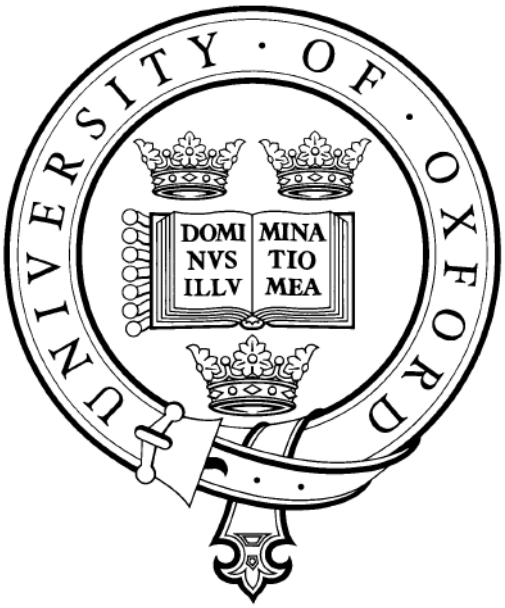
Automated Analysis of Spinal MRI using Deep Learning



Amir Jamaludin

St Hilda's College

Supervised by
Professor Andrew Zisserman
Doctor Timor Kadir

Visual Geometry Group
Department of Engineering Science
University of Oxford
Submitted: Michaelmas Term 2017

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Amir Jamaludin, St Hilda's College

Copyright © 2017
Amir Jamaludin
All Rights Reserved

Abstract

The objective of this thesis is the automation of radiological gradings in spinal lumbar Magnetic Resonance Images (MRIs). Solving this is extremely beneficial as this in a way would help in the standardization of gradings especially for back pain research. The output of the research done in this thesis would allow extremely fast readings of clinical scans which can potentially be useful in a large scale epidemiological study of spine-related diseases and aide clinical decision making.

First, we build a pipeline to automatically produce radiological gradings of spinal lumbar MRIs and also localize the predicted pathologies. We show that: (i) a Convolutional Neural Network (CNN) is able to predict multiple gradings at once, and we propose variants of the architecture including a multi-modal CNN that is able to take in both axial and sagittal or T1-weighted and T2-weighted scans; and (ii) a localization method that clearly shows pathological regions in the disc volumes using only a CNN trained for classification. The CNN is applied to a large corpus of standard clinical scan MRIs acquired from multiple machines via various scanning protocol, and is used to automatically compute intervertebral disc and vertebral body gradings for each MRI. We explore several radiological gradings: Pfirrmann grading, disc narrowing, upper/lower endplate defects, upper/lower marrow changes, spondylolisthesis, central canal stenosis, anterior/posterior disc bulging, and disc herniation. We report near human performances across all the gradings, and also visualize the evidence for these gradings localized on the original scans.

Then, since a significant proportion of patients scanned in a clinical setting have follow-up scans; we show that such longitudinal scans alone can be used as a form of "free" self-supervision for training a deep network. We demonstrate this self-supervised learning for the case of T2-weighted sagittal lumbar MRIs. This learning via self-supervision can act as a pre-training regime when labelled data is sparse. We show that the performance of the pre-trained CNN on the supervised classification task is (i) superior to that of a network trained from scratch; and (ii) requires far fewer annotated training samples to reach an equivalent performance to that of the network trained from scratch.

Finally, we show some preliminary results in mapping disc features learnt from radiological gradings to the Oswestry Disability Index (ODI) which is a measure of disability commonly used by back pain patients.

Acknowledgements

First of all, I would like to acknowledge my two bosses: Professor Andrew Zisserman and Dr Timor Kadir. I believe that the work in this thesis conducted over the course of these three years can easily be completed without me but would not have worked without Andrew or Timor.

I would also like to thank Professor Jeremy Fairbank and Dr Jill Urban for their guidance on the clinical aspect of this thesis throughout this DPhil. Special thanks to Professor Iain McCall and Dr Michele Battié for their work on the **Genodisc** dataset, and Professor Frances Williams for providing the **TwinsUK** data.

A big thank you to the members of the VGG especially Joon, Yujie, Meelis, Abhishek, Carlos, Aravindh, Mohsan and several others for making the lab interesting. I would particularly miss my weekly lunch/dinner date with Joon and Yujie.

Finally, I would like to express my gratitude to my family particularly my wife, Syarah, other members of my huge family which includes Stormie, Opah, Atuk, my parents, Aileen, Pok De, Mok De, Mak Ngah and all the other 'Mak/Mok's and 'Pak/Pok's for their support. Thank you.

Contents

Chapter 1

Introduction

Back pain is one of the most common health problems in the world. Palmer et al. (2000) estimated that the prevalence of back pain to be around 80% in the UK; meaning around four out of five people in the UK would likely to have experienced back pain. Even though it is so widespread, it remains to be quite a mystery. Burton et al. (2005) found out that around 85% of patients with back pain cannot be attributed to any pathology. This significant chunk of back pain patients are thus diagnosed with having non-specific back pain.

One current prevalent research area in the quest of demystifying back pain is the study trying to tie radiology to back pain. This is normally conducted via two main imaging techniques: Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans. Out of the two, MRI is more popular due to the fact that it is the only NICE-approved image modality for back pain diagnosis as specified by Savigny et al. (2009), and is often the go-to imaging technique in clinic. Radiological analysis of spinal MRIs is typically anatomically specific, with the most popular being the intervertebral discs i.e. the disc in between vertebral bodies (Battie et al. (2014)). Because of this, there exists several measures or radiological gradings designed to measure levels of disc degeneration unique to MRIs. These are often attached to

radiological reports of patients written by radiologists. In a sense, these radiologist-read radiological gradings, can be interpreted as features for regression to back pain.

However, to date there is still no clear relationship between back pain and any of the numerous radiological gradings; Brinjikji et al. (2015) hypothesized that many radiological gradings might just be an outcome of ageing and not associated with pain while MacGregor et al. (2004) found that in their study, MRI-based disc degeneration scores are the main prediction of back pain. One clear distinction between a lot of these studies is simple; there is a lack of coherent definition of the many radiological gradings e.g. an **endplate defect**, defects that can be seen in the vertebral endplate, can mean one thing to one group and can mean another thing to a different group; they can be Schmorl's nodes, vertebral fracture, etc.

In this thesis, we aim to automate the grading process which would entail in developing a system or algorithm to process spinal MRIs and outputting a series of quantitative radiological gradings. This automated grading system would aide researchers in dealing with a large amount of spinal MRIs with the added benefit of a more standardized definition of gradings; gradings produced by the system are invariant to outside factors i.e. little to no intra-observer variation. This is theory would solve one of the major problems in a large scale study involving radiology and back pain. Furthermore, an automated system can also support clinical decision making. As the utilization and interpretation of medical images continues to expand beyond the radiology department, as it has in cardiology and neurology, it is becoming increasingly important to supplement the typically qualitative radiological readings with objective quantitative methods. As an added benefit, our quantitative output would also be complete, instead of the normally sparse annotations in a standard radiological report.

1.1 Challenges

Variation in Radiological Gradings. The main challenge of learning from radiological information is the extremely grey line separating a normal and abnormal appearance. This results in a variation in readings of the same subject or even the same scan. One example of this was presented by Herzog et al., where a patient was sent to 10 different MRI centers and it was found that the agreement between the 10 different MRI examinations was very poor. As such, gradings are often accompanied by a measure of reliability between observers or readers, typically radiologists, called inter-observer variability. Similar variability exists for measuring the reliability of a single reader, called the intra-observer variability; comparing readings of the same scans read by the same radiologist. The range of this variability changes from one grading to the other. Thus, when a model learns a set of gradings there is a ceiling of performance which is the intra-variability of the radiologist.

Variation in Imaging. The second challenge is the variability of imaging. Clinical scans from different sources do not possess the same scanning characteristics. This results in a variety of field-of-view, scanning orientation, slice thickness, pixel spacing and others. These differences might result in a different grading e.g. a subject scan via one scanner might exhibit marrow change but there is no guarantee this is true when the same subject is scanned with a different scanner. A model trained to predict from these scans has to be robust to such variations.

1.2 Contributions and Thesis Outline

Datasets & Pre-Processing. Chapter 3. We look at the differences of the three main datasets: **Genodisc**, **TwinsUK**, and **OSCLMRIC**. In addition, we propose a pre-processing pipeline to detect tight bounding volumes of vertebral bodies and their

extent in terms of slices in a sagittal lumbar scan, which we later use for extracting intervertebral disc volumes in both sagittal and axial scans.

Predicting Radiological Gradings. Chapter 4. Here, we explore multiple classifications of radiological gradings from disc volumes. We show that by learning all the gradings simultaneously, in a multi-task manner, we achieve a significant performance boost compared to learning each grading by itself. We also show there exists a single best point to branch out in a multi-task CNN and this might differ depending on modality. We achieve near human performances across all the learnt radiological gradings and state-of-the-art results on some gradings when compared with other comparable automated methods.

Evidence Hotspots. Chapter 5. This chapter presents a method to localize evidence of prediction from a classification CNN; in our case we localize the pathological regions in the intervertebral disc volume corresponding to a unique radiological grading. We compare several methods to produce saliency maps and explored the possibility of producing bounding boxes from these maps. We find that test time augmentation is key in producing high quality evidence hotspots.

Predicting Radiological Gradings using Other Planes, Sequences & Raw MRIs. Chapter 6. In this chapter, we demonstrated that classification of certain gradings can be improved upon by adding either: (i) scans of a different plane e.g. axial scans, or (ii) different sequences e.g. T1-weighted scans. We also look at predicting disc gradings from raw MR volumes which we find to be slightly worse when compared to using disc volumes as input.

Self-Supervision. Chapter 7. We show that it is possible to train a CNN on

essentially free information; longitudinal information in our case, where we train a CNN to learn if a pair of vertebral bodies or intervertebral discs are from the same person. The weights of this CNN is then transferred to classification CNN and we found that performance is better than training a classification CNN from scratch.

1.3 Publications

The research conducted in this thesis has resulted in several peer-reviewed publications. The work of Chapter 3, Section 3.2, on refinement of vertebral body localization and labelling was part of a piece of work presented at the International Workshop on Computational Methods and Clinical Applications for Spine Imaging in MICCAI 2015 (Jamaludin et al., 2015b). The work of Chapter 3, Sections 3.3–3.4, on the vertebral body corner localization and extent of the vertebral body was part of the research on classifying Modic changes which was also presented (oral) at the same workshop (Jamaludin et al., 2015a); this work was awarded the best workshop paper award. The work on radiological grading prediction and evidence hotspots localization in Chapters 4–5, was presented at MICCAI 2016 (Jamaludin et al., 2016) and was awarded the MICCAI Young Scientist Award (awarded to 5 papers in MICCAI 2016) and the MICCAI Student Travel Award. Extensions of this work, also discussed in Chapters 4–5, was published in a Special Issue of Medical Image Analysis (Jamaludin et al., 2017b). The work conducted in Chapters 4–5 has also resulted in a paper which was published in the European Spine Journal (Jamaludin et al., 2017c), and was presented (oral) at the 44^{th} International Society for the Study of the Lumbar Spine Annual (ISSLS) Meeting and was awarded the ISSLS Prize in Bioengineering Science. Finally, the work on self-supervision in Chapter 7 was presented at the International Workshop on Deep Learning in Medical Image Analysis in MICCAI 2017 (Jamaludin et al., 2017a).

Additional Publication. Another paper published during the course of this DPhil is the work on generating videos of talking heads from audio and still images presented (oral) at BMVC 2017 (Chung et al., 2017); co-first author with Joon Son Chung.

Chapter 2

Background

In this chapter, we review suitable literature as background to the research conducted in this thesis. We begin, in Section 2.1, with a brief discussion of the anatomy of the human spine. Then we follow up, in Section 2.2, by discussing commonly used techniques in computer vision used in the analysis of the spine. This includes localization and labelling of anatomical parts and classification/regression of/to radiological gradings or measurements. Then, in Section 2.3, we discuss in brief the progression of deep learning in computer vision and more specifically deep learning in medical image computing. Finally, in Section 2.4, we discuss in brief, the specifics of lumbar MRIs.

2.1 Spinal Anatomy

Anatomical illustrations detailing the anatomy of the human spine can be seen in Figure 2.1.

Vertebrae. A normal human spine contains 33 vertebrae in total, though this number can vary. They are grouped into five main regions corresponding to the curves of the spinal column: cervical, thoracic, lumbar, sacrum and coccyx. There are seven cervical (C1–C7), twelve thoracic (T1–T12), five lumbar (L1-L5), five sacral (S1–S5),

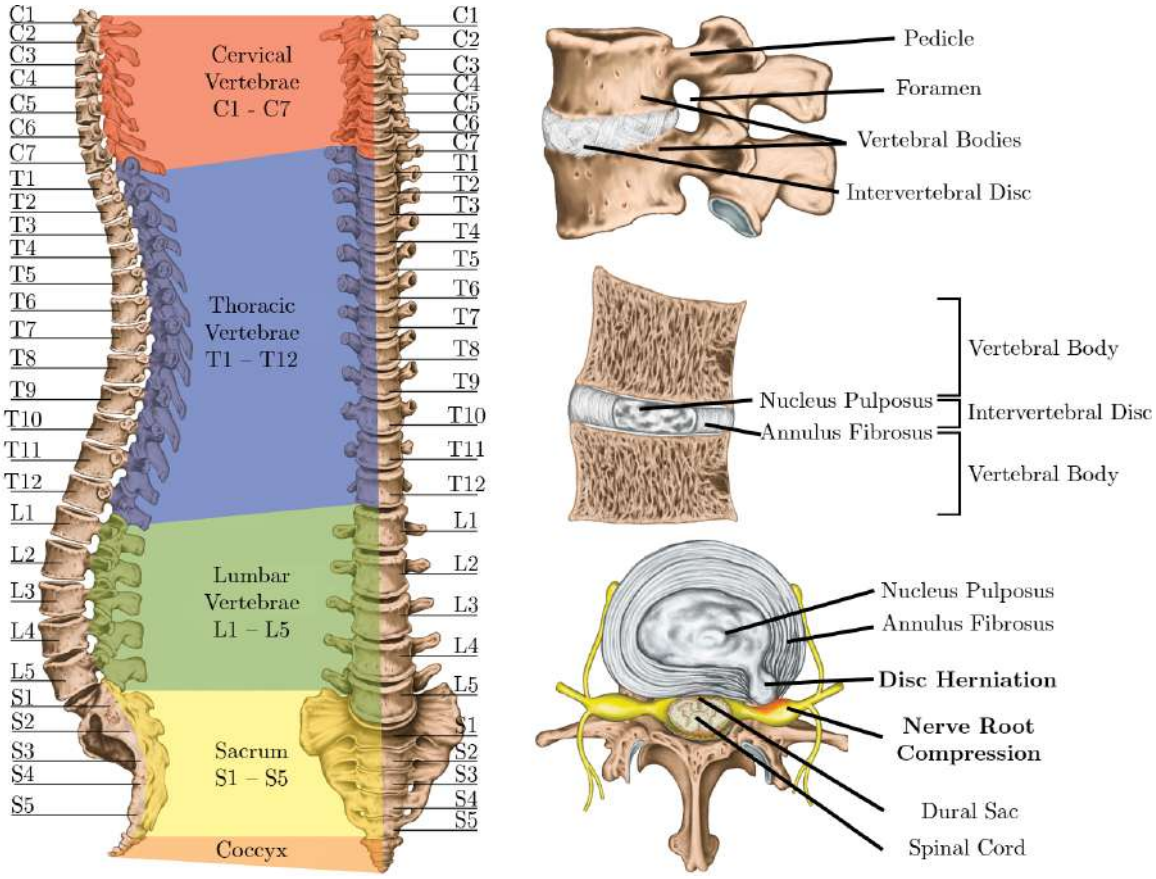


Figure 2.1: ***Illustrations of Spinal Anatomy****. **Left:** Sagittal and front coronal view of the spine. **Top Right:** Overview of an intervertebral disc and its vertebral bodies. **Middle Right:** Mid-sagittal view of an intervertebral disc. **Bottom Right:** Axial view of a disc with disc herniation and nerve root compression. Copyright: stihii/Shutterstock.*

and four coccygeal vertebrae. A single vertebra is made up of two main parts: the vertebral body, and the vertebral arch containing the pedicles and laminae.

