," 2019. https://cks.nice.org.uk/cvd-risk-assessment-and-management.

[55] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study," *BMJ* (May, 2017) j2099.

[56] P. W. Wilson, L. I. Kauppila, C. J. O'Donnell, D. P. Kiel, M. Hannan, J. M. Polak, and L. A. Cupples, "Abdominal aortic calcific deposits are an important predictor of vascular morbidity and mortality," *Circulation* **103** no. 1524-4539, (2001) 1529–1534.

[57] U. Hoffmann, J. M. Massaro, R. B. D'Agostino, S. Kathiresan, C. S. Fox, and C. J. O'Donnell, "Cardiovascular Event Prediction and Risk Reclassification by Coronary, Aortic, and Valvular Calcification in the Framingham Heart Study," *Journal of the American Heart Association* **5** no. 2, (Feb, 2016) .

[58] F. Bastos Gonçalves, M. T. Voûte, S. E. Hoeks, M. B. Chonchol, E. E. Boersma, R. J. Stolker, and H. J. M. Verhagen, "Calcification of the abdominal aorta as an independent predictor of cardiovascular events: a meta-analysis.," *Heart (British Cardiac Society)* **98** no. 13, (Jul, 2012) 988–94.

[59] M. J. Bolland, T. K. Wang, N. C. van Pelt, A. M. Horne, B. H. Mason, R. W. Ames, A. B. Grey, P. N. Ruygrok, G. D. Gamble, and I. R. Reid, "Abdominal aortic calcification on vertebral morphometry images predicts incident myocardial infarction," *Journal of Bone and Mineral Research* **25** no. 3, (Mar, 2010) 505–512.

[60] I. M. Van Der Meer, M. L. Bots, A. Hofman, A. I. Del Sol, D. A. M. Van Der Kuip, and J. C. M. Witteman, "Predictive Value of Noninvasive Measures of

Atherosclerosis for Incident Myocardial Infarction: The Rotterdam Study," *Circulation* **109** no. 9, (2004) 1089–1094.

[61] C. Walsh, "Abdominal aortic calcific deposits are associated with increased risk for congestive heart failure: The Framingham Heart Study," *American Heart Journal* **144** no. 4, (Oct, 2002) 733–739.

[62] S. a. E. Peters, H. M. den Ruijter, M. L. Bots, and K. G. M. Moons, "Improvements in risk stratification for the occurrence of cardiovascular disease by imaging subclinical atherosclerosis: a systematic review," *Heart* **98** no. 3, (2012) 177–184.

[63] R. H. Davies, C. J. Twining, P. D. Allen, T. F. Cootes, and C. J. Taylor, "Shape Discrimination in the Hippocampus Using an MDL Model," in *IPMI 2003: Information Processing in Medical Imaging*, pp. 38–50. 2003.

[64] T. F. Cootes and C. J. Taylor, "Active Shape Models — 'Smart Snakes'," in *BMVC92*, pp. 266–275. Springer London, London, 1992.

[65] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** no. 6, (Jun, 2001) 681–685.

[66] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition* **41** no. 10, (Oct, 2008) 3054–3067.

[67] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** no. 9, (Sep, 2015) 1862–1874.

[68] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes, "Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting," *IEEE Transactions on Medical Imaging* **32** no. 8, (Aug, 2013) 1462–1472.

[69] J. Thomson, T. O'Neill, D. Felson, and T. Cootes, "Automated Shape and Texture Analysis for Detection of Osteoarthritis from Radiographs of the Knee," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 127–134. Springer, Cham, 2015.

[70] P. Bromiley, T. Cootes, and J. Adams, "Localisation of Vertebrae on DXA Images Using Constrained Local Models with Random Forest Regression Voting," *Lecture Notes in Computational Vision and Biomechanics* **20** (2015) 235–240.

[71] C. J. Dean, J. R. Sykes, R. A. Cooper, P. Hatfield, B. Carey, S. Swift, S. E. Bacon, D. Thwaites, D. Sebag-Montefiore, and A. M. Morgan, "An evaluation of four CT–MRI co-registration techniques for radiotherapy treatment planning of prone rectal cancer patients," *The British Journal of Radiology* **85** no. 1009, (Jan, 2012) 61–68.

[72] A. Toga and P. Thompson, "The role of image registration in brain mapping," *Image and Vision Computing* **19** no. 1-2, (Jan, 2001) 3–24.