Intervertebral Disc. In between a pair of vertebrae are intervertebral discs. They are labelled according to their neighbouring vertebrae e.g. the L5-S1 intervertebral disc is the disc between the L5 and S1 vertebrae. The disc is made up of two main parts: a nucleus pulposus, and an annulus fibrosus. Though these parts are quite distinct, separation of these parts in the MRI is normally unclear. The region connecting the disc and the vertebral body is called the vertebral endplate.

Spinal Cord. The spinal cord is a bundle of nervous tissue normally involved in the transmission of signals from the brain to the body. A specific region of the spine controls a specific motor function e.g the L1-L4 region controls the flexing of the thigh. The spinal cord is surrounded by a membranous sheath called the dural sac.

2.2 Spinal Analysis in Computer Vision

Image analysis of the spine in computer vision is generally confined to the analyses of specific anatomical parts of the spine; with intervertebral discs (IVDs) and vertebral bodies (VBs) being the most popular. As such, more focus is placed on literature regarding IVD and VB analysis in this chapter as they play a bigger part in this thesis compared to other parts of the spine. The general approach of automated analysis can be broken down to two main stages which are (i) the localization step, and (ii) the classification or regression to radiological gradings or measurements.

2.2.1 Anatomical Localization

The current standard approach to localize the anatomical parts is to treat the problem as two main steps: (i) detection, and (i) labelling. This approach to anatomical localization came from a framework that used parts-based model for object recognition by Felzenszwalb and Huttenlocher (2005) and was first adapted for spinal imaging by Schmidt et al. (2007) to detect and label IVDs. The detection of the anatomical parts is necessary as spinal scans or images normally contain multiple elements per image which vary in their positions depending on imaging modality, field-of-view, type of scan and general interpersonal variation. Once detected, they are labelled thus distinguishing one anatomical part from another in a single image. This is especially important in spinal MRIs as VBs at different levels generally look alike.

2.2.1.1 Detection of Anatomical Parts

The detection of the parts are typically done per window basis i.e. via a sliding window detector. Several features have been proposed for the detection.

Correlation Filter. Aslan et al. (2010) proposed a correlation filter to automatically place seeds for VBs segmentation in axial scans. Overall, the technique fails to identify 15% of VBs in their dataset but are visually better than other segmentation of axial VBs.

Haar Wavelet. Huang et al. (2009) and Zhan et al. (2012) both proposed a Haar wavelet based appearance model for detecting anatomical parts. The main difference between the two is that Huang et al. (2009) focus on the appearance of VBs in sagittal scans while Zhan et al. (2012) detect both VBs and IVDs in transverse scans. The only minor issue with the system proposed by Zhan et al. (2012) is that the validation of the system is done on scout scans which are generally less beneficial that normal

scans for clinical diagnosis.

Histogram of Oriented Gradients (HOG). Lootus et al. (2013) and Oktay and Akgul (2013) proposed systems that focused on VBs detection via HOG features. The general results from both are comparable but the system proposed by Lootus et al. (2013) is fully automatic in 3D while the system of Oktay and Akgul (2013) operates in 2D midsagittal slices of the MRI scans.

The outputs of the detection can be bounding boxes which contain the anatomical parts, as in the works done by Huang et al. (2009) and Lootus et al. (2013), segmented region of the parts, shown by Aslan et al. (2010), or point predictions of the inner part of the anatomical parts, proposed by Oktay and Akgul (2013), and Zhan et al. (2012).

2.2.1.2 Labelling of Anatomical Parts

The anatomical parts localization is normally followed with a secondary step to label each part in the image, labelling the bounding boxes with levels of the VBs for example. This is normally done via a graphical model based on the geometrical layout of the spine and the appearance model of the VBs. Glocker et al. (2012) uses sub-volumes information around each vertebra for the appearance model and conditional probabilities of vertebrae positions for the overall layout of the spine which results in 85% lumbar vertebrae identification rate. Similarly, Lootus et al. (2013) uses the classifier score for each part as the appearance model and trained box thresholds that define the pairwise relationship between adjacent vertebrae as the spinal shape priors for lumbar identification. We used the method proposed by Lootus et al. (2013) as the basis of our MRI pre-processing in Chapter 3.

2.2.2 Classification/Regression to Radiological Measurements

With every anatomical parts relevant to the radiological measurements localized and labelled, we can now train a classifier or a regressor for each measurement. Though this is an important step in creating an automatic spinal analysis system, some works do rely on manual or semi-manual localization. The most important step prior to classification is feature extraction. There are only two types of information regarding the anatomical parts that are important for learning. These two types are shape information, e.g. height of the IVD, and signal information, e.g. the intensity variation of the VB. These methods of defining features by hand is now called shallow learning which is now superseded by automatically learning the features implicitly through the labels themselves via deep learning. We discuss these so-called shallow features here and we will discuss deep learning in next Section.

Human Measurements. Abbati et al. (2017) proposed two methods in their work to diagnose lumbar spinal stenosis. One of their methods is to regress to labels of stenosis via features marked up by radiologists. These features include quantitative features e.g. area of the spinal canal and qualitative features e.g. Pfirrmann gradings, Modic changes etc.

Shape. For herniation, Alomari et al. (2014) proposed a contour shape feature describing the level of herniation of the discs with a reported accuracy of 93.9%. However, the level of disc herniation is simplified to be a binary problem i.e. normal and herniated disregarding classes in between that are beneficial for clinical analysis. Similarly, for Scoliosis, Shen et al. (2014) proposed a geometric torsion feature for spinal deformities classification in bi-planar X-rays. The general idea of using parametric curve fitting works well for spinal deformities detection. Koompairojn et al. (2010) also described a shape-based feature but for spinal stenosis by using diame-

ters and lengths of various landmarks surrounding the spinal canal as features. They achieved an accuracy of 92.7% for central stenosis and 96.3% for lateral stenosis classifications. However, improvements can be made in their intensity-based localization of the spinal canal which is not robust.

Signal. Koh et al. (2012) proposed a signal based feature based on the ratio of disc, vertebrae and spinal cord pixels within the ROI to classify herniation with an accuracy of 99%. The whole process is not automatic but works really well with supervised human inputs for the ROI. However, the exclusion of shape features for a problem that is visually closely related to shape i.e. disc herniation, is questionable.

Both Shape and Signal. Lootus et al. (2014) proposed a method that utilised the histogram of signal information of the IVD alongside the IVD height as features to regress Pfirrmann grading with an accuracy of 85.9%. Ghosh et al. (2011) proposed a spatially-binned intensity feature couple with a ratio width and height of the IVD for herniation classification with a reported accuracy of 94.9%.

Overall, the current approach for classification and regression to radiological measurements is to fine tune hand-crafted features for specific problems. This is time consuming, especially if there are multiple gradings to classify, and is prone to failure from bias i.e. not robust. Robustness here means little to no degradation of performance when exposed to newer inputs.

2.3 Deep Learning

We heavily used Convolutional Neural Networks (CNNs), which is why we include literature concerning CNN as background. CNN is a type of artificial neural network

(ANN) designed to work on data that are spatially connected e.g. an image, a video, an audio file and is a part of the increasingly popular subset of machine learning now called deep learning (see LeCun et al. (2015)). A standard CNN is made up of five different layers: convolutional, pooling, dropout, non-linearity (typically ReLU) and fully-connected layers. In training, sets of weights or convolutional filters are learned such that the input becomes linearly separable in the last layer of the CNN.

2.3.1 CNNs in Natural Images

Various papers exist that use CNNs as their main tools for solving computer vision problems. Papers reviewed here are chosen based on the fact that they are in some way relevant to the research in the subsequent chapters.

Classification. The AlexNet architecture, first introduced by Krizhevsky et al. (2012), is one of the most well known CNN architecture for image classification. They achieved state-of-the-art performance on both the classification and localization tasks in the ImageNet Large Scale Visual Recognition Challenge (ILSVR) Russakovsky et al. (2015) back in 2012. They are among the first to utilize dropout as a mean of regularization in CNNs which is now standard in current and newer CNN architectures. Besides dropout, the authors also emphasized the importance of data augmentation to reduce overfitting. Their success boosted the popularity of deep learning in vision tasks resulting in deep-learning methods becoming the majority in ILSVR2013 and ILSVR2014. In ILSVR2013, another CNN-based method, by Zeiler and Fergus (2013), won the classification task, this time introducing a novel method to visualise the learned filter essentially using a deconvolutional network to help them choose the better architecture for the classification task. Similarly, in ILSVR2014, the top two entries were CNN-based, namely: 1) VGGNet by Simonyan and Zisserman (2015) where they implemented a network with really small (3×3) convolutional

layers and increased the depth of the network to a maximum of 19 weight layers, and 2) GoogLeNet by Szegedy et al. (2015) where they also implemented a really deep, 22 weight layers, CNN network with a new idea of using network-in-network inception modules which is essentially a better organisational approach to designing deep networks. Current best architectures include the ResNet family of networks by He et al. (2015) and the Squeeze-and-Excitation networks by Hu et al. (2017).

Segmentation. One of the best methods for semantic segmentation is the one proposed by Long et al. (2015) which is CNN-based. Their main contribution is the notion of fully convolution network for semantic pixelwise classification. The primary contribution is the novel notion of pooling shallow layers, layers close to the input image, with spatially "fine" features with "coarse" features from deeper layers. An interesting note is that they notice no improvement to the segmentation accuracy when data augmentation is added to for training which is unlike the result of Krizhevsky et al. (2012) for image clasification. Overall, they achieved a mean region intersection over union (IU) of 62.2% on the PASCAL VOC 2012 dataset. In comparison, the closest performing method, proposed by Hariharan et al. (2015) which uses the concept of per-pixel neuron activations in the overall CNN called hypercolumns to be the pixel descriptor, only managed a mean IU of 54.6% on the same dataset. Unlike classification which normally has datasets constructed to be balanced in labelled samples, segmentation networks have been shown to work better with a balanced loss, pixel-wise, as shown by Badrinarayanan et al. (2017). We adapted this notion of balancing via the losses as our data or any data in medical imaging are normally unbalanced.

Localization via Classification. Over the years several methods have been developed for visualizing saliency maps of predictions from CNN models trained on

classification tasks. In this work, we focus on three methods: (i) error backpropagation by Simonyan et al. (2014), which computes the gradient of the class score prediction with respect to the input image via backpropagation; (ii) guided backpropagation by Springenberg et al. (2015), which changes the backward pass of ReLU in a similar manner to the 'deconvnet' method by Zeiler and Fergus (2014); and, (iii) excitation backpropagation by Zhang et al. (2016), which changes the gradients of the convolutional and average pooling layers. Another method that works extremely well that is not covered in this work is the saliency via class activation maps (CAM) by Zhou et al. (2016) which is applicable principally to later layers in the CNNs (it is usually applied to the global average pooling prior to the final fully-connected layer). Discussions of what the visualizations mean and how they can be changed by architectural choices in the CNN models can be seen in the work by Mahendran and Vedaldi (2016).

2.3.2 CNNs in Medical Images

CNNs have started to be used more and more in medical vision problems and there exists in two main ways to start training a CNN: either (i) using transfer-learning ; or (ii) training a network from scratch.

Transfer Learning. This idea of reusing networks trained on other tasks is called transfer learning. Transfer learning in CNNs has been found to be extremely effective especially when the model has been pre-trained on a large dataset like ImageNet (see Deng et al. (2009)), and there have been several successes on using models pre-trained on natural images on medical images e.g. lung disease classification Shin et al. (2016) and to detect skin cancer Esteva et al. (2017). However, there are several differences between natural and medical images that makes using a network trained of natural images, e.g. ImageNet, less desirable for transfer learning to medical-based tasks. The

first major difference is the structure of data; medical images can be volumetric in nature but natural images tend to be made up of 3 colour channels. Another difference is the dynamic range, which is normally 24-bit for a natural image (8-bit per channel) but this is different for a medical image which can vary substantially e.g. 12-bit for CT scans and 12-bit or 16-bit for MRIs. Perhaps the most substantial difference is that medical images are very specific in appearance for each use case i.e. a dataset for lumbar MRIs is very homogeneous in appearance unlike ImageNet. This variation in the statistics of the data would mean it is appropriate to experiment transfer learning with scans of the same modality rather than using ImageNet-trained models.

Training from Scratch. We find that if the dataset is sufficiently large enough, it is almost always better to train from scratch. There is another way to mitigate this problem regarding the amount of data especially for segmentation tasks which is to treat the problem as a patch-based problem. One of the first successful application of this is the work by Cireşan et al. (2013) which won both MICCAI 2013 and ICPR 2012 mitosis detection challenges and the ISBI 2012 brain segmentation challenge. All three challenges were won using the same method which is one of the earliest CNN architectures that started using max-pooling layers as a mean to deal with translational variance of the input image; see Cireşan et al. (2012, 2013). Their method involves using a sliding window to get small image patches which are used as inputs to the CNN. The centre-most pixel is then classified as a certain class resulting either a boundary mask as segmentation output or a probability map for a bounding box prediction. Following up on that, Ronneberger et al. (2015) proposed a new network architecture comprised of both convolutional and deconvolutional filters, with skip connections connecting these two parts together, for segmentation in medical images. This is different to the proposed architecture of Long et al. (2015), where instead of pixel-wise classification, the CNN produces the segmentation map

via deconvolution. This results in a smoother segmentation mask but loses details if the segmentation region is small. They also introduced this notion of extremely aggressive training augmentation regime through elastic augmentation, though this requires strong (pixel-level) supervision. Since we do not have segmentation masks in any of our datasets, we find that adding extra tasks acts as extra supervisory signals which aide training a CNN.