[73] F. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11** no. 6, (Jun, 1989) 567–585.

[74] N. Ramesh, "Thresholding based on histogram approximation," *IEE Proceedings - Vision, Image, and Signal Processing* **142** no. 5, (1995) 271.

[75] A. Brink, "Minimum spatial entropy threshold selection," *IEE Proceedings - Vision, Image, and Signal Processing* **142** no. 3, (1995) 128.

[76] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282. IEEE Comput. Soc. Press, 1995.

[77] L. Breiman, "Random Forests," *Machine Learning* **45** no. 1, (2001) 5–32.

[78] J. R. Quinlan, "Induction of decision trees," *Machine Learning* **1** no. 1, (Mar, 1986) 81–106.

[79] L. Breiman, "Bagging predictors," *Machine Learning* **24** no. 2, (Aug, 1996) 123–140.

[80] P. Probst and A.-L. Boulesteix, "To Tune or Not to Tune the Number of Trees in Random Forest," *Journal of Machine Learning Research* **18** (2018) 1–18.

[81] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** no. 8, (1998) 832–844.

[82] E. Gatnar, "Dimensionality of Random Subspaces," in *Classification — the Ubiquitous Challenge*, pp. 129–136. Springer-Verlag, Berlin/Heidelberg, 2005.

[83] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–511–I–518. IEEE Comput. Soc, 2001.

[84] M. G. Roberts, T. F. Cootes, and J. E. Adams, "Automatic location of vertebrae on DXA images using random forest regression.," *Medical image computing*

and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **15** no. Pt 3, (2012) 361–8.

[85] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images," *NeuroImage* **57** no. 2, (Jul, 2011) 378–390.

[86] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression Forests for Efficient Anatomy Detection and Localization in CT Studies," in *MCV 2010: Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pp. 106–117. 2011.

[87] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans," in *MICCAI 2012: Medical Image Computing and Computer-Assisted Intervention*, pp. 590–598. 2012.

[88] V. Lempitsky, M. Verhoek, J. A. Noble, and A. Blake, "Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3D Echocardiography," in *FIMH 2009: Functional Imaging and Modeling of the Heart*, pp. 447–456. 2009.

[89] D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and S. J. Price, "Decision Forests for Tissue-Specific Segmentation of High-Grade Gliomas in Multi-channel MR," in *MICCAI 2012: Medical Image Computing and Computer-Assisted Intervention*, pp. 369–376. 2012.

[90] J. Mitra, P. Bourgeat, J. Fripp, S. Ghose, S. Rose, O. Salvado, A. Connelly, B. Campbell, S. Palmer, G. Sharma, S. Christensen, and L. Carey, "Lesion segmentation from multimodal MRI using random forest following ischemic stroke," *NeuroImage* **98** (Sep, 2014) 324–335.

[91] F. Khalifa, A. Soliman, A. C. Dwyer, G. Gimel'farb, and A. El-Baz, "A random forest-based framework for 3D kidney segmentation from dynamic contrast-enhanced CT images," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3399–3403. IEEE, Sep, 2016.

[92] J. Liu, J. Hoffman, J. Zhao, J. Yao, L. Lu, L. Kim, E. B. Turkbey, and R. M. Summers, "Mediastinal lymph node detection and station mapping on chest CT using spatial priors and random forest," *Medical Physics* **43** no. 7, (Jun, 2016) 4362–4374.

[93] F. Rosenblatt, "The Perception - A Perceiving and Recognizing Automation," *Tech. Rep. 85-460-1 (Cornell Aeronautical Laboratory)* (1957) .

[94] B. Widrow, "Generalization and Information Storage in Networks of Adaline Neurons," in *Self-Organizing Systems*, M. D. Yovits, G. T. Jacobi, and G. D. Goldstein, eds., pp. 435–461. Spartan Books, Washington DC, 1962.

[95] S. Rota Bulo and P. Kontschieder, "Neural Decision Forests for Semantic Image Labelling," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 81–88. IEEE, Jun, 2014.

[96] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bulo, "Deep Neural Decision Forests," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1467–1475. IEEE, Dec, 2015.

[97] X. Dong, C. J. Taylor, and T. F. Cootes, "Small Defect Detection Using Convolutional Neural Network Features and Random Forests," in *ECCV 2018: Computer Vision – ECCV 2018 Workshops*, pp. 398–412. 2019.