Multi-tasking. The idea of multi-tasking in deep learning has been explored in both natural images e.g. UberNet by Kokkinos (2016) and medical images e.g. Moeskops et al. (2016). In UberNet, one single network was trained on 7 different tasks ranging from detection to semantic segmentation. Kokkinos (2016) however found that training on multiple tasks proved to be detrimental as the overall performance is lower than a single task model. Moeskops et al. (2016) however found that training a network on three separate segmentation tasks with three different modalities to be as good as training a single network. It can be hypothesized that multi-tasking is only useful when both the tasks and the modalities used in training are quite related.

Self-Supervision. Another solution to low amount of data for training is to actually learn from the data itself, normally called self-supervision. Models trained using only information contained within an image as a supervisory signal have been proven to be effective feature descriptors e.g. Doersch et al. (2015) showed that a CNN trained to predict relative location of pairs of patches, essentially learning spatial context, is better at a classification task (after fine-tuning) then a CNN trained from scratch. We use a somewhat related task, where we predict where each input lies in the overall scan (the level of the VBs) in Chapter 7. The task of learning scans from the same unique identity is related to slow-feature learning in videos Mobahi et al. (2009); Wang and Gupta (2015); Wiskott and Sejnowski (2002). One natural

images example is the SyncNet by Chung and Zisserman (2016) where they trained a network to align video and audio signals trained only on artificially desynchronized samples; no hard ground truth labels. Similarly, the simplest way to pre-train a network can be viewed as a self-supervision task as Abbati et al. (2017) where they first train an auto-encoder to generate the input and uses the intermediate encoded output as input to another classifier for grading lumbar spinal stenosis.

2.4 Lumbar MRIs

There are two main types of MRI sequences used in this thesis namely, T1-weighted and T2-weighted MRIs. In short, T1-weighted scans are produced using short repetition and echo times, TR and TE respectively, while T2-weighted scans are produced using longer TR and TE. These distinctions in acquisition times cause visual differences of the tissues in both sequences. In lumbar scans, this means the spinal canal appears brighter in T2-weighted scans than T1-weighted scans. Similarly, inflammations appear darker in T1 and brighter in T2 e.g. marrow changes of the vertebral bodies. These differences result in variation of scanning protocols across difference centres for acquisition of MR scans though T2-weighted scans tend to be more popular.

Chapter 3

Datasets & Pre-Processing

In this chapter, we describe the datasets that will be used in the subsequent chapters, in Section 3.1, followed by the pre-processing pipeline to detect tight bounding volumes of vertebral bodies, in Sections 3.2–3.3, and the extent of the vertebral bodies, in Section 3.4.

3.1 Datasets

There are three main datasets used in this thesis: (i) **Genodisc**, (ii) **TwinsUK**, and (iii) **OSCLMRIC**. Though they are all datasets containing spinal lumbar MRIs, they were sourced from different centers and thus contain variations in the acquisition protocol which we will delve into, dataset to dataset.

3.1.1 Genodisc

The **Genodisc** dataset consists of 2635 subjects diagnosed with back pain. Of the 2635, only 2079 were assessed and scored by a radiologist and of those 2079, only 2009 possess readable scans. Each scan contains up to six lumbar discs per subject, with the majority having a complete field-of-view of the lumbar region and includ-

ing six discs. In total there are 12018 discs, from 2009 subjects, each possessing several radiological gradings. The dataset was collected by the Genodisc Project in 2009 (`www.physiol.ox.ac.uk/genodisc`). **Genodisc** recruited "patients who seek secondary care for their back pain or spinal problem", and recorded their MR scans and clinical information. The scans were sourced from multiple centres: England (Oxford, Oswestry, Kettering), Hungary (Budapest), Slovenia (Ljubljana) and Italy (Milan). They came from different machines, and were acquired with a variety of acquisition protocols following the clinical standard of each centre. Because of this, differences in the scans can be quite noticeable e.g. the slice thickness of the sagittal scans may vary substantially. All the subjects were scanned in the supine position, lying on their backs. 54.6% of the subjects are females. Some details of the subjects and their scans can be see in Table 3.1.

	Mean	**Median**	**Range**
Age	50.1	50.0	14.0 – 87.0
Weight	77.8	80.0	43.0 – 130.0
Field Strength (Tesla)	1.3	1.5	0.2 – 3.0
Sagittal Slice Thickness (mm)	4.1	4.0	1.2 – 6.0
Sagittal Pixel Spacing (mm)	0.7	0.6	0.3 – 1.6
Sagittal Slice Gap (mm)	4.8	4.7	0.0 – 8.8
Sagittal Slice Count	12.3	11.0	8.0 – 25.0
Axial Slice Thickness (mm)	4.5	4.0	1.3 – 10.0
Axial Pixel Spacing (mm)	0.5	0.5	0.2 – 1.5
Axial Slice Gap (mm)	5.3	5.0	0.0 – 46.8
Axial Slice Count	12.5	12.0	9.0 – 22.0

Table 3.1: ***Subject & Scan Details – Genodisc.*** *Details of the subjects and scans in the* ***Genodisc*** *dataset that possess MRIs and were read by a radiologist.*

A set of scans, or each series, from a subject contains multiple type of scans: T1-weighted axial, T1-weighted sagittal, T2-weighted axial and T2-weighted sagittal. The axials scan are normally only available for the bottom three lumbar discs, L3-L4, L4-L5, and L5-S1. The slices of the axial scans can be obtained in two different orientations (see Figure 3.1): (i) in blocks where all axial slices are parallel to each

other, or (ii) per disc basis where slices pertaining to each disc are aligned to be parallel to the disc orientation. T2-weighted sagittal scans are used the most in this thesis as most radiological gradings for the spine can be assessed via T2-weighted sagittal scans, and these often emphasize clinically popular gradings compared to T1-weighted scans.

3.1.1.1 Radiological Gradings

The scans were assessed by a single expert spinal radiologist, Professor Iain McCall, who produced a range of global, i.e. the whole spine, and local, i.e. per disc, gradings. Overall, there are 16 different radiological scores where each one has its own specific scale of measure as shown:

1. **Pfirrmann grading [S]:** 1 – 5 (Pfirrmann et al. (2001))
2. **Disc narrowing [S]:** 0 – 3 (normal, slight, moderate, severe)
3. **Annular tears (HIZ) [S & A]:** 0 – 2 (absent, present, contiguous)
4. **Anterior disc bulging [S & A]:** 0 – 3 (normal, slight, moderate, severe)
5. **Posterior disc bulging [S & A]:** 0 – 3 (normal, slight, moderate, severe)
6. **Disc herniation [S & A]:** 0 – 3 (normal, slight, moderate, large)
7. **Location of herniation [S & A]:** Central (**C**), Posterolateral R or L (**PLR** or **PLL**), Foraminal R or L (**FR** or **FL**), and Other (**O**)
8. **Type of herniation [S & A]:** 1 – 3 (protrusion, extrusion, sequestration)
9. **Nerve root compression (herniation) [A]:** 0 – 3 (none, touching, displaced, compressed)
10. **Foraminal stenosis [S & A]:** 0 – 3 (absent, mild, moderate, severe)

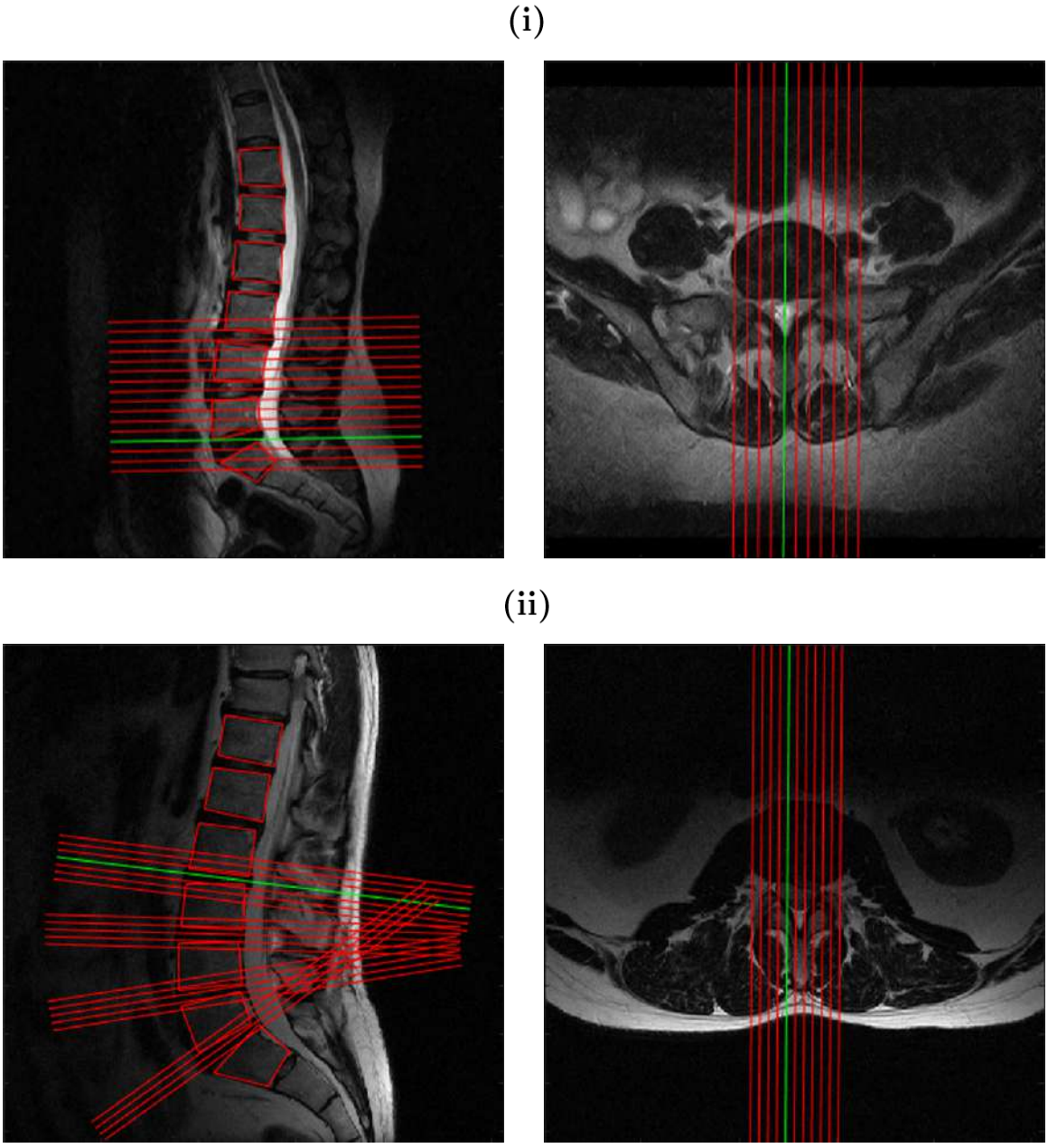


Figure 3.1: ***Examples of T2-weighted Axial & Sagittal Scans****. The two series of scans have different orientations of axial slices here overlaid on top of each corresponding mid-sagittal slice in the same series.* ***(i)*** *Axial slices oriented in blocks which is more common in* ***Genodisc****, and* ***(ii)*** *axial slices oriented in parallel to each disc, this case possesses axials of four lumbar discs instead of three which is the common amount for other subjects in* ***Genodisc****.*

11. **Nerve root compression (foraminal) [A]:** 0 – 3 (none, touching, displaced, compressed)

12. **Central canal stenosis [S & A]:** 0 – 3 (absent, mild, moderate, severe)

13. **Spondylolisthesis [S]:** 0 – 4 (0%, 25%, 50%, 75%, 100% of vertebral body sagittal plane width)

14. **Endplate defects [S]:** 0 – 3 (normal, slight, moderate, severe)

15. **Modic changes [S]:** 1 – 3 (Modic et al. (1988))

16. **Facet joint arthropathy [S & A]:** 0 – 3

Most of the scores were acquired by assessing both sagittal and axial scans **[S & A]**, five needed only sagittal **[S]**, and two needed only axial **[A]** scans. Out of these 16 gradings, we experimented on the following gradings: **Pfirrmann grading** (Figure 3.3), **disc narrowing** (Figure 3.4), **central canal stenosis** (Figure 3.5), **spondylolisthesis** (Figure 3.6), **endplate defects** (Figure 3.7), **modic changes** or **marrow changes** (Figure 3.8) in Chapter 4 and Chapter 5, with the addition of **anterior disc bulging** (Figure 3.9), **posterior disc bulging** (Figure 3.10), and **disc herniation** (Figure 3.11) in Chapter 6. **Endplate defects** and **marrow changes** are further divided into a classification of lower and upper endplate gradings such that the two gradings can be viewed as four separate binary classification tasks. Figure 3.2 gives the grade distribution of the radiological gradings. Since severe gradings are extremely rare, all the gradings used in Chapter 4 are binarized to just two classes i.e. normal and abnormal except for **Pfirrmann grading** and **disc narrowing**. We, also experimented with binarized classes for **anterior disc bulging**, **posterior disc bulging**, and **disc herniation** in Chapter 6 as the numbers for all the pathological cases, grades 1 to 3, are low.