[98] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On Using Very Large Target Vocabulary for Neural Machine Translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1–10. Association for Computational Linguistics, Stroudsburg, PA, USA, 2015.

[99] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships," *Journal of Chemical Information and Modeling* **55** no. 2, (Feb, 2015) 263–274.

[100] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics* **5** no. 4, (Dec, 1943) 115–133.

[101] D. Hebb, ": The Organization of Behavior," in *The Organisation of Behavior; A Neuropsychological Theory*. NY: Wiley, 1949.

[102] M. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969.

[103] D. Rumelhart, G. Hinton, and R. Williams, "Learning Internal Representations by Error Propagation," in *Readings in Cognitive Science*, pp. 399–421. Elsevier, 1988.

[104] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521** no. 7553, (May, 2015) 436–444.

[105] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **06** no. 02, (Apr, 1998) 107–116.

[106] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," *Proceedings of Machine Learning Research* **15** (2011) 315–323.

[107] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Neural Information Processing Systems* **25** (May, 2012) .

[108] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science* **8691** (2014) 346–361.

[109] D. C. Ciresan, A. Giusti, and L. M. Gambardella, "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, pp. 2843–2851. Curran Associates Inc., 2012.

[110] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440. IEEE, Jun, 2015.

[111] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2528–2535. IEEE, Jun, 2010.

[112] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, Cham, 2015.

[113] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics* **22** no. 3, (Sep, 1951) 400–407.

[114] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," *The Annals of Mathematical Statistics* **23** no. 3, (Sep, 1952) 462–466.

[115] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186. Physica-Verlag HD, Heidelberg, 2010.

[116] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature* **323** no. 6088, (Oct, 1986) 533–536.

[117] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O(1/kˆ2)," *Soviet Mathematics Doklady* **27** no. 2, (1983) 372–376.

[118] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in Optimizing Recurrent Networks," `arXiv:1212.0901`.

[119] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, eds., pp. 1139—-1147. PMLR, Atlanta, Georgia, USA, 2013.

[120] R. A. Jacobs, "Increased rates of convergence through learning rate adaptation," *Neural Networks* **1** no. 4, (Jan, 1988) 295–307.

[121] C. Darken, J. Chang, and J. Moody, "Learning rate schedules for faster stochastic gradient search," in *Neural Networks for Signal Processing II Proceedings of the 1992 IEEE Workshop*, vol. 3, pp. 3–12. IEEE, 1992.

[122] B. K. Singh, K. Verma, and A. Thoke, "Adaptive Gradient Descent Backpropagation for Classification of Breast Tumors in Ultrasound Imaging," *Procedia Computer Science* **46** (2015) 1601–1609.

[123] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," arXiv:1212.5701.

[124] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (Dec, 2014) 1–15.

[125] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research* **13** (2012) 281–305.

[126] J. Mockus, "On the Bayes Methods for Seeking the Extremal Point," *IFAC Proceedings Volumes* **8** no. 1, (Aug, 1975) 428–431.

[127] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems* **2** (2012) 2951–2959.

[128] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems* (2011) 2546–2554.

[129] C. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** no. 12, (1998) 1342–1351.

[130] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting," *IEEE Transactions on Information Theory* **58** no. 5, (May, 2012) 3250–3265.

[131] H. J. Kushner, "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise," *Journal of Basic Engineering* **86** no. 1, (Mar, 1964) 97–106.

[132] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *Journal of Global Optimization* **21** (2001) 345–383.

[133] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization* **13** (1998) 455–492.

[134] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology* **17** no. 1, (2019) 26–40.

[135] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields," *Remote Sensing* **9** no. 7, (2017) 1–19.

[136] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters* **15** no. 5, (May, 2018) 749–753.

[137] J. McGlinchy, B. Johnson, B. Muller, M. Joseph, and J. Diaz, "Application of UNet Fully Convolutional Neural Network to Impervious Surface

Segmentation in Urban Environment from High Resolution Satellite Imagery," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3915–3918. IEEE, Jul, 2019.

[138] H. He, D. Yang, S. Wang, S. Wang, and Y. Li, "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss," *Remote Sensing* **11** no. 9, (2019) 1–16.

[139] L.-A. Tran and M.-H. Le, "Robust U-Net-based Road Lane Markings Detection for Autonomous Driving," in *2019 International Conference on System Science and Engineering (ICSSE)*, pp. 62–66. IEEE, Jul, 2019.