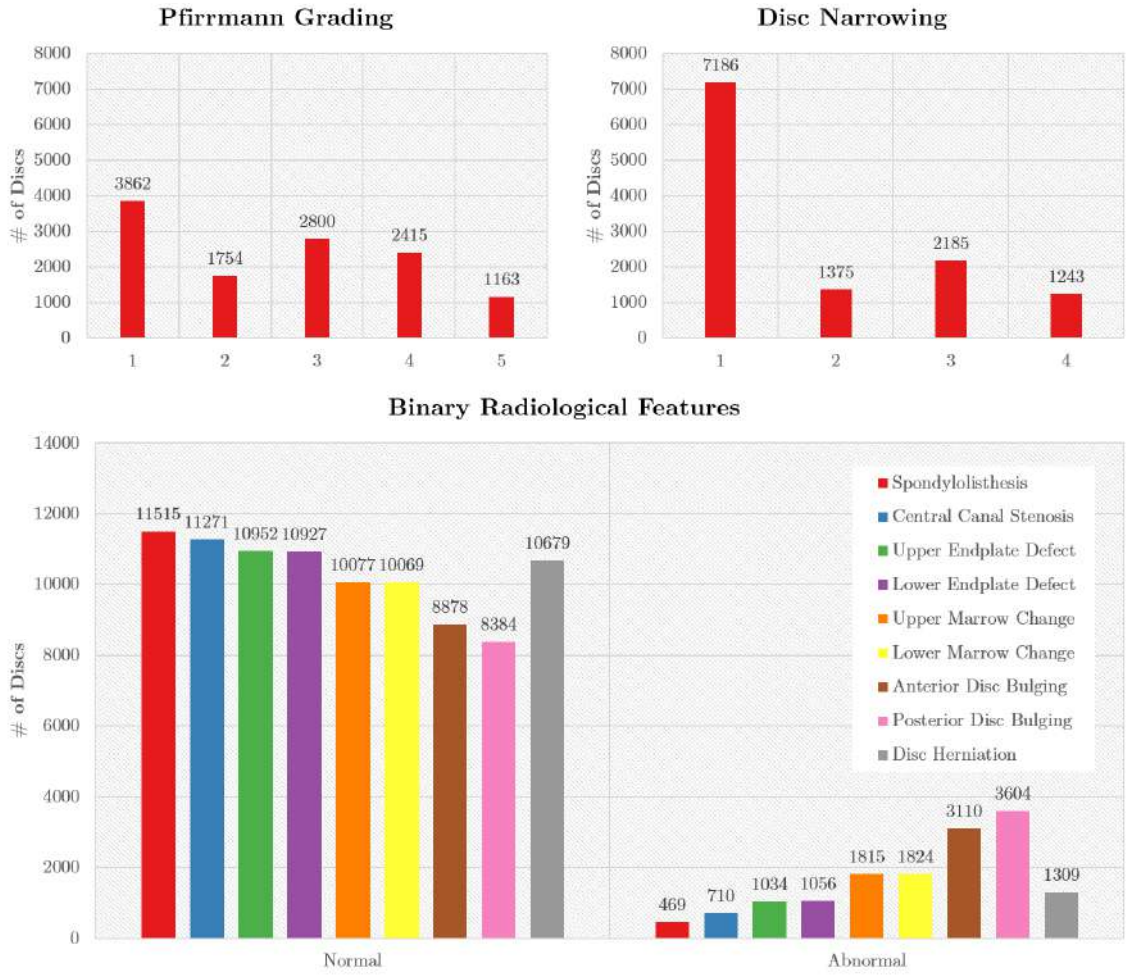


Figure 3.2: ***Distribution of The Radiological Gradings (per disc).*** *Both* ***endplate defects*** *and* ***marrow changes*** *have two separate gradings, one each for both the upper and lower endplate regions. Note, there is a total of 12018 discs but since there are missing labels (independent of grading), the totals of labelled discs shown in the table for each grading are different.* ***Pfirrmann grading*** *and* ***disc narrowing*** *are multi-class while the other gradings are binary.*

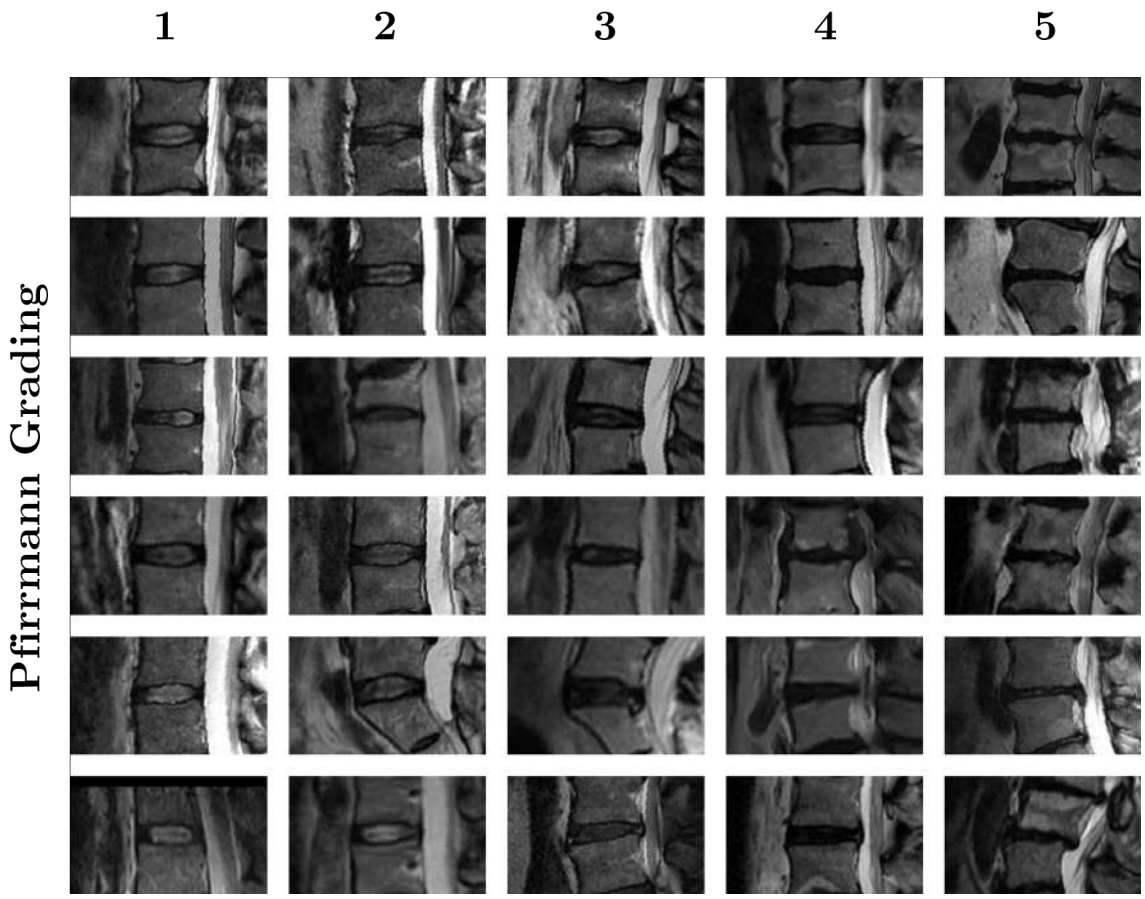


Figure 3.3: ***Examples of Pfirrmann Grading.*** ***Pfirrmann grading*** *is a grading system of disc degeneration using criteria of disc signal heterogeneity, brightness of the nucleus and disc height; 5 grades.*

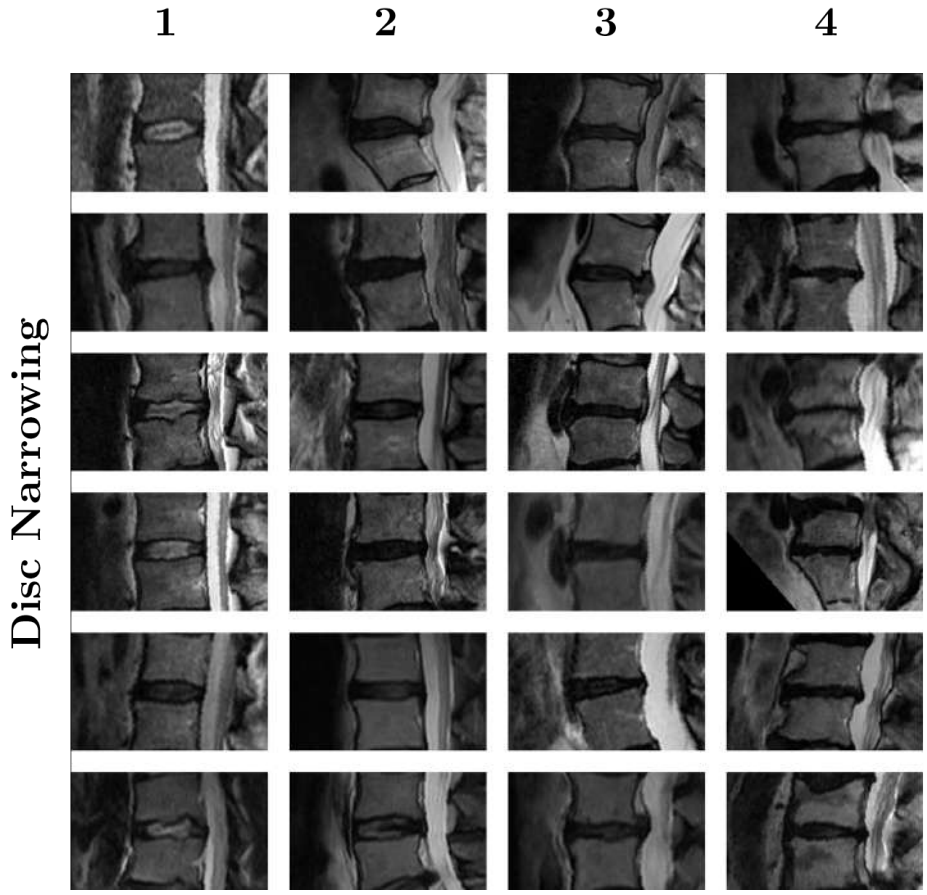


Figure 3.4: ***Examples of Disc Narrowing.*** ***Disc narrowing*** *is defined as a multi-class measurement of the disc heights; 4 grades.*

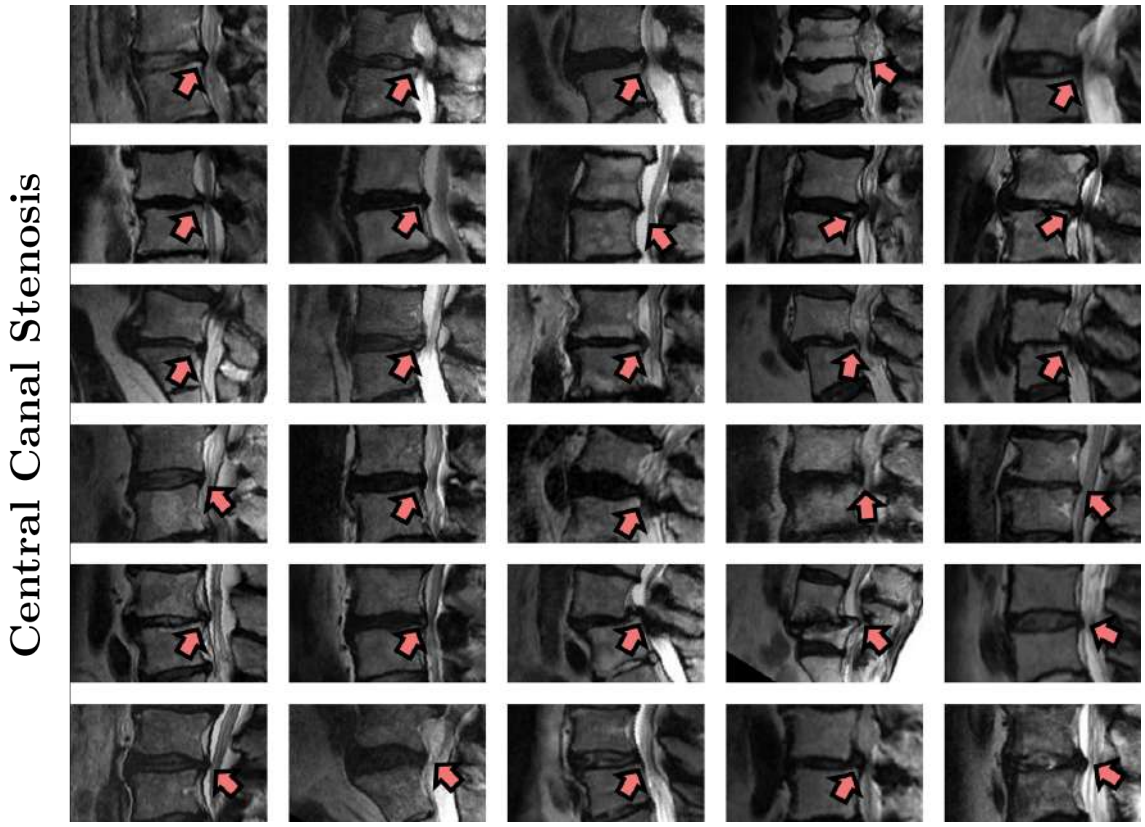


Figure 3.5: ***Examples of Central Canal Stenosis***. ***Central canal stenosis*** *is the constriction of the central canal in the region adjacent to each intervertebral disc. The grading is based on assessment of both sagittal and axial images. We only look at a binary 'presence' or 'absence' of stenosis in the sagittal scans.*

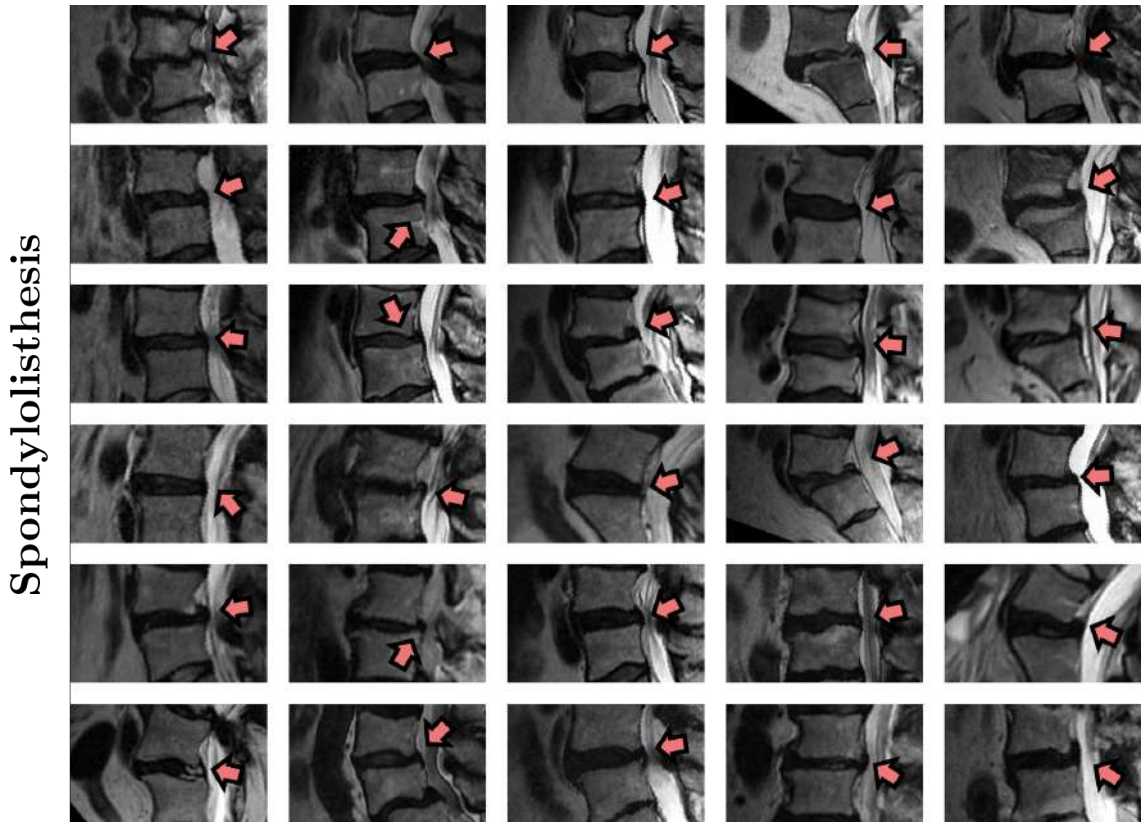


Figure 3.6: ***Examples of Spondylolisthesis***. ***Spondylolisthesis*** *is a binary measure of the vertebral slip; i.e. is the pair of vertebrae above and below a disc in-line or has it slipped?*

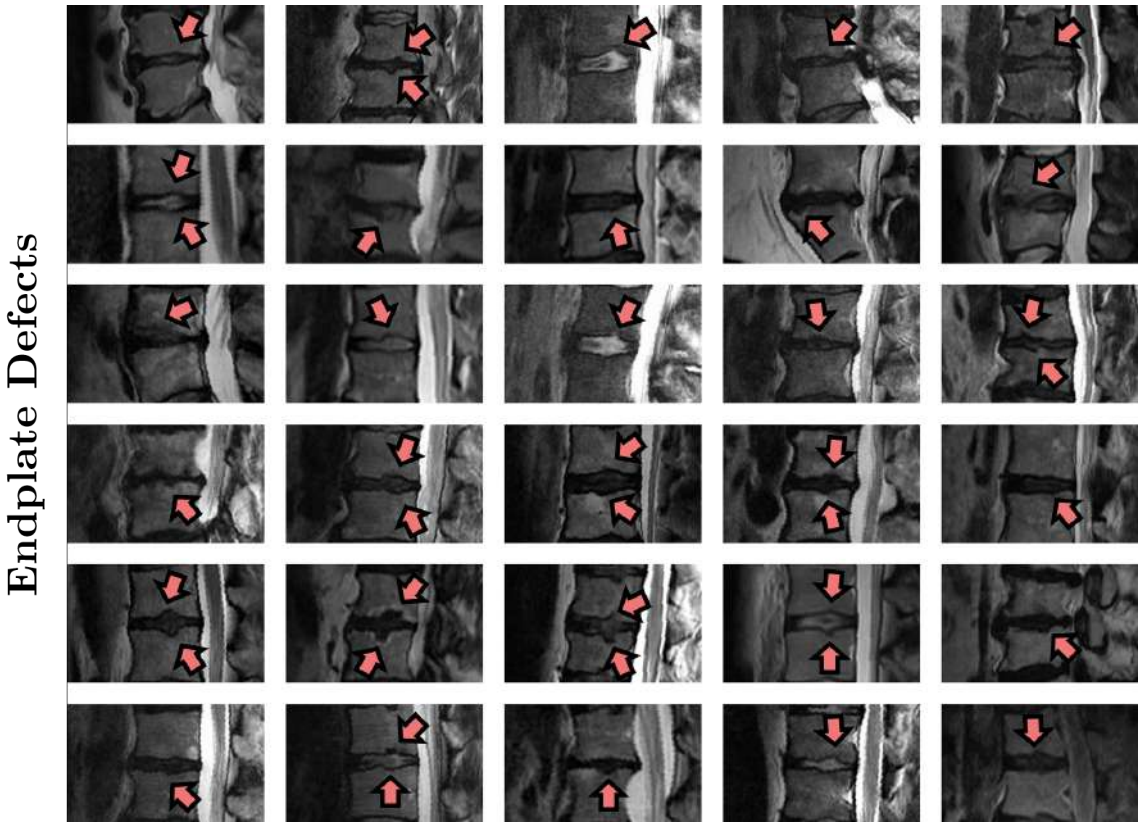


Figure 3.7: ***Examples of Endplate Defects***. ***Endplate defects*** *are deformities of the endplate regions, both upper and lower, with respect to the intervertebral disc.*

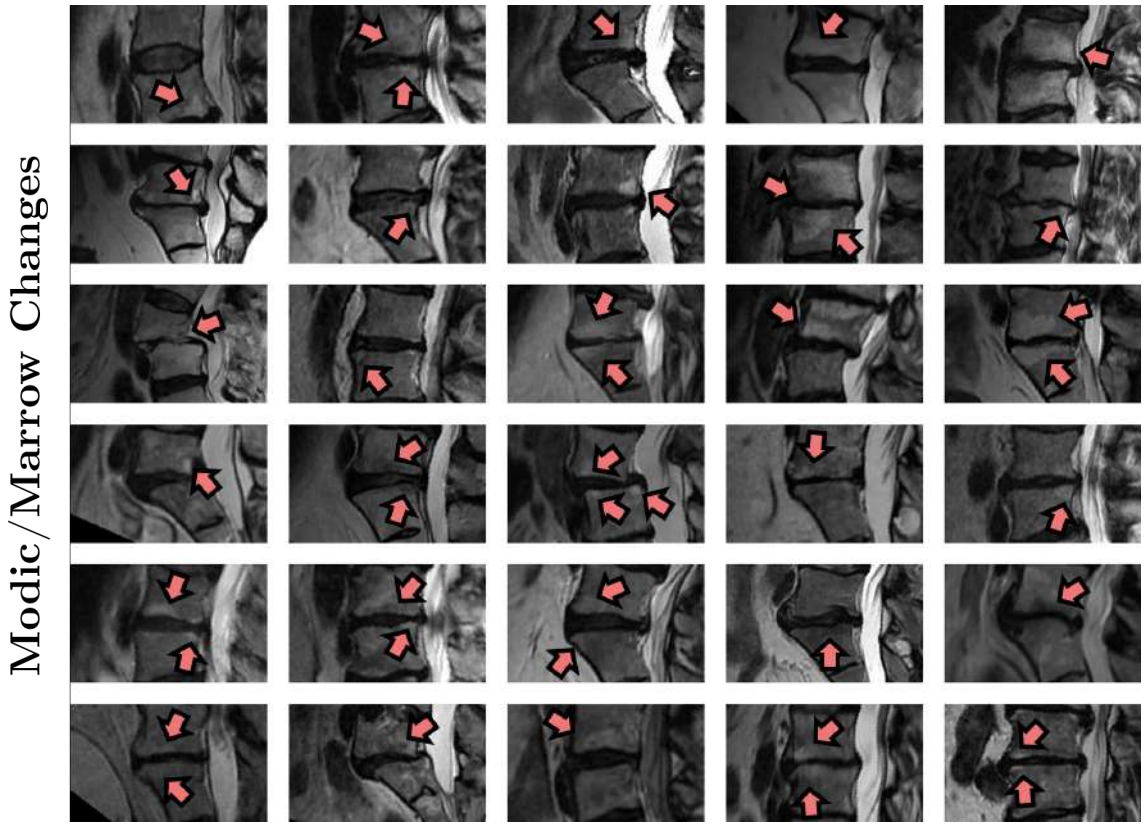


Figure 3.8: ***Examples of Marrow Changes***. ***Marrow changes*** *are visible signal variations that occur in the vertebral endplate regions, upper or lower. It can show up as both hyper- or hypo-intense signals of the vertebral bodies.*

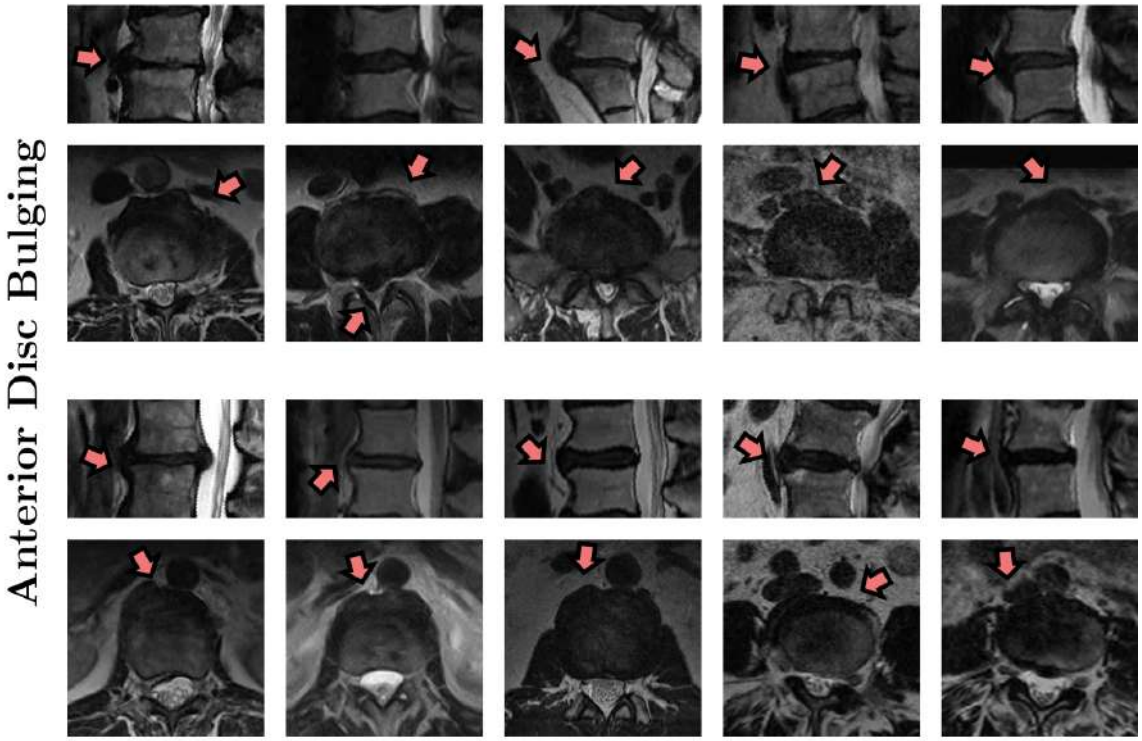


Figure 3.9: ***Examples of Anterior Disc Bulging***. ***Anterior disc bulging*** *is the protrusion of the frontal part of the intervertebral disc. Normally graded by viewing both sagittal and axial scans. The top row is the sagittal scan and the bottom is the axial view of the same disc.*

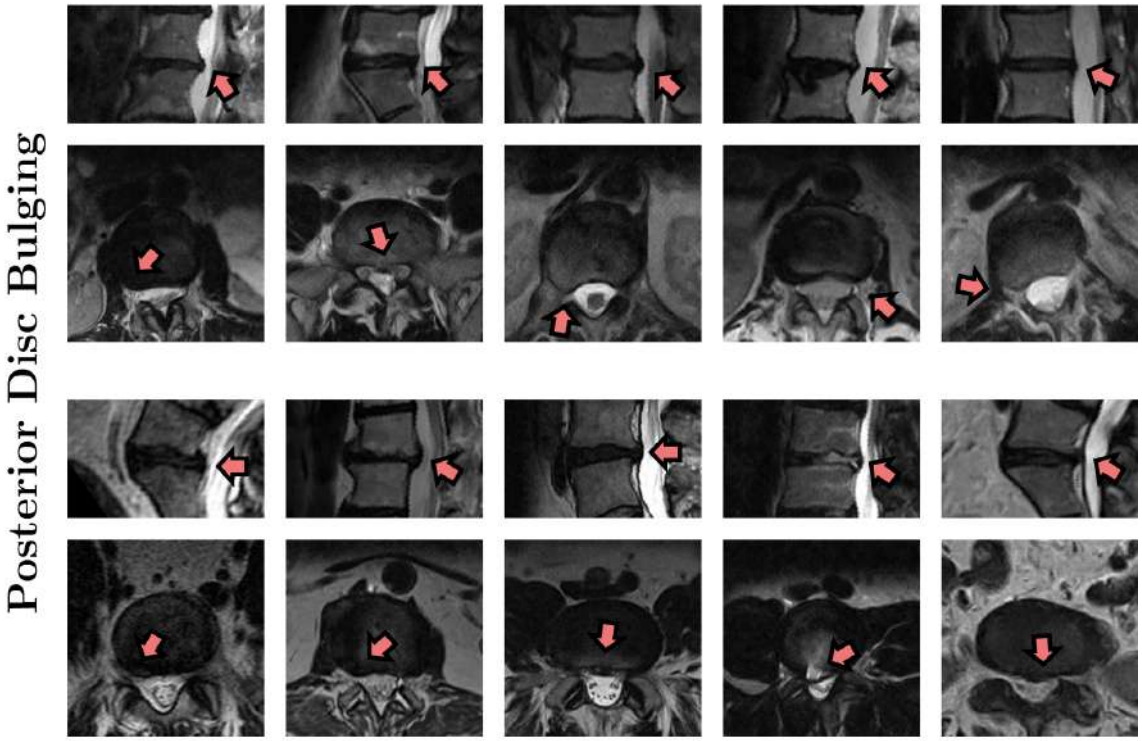


Figure 3.10: ***Examples of Posterior Disc Bulging***. ***Posterior disc bulging*** *is the protrusion of the back of the intervertebral disc. Normally graded by viewing both sagittal and axial scans. The top row is the sagittal scan and the bottom is the axial view of the same disc.*

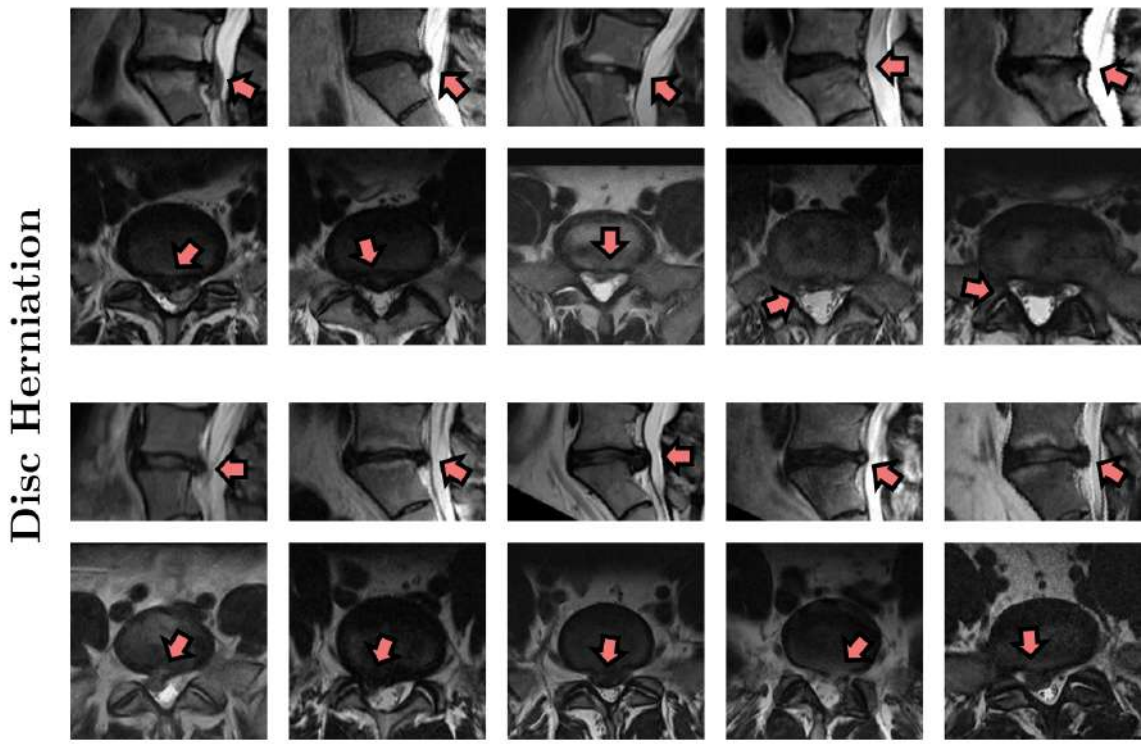


Figure 3.11: ***Examples of Disc Herniation.*** ***Disc herniation*** *is similar to* ***posterior disc bulging*** *in Figure 3.10 with a minor distinction in that the annulus of the disc is ruptured and there is displacement of disc material in a herniated case. Normally graded by viewing both sagittal and axial scans. The top row is the sagittal scan and the bottom is the axial view of the same disc.*

3.1.2 TwinsUK

The **TwinsUK** dataset was collected by the Department of Twin Research, King's College London (`http://www.twinsuk.ac.uk`). The dataset consists of 1016 subjects. Unlike the **Genodisc** dataset which was limited to patients with back pain, the **TwinsUK** dataset collected subjects without any exclusionary criteria. Each scan in **TwinsUK**, contains up to five lumbar discs per subject from L1-L2 to L5-S1, missing the T12-L1 disc in **Genodisc**. In total there are 6327 discs (with gradings), from 920 subjects, with radiological gradings. The scans were collected in a more consistent manner than **Genodisc**.