[140] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, E. Gomez, X. Hu, E. Humphrey, and E. Benetos, eds., pp. 334—-340. 2018.

[141] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, A. Dovzhenko, O. Tietz, C. Dal Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, and O. Ronneberger, "U-Net: deep learning for cell counting, detection, and morphometry," *Nature Methods* **16** no. 1, (Jan, 2019) 67–70.

[142] X. Gao, Y. Cai, C. Qiu, and Y. Cui, "Retinal blood vessel segmentation based on the Gaussian matched filter and U-net," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5. IEEE, Oct, 2017.

[143] R. Asgari, S. Waldstein, F. Schlanitz, M. Baratsits, U. Schmidt-Erfurth, and H. Bogunović, "U-Net with Spatial Pyramid Pooling for Drusen Segmentation

in Optical Coherence Tomography," *Ophthalmic Medical Image Analysis. OMIA 2019. Lecture Notes in Computer Science.* **11855** (2019) 77–85.

[144] 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)*, pp. 565–571. IEEE, Oct, 2016.

[145] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *IEEE Transactions on Medical Imaging*, vol. 39, pp. 3–11. Springer International Publishing, Jun, 2018.

[146] W. Chen, Y. Zhang, J. He, Y. Qiao, Y. Chen, H. Shi, E. X. Wu, and X. Tang, "Prostate Segmentation using 2D Bridged U-net," in *2019 International Joint Conference on Neural Networks (IJCNN)*, no. July, pp. 1–7. IEEE, Jul, 2019.

[147] A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, and X. Lladó, "Acute and sub-acute stroke lesion segmentation from multimodal MRI," *Computer Methods and Programs in Biomedicine* **194** (Oct, 2020) 105521.

[148] M. Livne, J. Rieger, O. U. Aydin, A. A. Taha, E. M. Akay, T. Kossen, J. Sobesky, J. D. Kelleher, K. Hildebrand, D. Frey, and V. I. Madai, "A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients With Cerebrovascular Disease," *Frontiers in Neuroscience* **13** no. Feb, (Feb, 2019) 1–13.

[149] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks," *Communications in Computer and Information Science Medical Image Understanding and Analysis* (2017) 506–517.

[150] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal* **27** no. 3, (Jul, 1948) 379–423.

[151] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science*, vol. 9901, pp. 424–432. Jun, 2016.

[152] H. Hwang, H. Z. U. Rehman, and S. Lee, “3D U-Net for Skull Stripping in Brain MRI,” *Applied Sciences* **9** no. 3, (Feb, 2019) 569.

[153] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-Decem, pp. 770–778. IEEE, Jun, 2016.

[154] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science* **9908** (2016) 630–645.

[155] X. Xiao, S. Lian, Z. Luo, and S. Li, “Weighted Res-UNet for High-Quality Retina Vessel Segmentation,” in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 327–331. IEEE, Oct, 2018.

[156] G. Bortsova, G. van Tulder, F. Dubost, T. Peng, N. Navab, A. van der Lugt, D. Bos, and M. De Bruijne, “Segmentation of Intracranial Arterial Calcification with Deeply Supervised Residual Dropout Networks,” in *Medical Image Computing and Computer Assisted Intervention MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science*, vol. 10435, pp. 356–364. 2017.

[157] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* **2017-July** (2017) 1175–1183.

[158] Z. Qiang, S. Tu, and L. Xu, "A k-Dense-UNet for Biomedical Image Segmentation," in *Intelligence Science and Big Data Engineering. Visual Data Engineering. IScIDE 2019. Lecture Notes in Computer Science*, vol. 11935, pp. 552–562. Springer International Publishing, 2019.

[159] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-unet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quantitative Imaging in Medicine and Surgery* **10** no. 6, (2020) 1275–1285.

[160] J. Dolz, I. Ben Ayed, and C. Desrosiers, "Dense Multi-path U-Net for Ischemic Stroke Lesion Segmentation in Multiple Image Modalities," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science*, vol. 11383, pp. 271–282. Springer International Publishing, 2019.

[161] I. Isgum, B. van Ginneken, and M. Prokop, "A pattern recognition approach to automated coronary calcium scoring," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 746–749 Vol.3. IEEE, 2004.

[162] I. Isgum, B. van Ginneken, A. Rutten, and M. Prokop, "Automated coronary calcification detection and scoring," in *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, vol. 2005, pp. 127–132. IEEE, 2005.