As well as a baseline scan for each subject, 423 subjects have follow-up scans (circa 2007-2009) taken 8-12 years after the original baseline (circa 1995-2000). A majority of the subjects with the follow-up scans have only two scans (one baseline and one follow-up) while a minority have three (one baseline and two follow-ups). The baseline scans were taken with a 1.0-Tesla scanner while the follow-up scans were taken with a 1.5-Tesla machine but both adhered to the same scanning protocol (slice thickness, times to recovery and echo, TR and TE). Although, the baseline and follow-up came from two different machines but due to this enforcement of similar scanning protocol, differences in the scans are quite small. Similar to **Genodisc**, the subjects were scanned in the supine position. 96.7% of the subjects are females. Some details of the subjects and their scans can be see in Table 3.2. Only T2-weighted sagittal scans were collected for each subject.

3.1.2.1 Radiological Grading

The **TwinsUK** subjects were graded with a measure of **disc degeneration** (Figure 3.12 shows some examples), not dissimilar to **Pfirrmann grading**; graded from 1 to 4 according to severity, instead of 1 to 5 in **Pfirrmann grading**. The gradings were annotated by a clinician and were done on a per disc basis: from L1-L2 to

	Mean	Median	Range
Age	56.1	56.0	19.0 – 79.0
Weight	67.6	65.0	44.0 – 128.0
Field Strength (Tesla)	1.1	1.0	1.0 – 1.5
Sagittal Slice Thickness (mm)	4.0	4.0	4 – 5
Sagittal Pixel Spacing (mm)	0.9	0.9	0.8 – 1.2
Sagittal Slice Gap (mm)	4.4	4.4	0.4 – 5.5
Sagittal Slice Count	11.0	11.0	7.0 – 16.0

Table 3.2: ***Subject & Scan Details – TwinsUK.*** *Details of the subjects and scans in the* ***TwinsUK*** *dataset that possess MRIs and disc gradings.*

L5-S1 discs (5 discs per subject). 920 of the 1016 subjects possess gradings of **disc degeneration**. We conducted experiments using **disc degeneration** in Chapter 7.

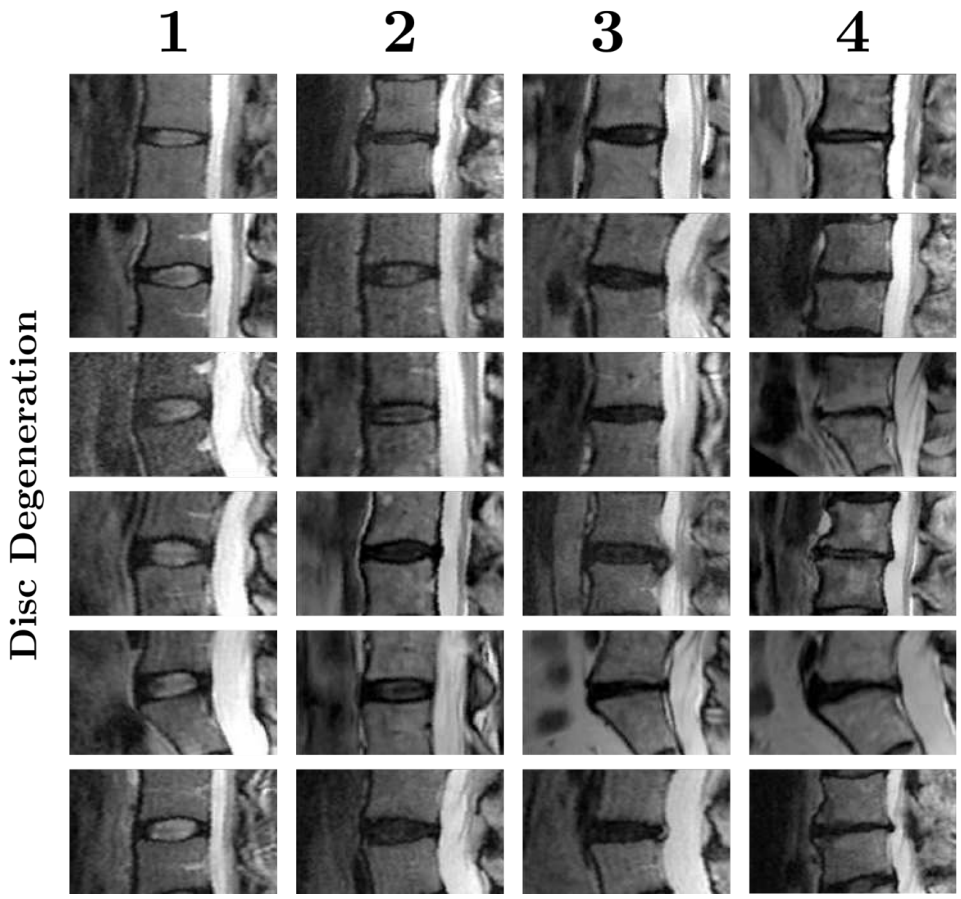


Figure 3.12: ***Examples of Disc Degeneration. Disc degeneration*** *in* ***TwinsUK*** *is similar to* ***Pfirrmann grading*** *in* ***Genodisc*** *(Figure 3.3). Grade one discs are normal or healthy discs with little to no degeneration and grade two to four are degenerated disc, with the highest level of degeneration at grade four.*

3.1.3 OSCLMRIC

OSCLMRIC or Oxford Secondary Care Lumbar MRI Cohorts is a collection of lumbar MRI scans from 1689 patients. Unlike **Genodisc**, where subjects commonly only possess a single series of scans, subjects in **OSCLMRIC** may contain more than a single series, similar to **TwinsUK**. However, the number of series of scans in **OSCLRMRIC** are not constrained to just two or three. 530 subjects (that have T2 sagittal scan) have more than one scans; 366 possessing 2 scans, 104 with 3 scans, 35 with 4 scans, 11 with 5 scans, 8 with 6 scans, and 6 subjects with more than 6 scans. Radiological gradings of the scans are limited to just reports by the radiologists during the clinic; 83 radiologists in total. 57.8% of the subjects are females. Like **Genodisc**, there is no standardization of acquisition protocols (Table 3.3). There are also multiple type of scans: T1-weighted axial, T1-weighted sagittal, T2-weighted axial, T2-weighted sagittal and other sequences e.g. STIR, TIRM, and SPIR in both axial and sagittal views.

	Mean	**Median**	**Range**
Age	51.8	51.0	16.0 – 92.0
Weight	81.7	80.0	31.8 – 163.0
Field Strength (Tesla)	1.5	1.5	0.4 – 3.0
Sagittal Slice Thickness (mm)	3.9	4.0	1.0 – 6.0
Sagittal Pixel Spacing (mm)	0.7	0.7	0.4 – 1.3
Sagittal Slice Gap (mm)	4.5	4.4	0.1 – 7.0
Sagittal Slice Count	14.7	15.0	11.0 – 24.0
Axial Slice Thickness (mm)	4.1	4.0	3.0 – 7.0
Axial Pixel Spacing (mm)	0.6	0.6	0.3 – 1.1
Axial Slice Gap (mm)	4.5	4.4	0.0 – 21.1
Axial Slice Count	19.3	17.0	3.0 – 80.0

Table 3.3: ***Subject & Scan Details – OSCLMRIC.*** *Details of the subjects and scans in the* ***OSCLMRIC*** *dataset.*

3.1.3.1 Oswestry Disability Index

Unlike **Genodisc** and **TwinsUK**, subjects in **OSCLMRIC** not only possess MRI scans but also a measure of disability called the Oswestry Disability Index (**ODI**) by Fairbank and Pynsent (2000). Each **ODI** questionnaire is made up of 10 separate sections, each scored from 0 to 5. The completed sections are tallied up and the final score is calculated from the sum of the scores over the sum of possible scores; the possible score is made up of the possible maximum scores of completed sections. Figure 3.13 shows an example of a completed questionnaire in the dataset. See `http://mapi-trust.org/questionnaires/odi/` for more details. We look at the relationship between MRI and disability using the MRIs in **OSCLMRIC** and the corresponding **ODI** scores in Chapter 8

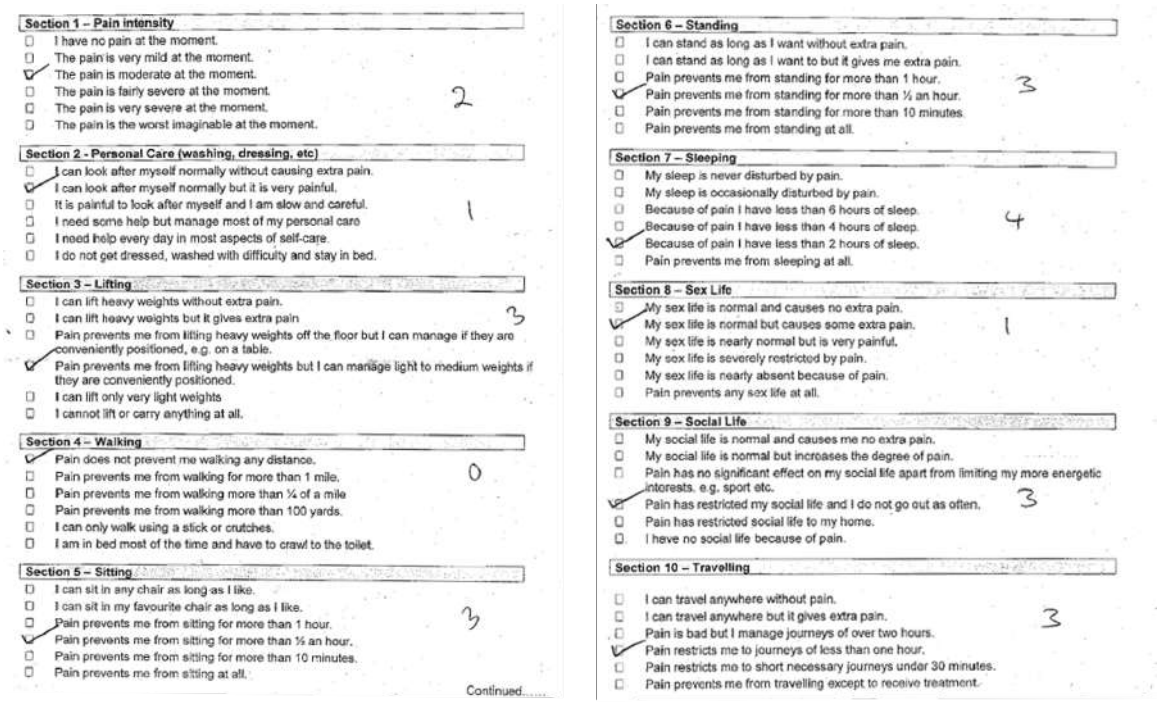


Figure 3.13: ***Oswestry Disability Index Questionnaire****. An example* ***ODI*** *questionnaire by Fairbank and Pynsent (2000) completed by the subjects in the* ***OSCLMRIC*** *dataset. Total* ***ODI*** *score of this example is 46. This is version 2.1a of the ODI questionnaire.*

3.1.4 What Makes a Good Dataset

Throughout this thesis, we explore three main datasets, all containing lumbar MR scans with labels mostly attached to intervertebral discs. The datasets were collected retrospectively and as such, the number of scans are limited to whatever was collected beforehand. If we were to conduct a prospective study, it would be beneficial to know how to construct an adequate dataset, especially a dataset concerning lumbar MRIs and the corresponding radiological measurements.

The most important measure is perhaps the number of subjects in the dataset. As we discuss further along in subsequent Chapters, most prominently in Chapter 7, the number of subjects or more specifically the number of available scans directly affect the performance of the classifier. However, if transfer learning is used, we find that a small number of scans is good enough to train a good classifier, typically around 200 scans for a disc-based measurement. If transfer learning is not feasible or the intended task is far removed from any learned models, the size of the new dataset has to be considerably bigger, perhaps as big as **Genodisc**. Another good idea is to have supplementary measurements, not directly similar to the main measurement, as discussed in Chapter 4. We find that the more additional measurements or tasks added, the better the performance of the classifier. So, instead of just collecting more scans, it is possibly easier to introduce extra supervision to already available data.

3.2 Pre-Processing – Vertebral Body Localization

This section presents several improvements to the previous vertebral body localization system in T2-weighted sagittal scans described by Lootus (2015); the improvements are both for the parts detection and graphical model fit. The experiments for this step were evaluated on a subset of **Genodisc**, with 368 subjects with an 80:80:208 train:validation:test split, similar to that used by Lootus (2015). The best perfor-

mance with the new improvements is a labelling rate of 95.6% which is roughly 10% better than what was previously observed. Localizing these vertebral bodies is an important step in predicting disc gradings which is used in Chapters 4, 6, 7, and 8.

The input for pre-processing is the raw MR volume and the outputs are detections of the vertebral bodies and their level labels. The localization can be broken down into two separate stages: (i) parts detection, and (ii) graphical model fit. A good overview of the vertebral body localization can be seen in Fig. 3.14.

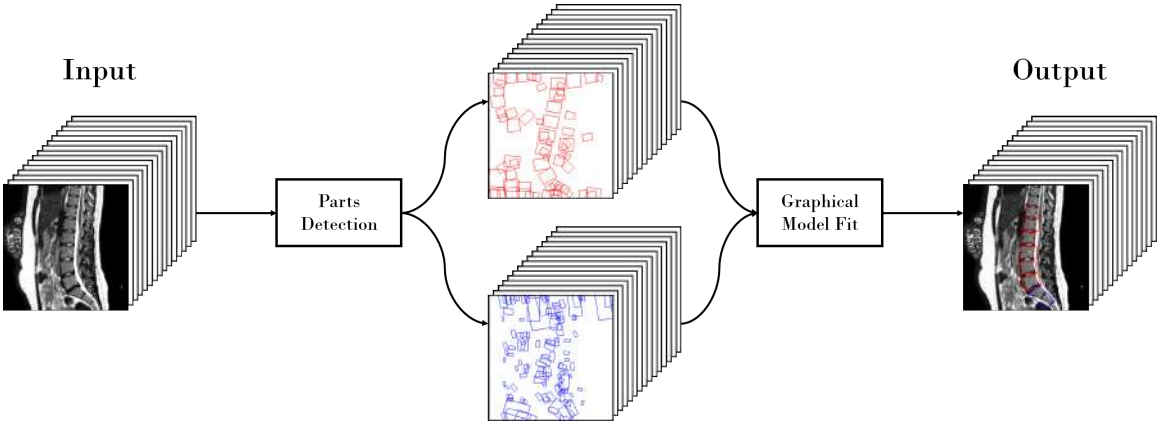


Figure 3.14: ***Vertebral Body Localization Pipeline****. The parts detection stage uses two different templates and a sliding window detector on the same scan. The results of which are two distinct pool of candidates (non-sacral vertebrae in red and sacrum in blue) for every slice in the scan. These candidates are then parse through to fit a trained graphical model, the outputs of which are 7 detected bounding boxes of the vertebral bodies.*

3.2.1 Parts Detection

A single slice of the input MRI undergoes two different parts detection based on the Deformable Part Model (DPM) framework of Felzenszwalb et al. (2010); one for the generic non-sacral vertebrae (T12–L5) and one for the sacrum (S1–S2). This is done for all the slices of the MR scan. The DPM framework is based on the Histogram of Oriented Gradients (HOG) feature descriptor. This type of descriptor is deemed optimal for vertebrae localization since vertebrae are near rectangular and possess generally the same shape especially when viewed sagittally. The HOG feature vectors

used in the DPM framework are 31-dimensional (with a 9-bin histogram): 27 for orientations (18 contrast sensitive + 9 contrast insensitive) and 4 for the gradient energy of the cells around the origin.

3.2.1.1 HOG Templates

Prior to detection, two HOG templates are trained from the minimum area bounding rectangles that came from the original ground truth labels as positive samples. The rectangles of the ground truth labels are taken from the mid-sagittal slice since generally they possess the best looking vertebrae suitable for training. Each scan in the training set provides around one example for sacrum (both S1 and S2 vertebrae) and between 5 to 8 examples for non-sacral vertebral bodies (T10–L5); at test time we only localize T12 to L5. For the detection of the non-sacral vertebral bodies, four distinct HOG templates are trained, all with the same cell height but different cell width, while for the sacrum detection, only one fixed sacrum template is trained. It was found that a single template for sacrum is sufficient while four templates are necessary to capture the variance of the non-sacral vertebrae. Figure 3.15 shows the templates at their default orientations. The negative samples are trained similarly by masking the ground truth rectangles from the sample space.

3.2.1.2 Sliding Window Detection

The sliding window detection is performed per slice for every slice of a given input and outputs the same amount of slices with two different pool of candidates. In Peleg et al. (2007), it was shown that a normal human sacrum ranges between 18° – 81° with respect to the transverse plane (where 0° is horizontally to the left in a sagittal view) which roughly translates to a sacrum HOG template rotation between −9° to −72°. Since the end goal of the spinal analysis system is to detect problematic spines, it is safe to assume the detection step has to take into account a wider range of sacrum

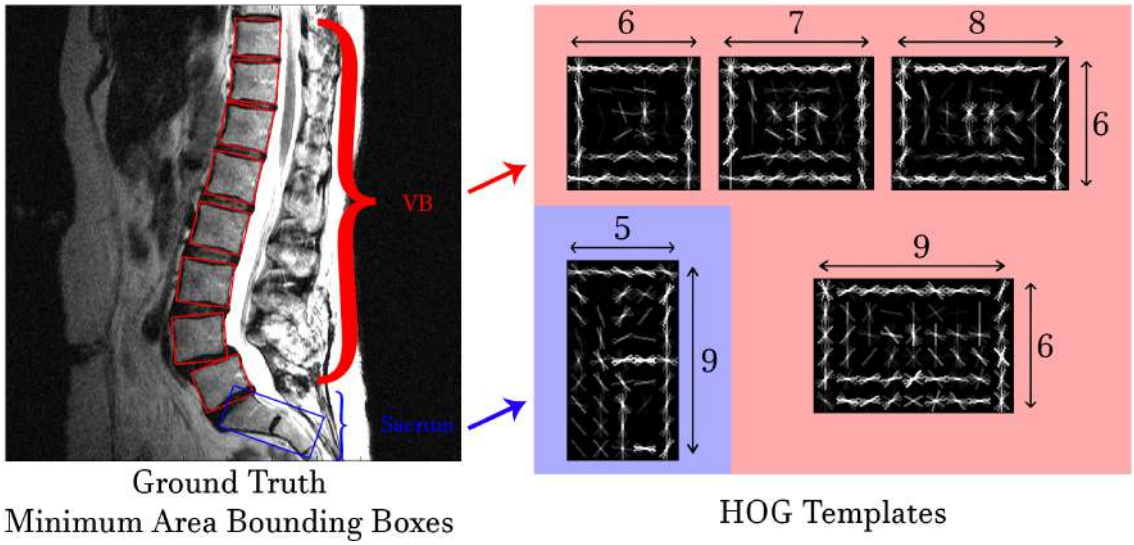


Figure 3.15: ***HOG Templates.*** *The ground truth bounding boxes are used to train two different types of HOG templates: sacrum and non-sacral vertebrae. Because the general shape of the vertebrae in the non-sacral vertebrae category are more varied compared to the sacrum, 4 different templates are used with varying aspect ratios in the sliding window detector compared to the singular template for sacrum. The unit of length of the templates is in cells where one cell is* 8×8 *pixels. VB here means non-sacral vertebrae.*

orientation. To be exact, the trained HOG templates are used in the sliding window detector with varying scales and orientations as shown:

1. Non-sacral vertebrae Orientation: $-45°$ to $45°$ with $5°$ step size
2. Sacrum Orientation: $-90°$ to $10°$ with $5°$ step size
3. Scaling: 0.5 to 2.0 times the original size

The rotation is achieved by rotating the image or slice itself instead of the templates to make it easier to use with the DPM framework but in essence it should behave similarly. Since the scale of the feature pyramid is untrained and MR images are generally noisy, there exists several outlier candidates that are detected in terms of size in both pools and these are deleted, based on the areas of the candidates, prior to model fitting.

3.2.2 Graphical Model Fit

Here the two pools of candidates are parsed and the best fit to the graphical model is found based on the best possible combination of detected scores and the geometrical costs. Explicitly, the best output of a fit contains the localization of the vertebrae which come from the bounding boxes chosen from the pools and also the labels of said vertebrae which roughly translate to the candidates' vertebral level, l_i, from S1-S2 to T12. The first node of graphical model contains the localization of the S1-S2 vertebrae, the second node contains the localization of the 5th lumbar, L5, and so on. The slice position of the candidates in the pools are ignored when finding the best fit but the slice index of the fitted vertebrae can be obtained if need be. The chain graph model layout of the spine can be described as follows:

$$L = (l_1, l_2..., l_n) \tag{3.1}$$

where in the case of just the lumbar spine, l_1 = S1-S2 sacrum and $l_n = l_7$ = T12 or the 12th thoracic vertebra. To be more specific, the vertebra candidate v_i fitted into l_1 will only come from the sacrum candidate pool (V_1) while the rests will come from the non-sacral pool (V_2). The best possible configuration or fit L^* can be expressed as:

$$L^* = \underset{L}{\operatorname{argmin}} \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{v_{i,j} \in V_{1,2}} d_{ij}(l_i, l_j) \right) \tag{3.2}$$

where m_i is the match cost or the score of v_i from the output of the parts detection and d_i is the deformation cost or the pairwise penalties associated with specific v_i and v_j at specific nodes l_i and l_j. The deformation cost or pairwise penalties is the summation of four box functions (B_k) corresponding to four different geometrical properties of adjacent pairs of vertebrae candidates in the chain:

$$d_{i,j} = B_1(x_i - x_j) + B_2(y_i - y_j) + B_3(A_i/A_j) + B_4(\theta_i - \theta_j) \tag{3.3}$$

These four properties are: the change in x-coordinate, the change in y-coordinate, the area ratio and the change in orientation. All the box functions take on the same high constant value if their properties or arguments are outside the trained boundary values and vice versa. The trained boundaries are the minimum and maximum values of the functions when applied to the ground truth bounding rectangles in the training set. A candidate can not be in the same node more than once in a model fit. This is crucial since all the candidates for all the nodes apart from the singular parent node, l_1, come from the non-sacral candidates pool. Without this specification, one candidate might occupy more than one node. This is an inherent weakness of using one generic pool for multiple nodes. This new rule adds a new condition to the minimization problem posed in equation 3.2 slightly:

$$d_{i,j} = \infty \text{ if } v_i = v_j \in V_2 \tag{3.4}$$

To find the minimum cost configuration in equation 3.2, a standard Viterbi algorithm is used which operates in the order of $O(nh^2)$ where n is the number of nodes (which in the case of just the lumbar model, $n = 7$) and h is the number of candidates.

3.3 Pre-Processing – Tighter Bounding Volumes

Loose bounding boxes can hinder good alignment of the disc image for classification. Thus it is desirable to reduce the variation of fit of the bounding boxes, both intra-scan and inter-scan, which can be seen in Figure 3.16.

We propose a finer localization post-processing of these bounding boxes such that the resulting bounding boxes are more consistent and tightly aligned with the vertebral bodies. This can be done by training a regressor to regress to the corner points of the vertebral bodies. We experimented with two regression methods: (i) the supervised descent method (SDM) by Xiong and De la Torre (2013), and (ii) a

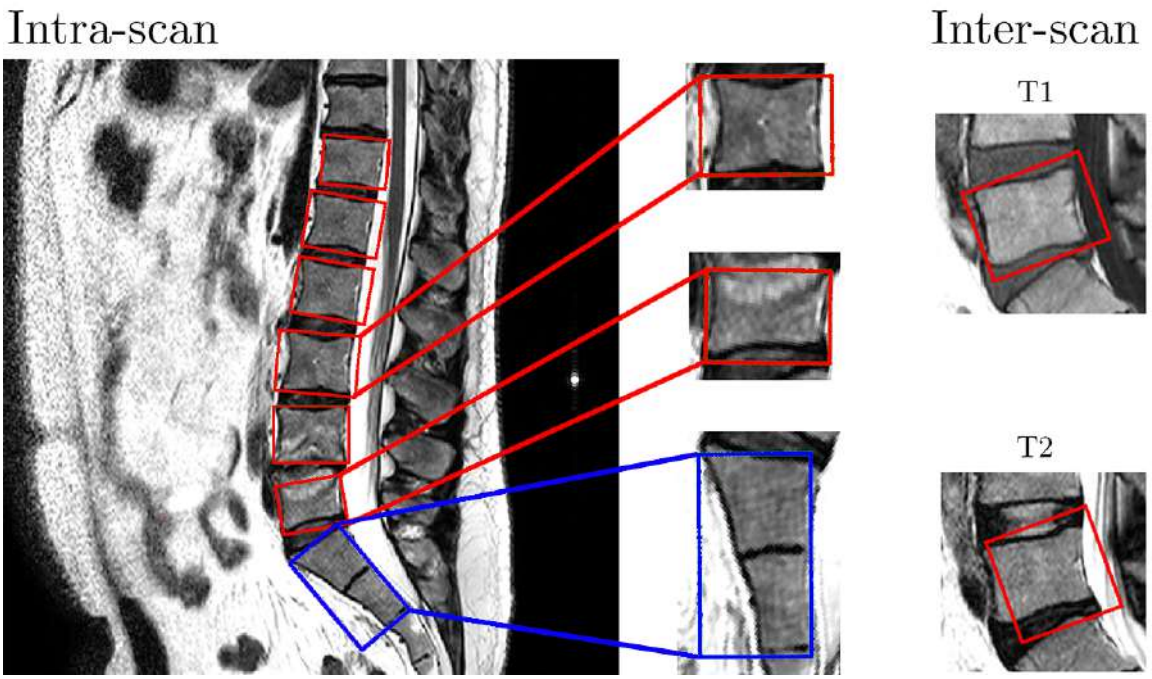


Figure 3.16: ***Variability in Detection.*** *Shown in n example output using the vertebral body localization method. The enlarged set of images of the vertebral body show the variation of fit of the bounding boxes to the vertebral body.* ***Intra-scan variability:*** *Note the sacrum bounding box (**Bottom**) contains both S1 and S2. The L3 bounding box (**Top**) contains all of the vertebra but is slightly loose while the L5 bounding box (**Middle**) is missing the upper endplate of the vertebra.* ***Inter-scan variability:*** *Both images show the same L5 vertebral body but different sequences (T1-weighted and T2-weighted), and their respective bounding boxes.*

CNN regressor. Training, validation and test samples are identical to the ones in Section 3.2. Examples of corner localized vertebral bodies given its raw bounding boxes can be seen in Figure 3.17. The corner localization is done on all slices and all lumbar discs in a given scan.