[163] Y. Xie, M. D. Cham, C. Henschke, D. Yankelevitz, and A. P. Reeves, "Automated coronary artery calcification detection on low-dose chest CT images," in *Medical Imaging 2014: Computer-Aided Diagnosis*, S. Aylward and L. M. Hadjiiski, eds., vol. 9035, p. 90350F. Mar, 2014.

[164] U. Kurkure, D. R. Chittajallu, G. Brunner, Y. H. Le, and I. A. Kakadiaris, "A supervised classification-based method for coronary calcium detection in non-contrast CT," *The International Journal of Cardiovascular Imaging* **26** no. 7, (Oct, 2010) 817–828.

[165] S. C. Saur, H. Alkadhi, L. Desbiolles, G. Székely, and P. C. Cattin, "Automatic Detection of Calcified Coronary Plaques in Computed Tomography Data Sets," *Med Image Comput Comput Assist Interv.* **11** (2008) 170–177.

[166] R. Shahzad, M. Schaap, T. van Walsum, S. Klien, A. C. Weustink, L. J. van Vliet, and W. J. Niessen, "A patient-specific coronary density estimate," in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 9–12. IEEE, 2010.

[167] G. Brunner, D. R. Chittajallu, U. Kurkure, and I. A. Kakadiaris, "Toward the automatic detection of coronary artery calcification in non-contrast computed tomography data," *The International Journal of Cardiovascular Imaging* **26** no. 7, (Oct, 2010) 829–838.

[168] I. Isgum, M. Prokop, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Automatic Coronary Calcium Scoring in Low-Dose Chest Computed Tomography," *IEEE Transactions on Medical Imaging* **31** no. 12, (Dec, 2012) 2322–2334.

[169] I. Išgum, A. Rutten, M. Prokop, M. Staring, S. Klein, J. P. W. Pluim, M. A. Viergever, and B. van Ginneken, "Automated aortic calcium scoring on

low-dose chest computed tomography," *Medical Physics* **37** no. 2, (Jan, 2010) 714–723.

[170] S. Kurugol, C. E. Come, A. A. Diaz, J. C. Ross, G. L. Kinney, J. L. Black-Shinn, J. E. Hokanson, M. J. Budoff, G. R. Washko, and R. San Jose Estepar, "Automated quantitative 3D analysis of aorta size, morphology, and mural calcification distributions," *Medical Physics* **42** no. 9, (2015) 5467–5478.

[171] G. Santini, D. D. Latta, N. Martini, G. Valvano, A. Gori, A. Ripoli, C. L. Susini, L. Landini, and D. Chiappino, "An automatic deep learning approach for coronary artery calcium segmentation," *EMBEC & NBC 2017. IFMBE Proceedings* **65** (2018) 374–377.

[172] G. González, G. R. Washko, R. S. J. Estépar, M. Cazorla, and C. Cano Espinosa, "Automated Agatston score computation in non-ECG gated CT scans using deep learning," in *Medical Imaging 2018: Image Processing*, E. D. Angelini and B. A. Landman, eds., vol. 176, p. 91. SPIE, Mar, 2018.

[173] N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Isgum, "Automatic Calcium Scoring in Low-Dose Chest CT Using Deep Neural Networks With Dilated Convolutions," *IEEE Transactions on Medical Imaging* **37** no. 2, (Feb, 2018) 615–625.

[174] B. Liu and B. Hua, "Semi-supervised semantic image segmentation using dual discriminator adversarial networks," in *Eleventh International Conference on Digital Image Processing (ICDIP 2019)*, X. Jiang and J.-N. Hwang, eds., vol. 1117907, p. 54. SPIE, Aug, 2019.

[175] P. M. Graffy, J. Liu, S. O'Connor, R. M. Summers, and P. J. Pickhardt, "Automated segmentation and quantification of aortic calcification at

abdominal CT: application of a deep learning-based algorithm to a longitudinal screening cohort," *Abdominal Radiology* **44** no. 8, (2019) 2921–2928.

[176] F. Lauze and M. de Bruijne, "Toward automated detection and segmentation of aortic calcifications from radiographs," in *Proc. SPIE 6512, Medical Imaging 2007: Image Processing*, J. P. W. Pluim and J. M. Reinhardt, eds., vol. 651239-7. 2007.

[177] K. Petersen, M. Nielsen, and S. S. Brandt, "A Static SMC Sampler on Shapes for the Automated Segmentation of Aortic Calcifications," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6314 LNCS, pp. 666–679. 2010.

[178] K. Petersen, M. Ganz, P. Mysling, M. Nielsen, L. Lillemark, A. Crimi, and S. S. Brandt, "A Bayesian Framework for Automated Cardiovascular Risk Scoring on Standard Lumbar Radiographs," *IEEE Transactions on Medical Imaging* **31** no. 3, (Mar, 2012) 663–676.

[179] K. Elmasri, Y. Hicks, X. Yang, X. Sun, R. Pettit, and W. Evans, "Automatic Detection and Quantification of Abdominal Aortic Calcification in Dual Energy X-ray Absorptiometry," *Procedia Computer Science* **96** (2016) 1011–1021.

[180] D. Kuh, M. Pierce, J. Adams, J. Deanfield, U. Ekelund, P. Friberg, A. K. Ghosh, N. Harwood, A. Hughes, P. W. Macfarlane, G. Mishra, D. Pellerin, A. Wong, A. M. Stephen, M. Richards, and R. Hardy, "Cohort Profile: Updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research," *International Journal of Epidemiology* **40** no. 1, (Feb, 2011) e1–e9.

[181] R. L. Prince, A. Devine, S. S. Dhaliwal, and I. M. Dick, "Effects of Calcium

Supplementation on Clinical Fracture and Bone Structure," *Archives of Internal Medicine* **166** no. 8, (Apr, 2006) 869.

[182] P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes, "Fully automatic localisation of vertebrae in CT images using random forest regression voting," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10182 LNCS** (2016) 51–63.

[183] J. E. Adams, "Radiogrammetry and Radiographic Absorptiometry," *Radiologic Clinics of North America* **48** no. 3, (May, 2010) 531–540.

[184] P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes, "Classification of Osteoporotic Vertebral Fractures Using Shape and Appearance Modelling," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10734 LNCS, pp. 133–147. 2018.

[185] A. V. Pavlova, F. R. Saunders, S. G. Muthuri, J. S. Gregory, R. J. Barr, K. R. Martin, R. J. Hardy, R. Cooper, J. E. Adams, D. Kuh, and R. M. Aspden, "Statistical shape modelling of hip and lumbar spine morphology and their relationship in the MRC National Survey of Health and Development," *Journal of Anatomy* **231** no. 2, (2017) 248–259.

[186] J. R. Lewis, J. T. Schousboe, W. H. Lim, G. Wong, K. E. Wilson, K. Zhu, P. L. Thompson, D. P. Kiel, and R. L. Prince, "Long-Term Atherosclerotic Vascular Disease Risk and Prognosis in Elderly Women With Abdominal Aortic Calcification on Lateral Spine Images Captured During Bone Density Testing: A Prospective Study," *Journal of Bone and Mineral Research* **33** no. 6, (2018) 1001–1010.

[187] L. Lefkovits, S. Lefkovits, and L. Szilágyi, "Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries," *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2016. Lecture Notes in Computer Science* **10154** (2016) 88–99.

[188] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015. `https://www.tensorflow.org/`.

[189] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* no. ICML'15, (2015) 448–456.

[190] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How Does Batch Normalization Help Optimization?," *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (May, 2018) 2483–2493.

[191] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation," `arXiv:1801.05746`.

[192] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao, "Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical

Images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2074–2083. IEEE, Jun, 2019.

[193] L. A. Chaplin and T. F. Cootes, "Automated scoring of aortic calcification in vertebral fracture assessment images," in *Medical Imaging 2019: Computer-Aided Diagnosis*, H. K. Hahn and K. Mori, eds., vol. 1095038, p. 116. SPIE, Mar, 2019.

[194] S. S. Raut, S. Chandra, J. Shum, and E. A. Finol, "The Role of Geometric and Biomechanical Factors in Abdominal Aortic Aneurysm Rupture Risk Assessment," *Annals of Biomedical Engineering* **41** no. 7, (Jul, 2013) 1459–1477.

[195] F. Lareyre, C. Adam, M. Carrier, C. Dommerc, C. Mialhe, and J. Raffort, "A fully automated pipeline for mining abdominal aortic aneurysm using image segmentation," *Scientific Reports* **9** no. 1, (Dec, 2019) 13750.