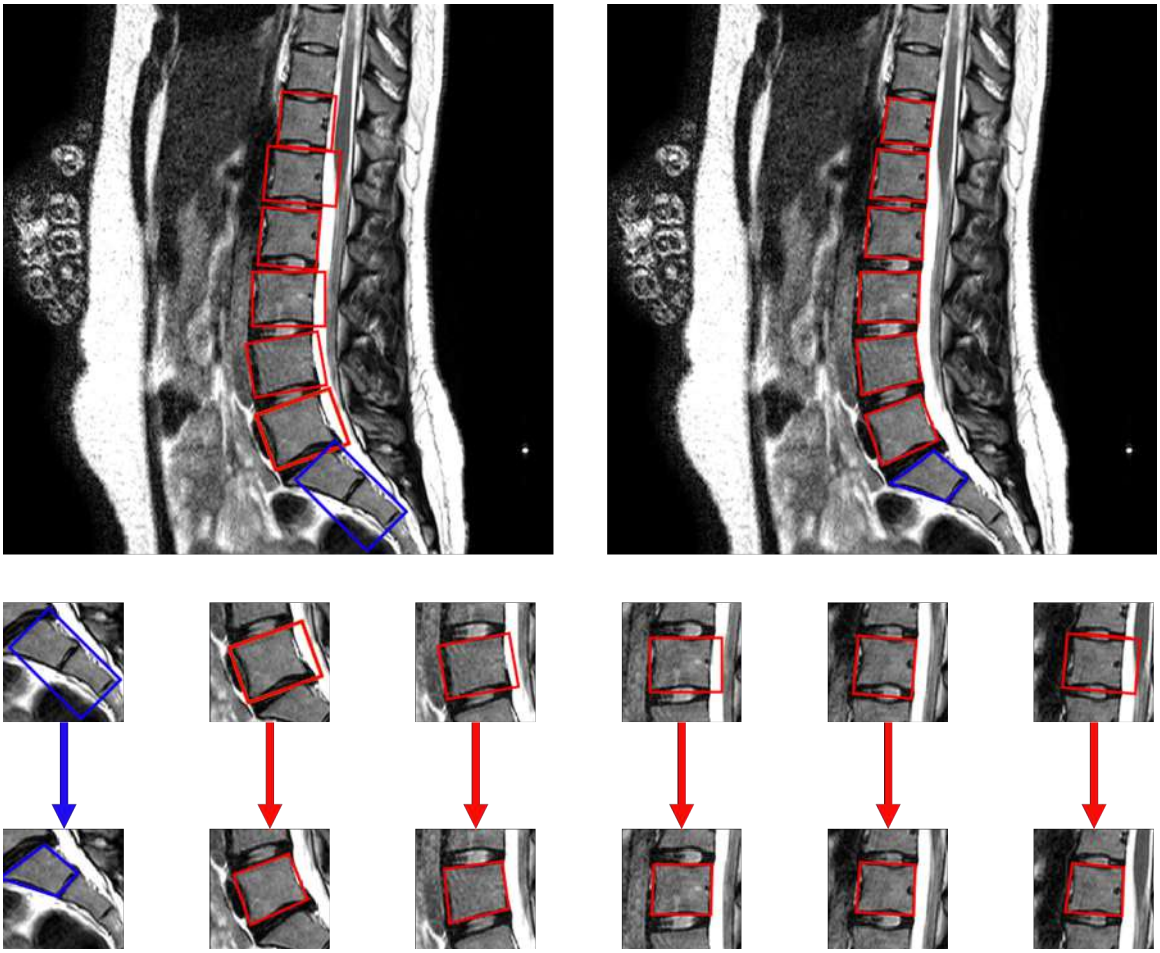


Figure 3.17: ***Outputs of Corner Points Regression. Top Left:*** *Examples of input detected bounding boxes.* ***Top Right:*** *Outputs of the corner localization. Note the sacrum bounding box is now more specific, containing only S1 instead of both S1 and S2, and the variation of bounding boxes fit is reduced.* ***Bottom:*** *Close-ups of the vertebral bodies detection in the first row and the corresponding quadrilaterals from the detected corner points in the second row.*

3.3.1 Supervised Descent Method

The supervised descent method (SDM) which was originally developed for the detection of facial landmarks is adapted to the regress to corner points of vertebral bodies. The input to this stage is the image together with its bounding box from the vertebrae detection and labelling stage and the output is an irregular quadrilateral, with a tighter fit around the vertebral bodies. At test time, the algorithm works by initialising to the corner points of the loose bounding boxes which are iteratively updated to the actual corner points of the vertebral bodies via regression based on SIFT features around the points. The regression is done iteratively and can be represented as:

$$x_{k+1} = x_k + \Delta x_k \tag{3.5}$$

where

$$\lim_{k \to \infty} x_k = x_* \tag{3.6}$$

x_0 are the 4 points of the loose bounding boxes and x_* are the vertebral corner points. After several iterations, k = 10 works well in our dataset, the regression is stopped and the final points become the new corner points describing the best quadrilateral fit for the vertebra.

Training the regressor is posed as a minimisation task of Δx:

$$\underset{\Delta x}{\operatorname{argmin}} f(x_0 + \Delta x) = \| \mathbf{h}(\mathbf{d}(x_0 + \Delta x)) - \phi_* \|_2^2 \tag{3.7}$$

where $\mathbf{h}$ is the feature transformation, which in this case is SIFT, of that point and

$$\phi_* = \mathbf{h}(\mathbf{d}(x_*)) \tag{3.8}$$

represents the SIFT features at the ground truth. To overcome overfitting, ridge

regression was used. The regularizer, λ, is a general singular value penalty imposed per iteration. Since the size of the vertebral bodies in the dataset vary, a normalisation step has to be conducted prior to regression. An individual vertebral body is resized according to the height, V_h, of its bounding box and the image is translated such that the centre of the bounding box is the point of origin.

Two sets of models are trained for corner localization, each set having a single regressor for the standard non-sacral vertebrae or normal vertebral bodies (VB), T12 to L5, and a more specific S1 regressor. The first set, termed the 1^{st} stage, is for more coarse localization, and the second set, termed the 2^{nd} stage which produces the final outputs is for a more precise corner regression. Both sets of models are trained with the same ground truth used in training the detection and labelling system.

Overall, 4274 T12-L5 and 720 S1 vertebrae were used in training the VB and S1 models respectively. There are only 4 parameters that have to be optimised; they are: the normalised height of the VB, V_h, the size of the SIFT patch, the regularizer of the ridge regression, λ, and the number of iterations, k. They are optimized such that they minimize the error on the validation set.

3.3.2 CNN Regression

Similar to the SDM, the end-goal is the regression of the four corner points of each vertebral body, producing a tighter quadrilateral, given an initial loose bounding box detection. For each vertebral body, the detected bounding box is expanded to be 50% larger than the initial detection, resized to be 224×224 in dimension, and passed through a CNN for regression of the four corner points (8 regression outputs i.e. $N = 8$). We employ a simple CNN architecture with three convolutional, **Conv**, layers and two fully-connected, **FC**, layers as seen in Figure 3.18.

The CNN was trained using the same samples with the SDM with an L_2 loss:

$$\mathcal{L} = \sum_{n=1}^{N} (\hat{y}_n - y_n)^2 \tag{3.9}$$

where $\hat{y}_n$ is the ground truth and y_n is the prediction. Unlike the SDM, there is only a single model and a single stage for the CNN regressor i.e. there is only one CNN model predicting the corners of the non-sacral vertebrae (VB) and the S1 sacrum.

Design choices for the CNN in this Section and subsequent Chapters are inspired by standard practices set forth by specific works, particularly the size of on the image input used by AlexNet Krizhevsky et al. (2012), the size of the kernels in VGG-M Chatfield et al. (2014), and several sane values for hyperparameters used by VGG-16/19 Simonyan and Zisserman (2015). These design choices are tweaked for every network used in this thesis and validated empirically for each experiment.

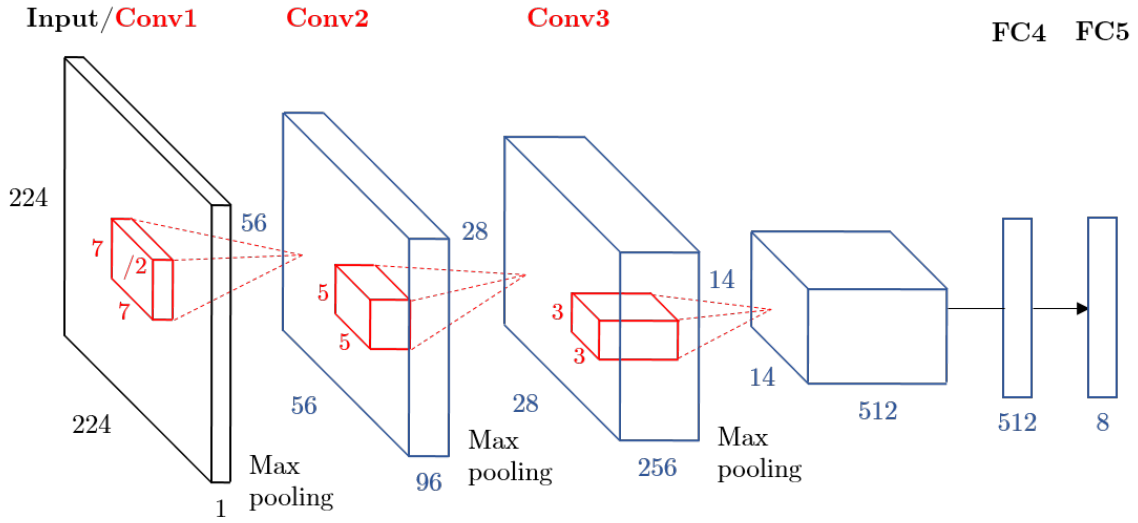


Figure 3.18: ***CNN Regressor.*** *Shown is the CNN architecture of the regression CNN made up of three convolutional layers and two fully-connected layers.*

3.3.3 Corner Regression Results

Figure 3.19 highlights the distances, in mm, to the ground truth corner points of each vertebral body from: (i) corners of the initial and loose bounding boxes, (ii) corners from the SDM and (iii) corners from CNN regressor. In general, for the CNN regressor which works best, 98.4% of the S1 vertebrae and 96.5% of the VB have errors less than 2.8mm, roughly 1 pixel off per corner, which is a considerable improvement over the original bounding box detections with none of the S1 vertebrae and 13.9% of the VB at the same error threshold. Comparable performance can be seen for the SDM output with 93.8% for S1 and 94.5% for VB.

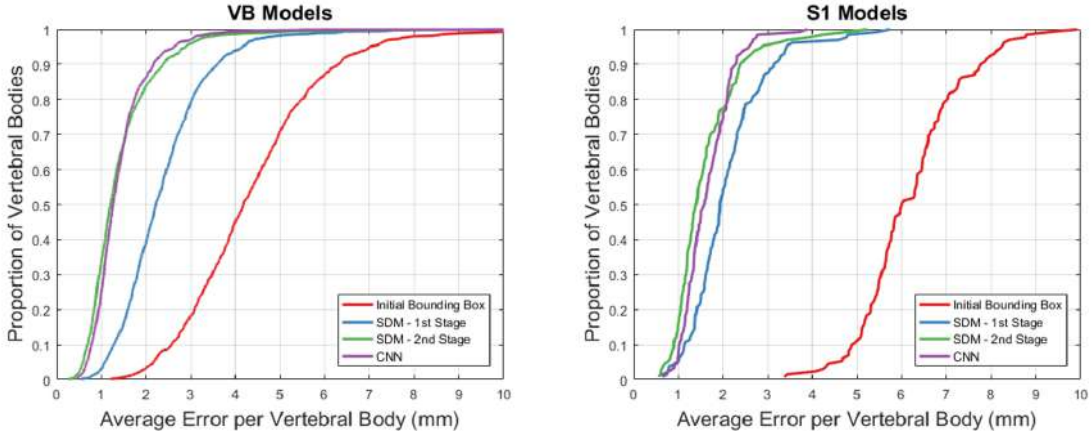


Figure 3.19: ***Comparison of Various Regression Methods. Left:*** *Error per vertebral body of the VB regression.* ***Right:*** *Error per vertebral body of the S1 regression. The red, green, blue, and purple plots represent the errors of the raw initial bounding boxes, 1^{st} stage of the SDM, 2^{nd} stage of the SDM, and the CNN regressor respectively. Both stages of the SDM outperform the initial bounding box considerably. This is especially true for S1 which has a larger fit variance of the bounding boxes compared to VB. The CNN method slightly outperforms the SDM while only requiring one model for VB and S1 regressions.*

3.4 Pre-Processing – Extent of Vertebral Bodies

This pre-processing step is to determine the 3D extent of the vertebral bodies from the 2D quadrilaterals in each sagittal slice; this requires determining where the original detections should start and end slice-wise in a sagittal scan. This is important

since the positions of the vertebrae in a scan are initially unknown and there exist slices which contain only partial volumes of the vertebrae, mostly containing tissue. Knowing the extent would also aide us align the volumes and mitigate slice errors from pathological cases e.g. scoliosis where slices of the bodies can be misaligned in the sagittal plane. To this end we utilise a classifier to distinguish non-vertebra and vertebra quadrilaterals. We follow a standard image classification scheme, discussed by Chatfield et al. (2011), where the sequential steps are: (1) dense SIFT feature extraction over the quadrilaterals, (2) Fisher Vector (FV) encoding of the features, (3) spatial tiling of the features in the image and (4) classification via linear SVM. This is done on a per slice basis on every slice.

For classification of vertebra and non-vertebra, the same patients that trained the corner regressors were used; a total of 66,556 VB quadrilaterals (39,309 vertebrae and 27,247 non-vertebrae) with 50:50 train and test split. In general, this step performs well with an accuracy of 95.6% at distinguishing vertebra and non-vertebra quadrilaterals. Since we now know which slices contain vertebral bodies, we can trivially transform a series of quadrilaterals in different slices to a bounding volume for each vertebral body.

3.5 Summary

In this chapter, we have discussed all the datasets that will be used throughout this thesis namely the **Genodisc**, **TwinsUK**, and **OSCLMRIC** datasets. We have also shown the necessary pre-processing used for the detecting tight bounding volumes of the vertebral bodies widely used in the subsequent chapters. In the next chapter, Chapter 4, we look at predicting radiological gradings attached to the intervertebral discs in the **Genodisc** dataset.

Chapter 4

Predicting Radiological Gradings

In this chapter, we discuss a CNN-based framework to classify radiological gradings in T2-weighted sagittal lumbar MRIs. The method is trained, validated and tested on the **Genodisc** dataset (see Chapter 3). Given an input scan, the trained model predicts, for each of the six lumbar intervertebral discs, eight separate classifications corresponding to eight different radiological gradings: (i) **Pfirrmann grading**, (ii) **disc narrowing**, (iii) **upper endplate defects**, (iv) **lower endplate defects**, (v) **upper marrow changes**, (vi) **lower marrow changes**, (vii) **spondylolisthesis**, and (viii) **central canal stenosis**. Instead of training a model for each classification task, one model per radiological grading, we instead learn a single model that predicts all the gradings simultaneously. To this end, we have devised our own methodology of multi-task learning with unbalanced labels which can be easily extended to other tasks or modalities, not just disc gradings of spinal MRIs.

This chapter can be broken down into three main sections; the first one being the definition of the classification task and details of the input disc volume in Section 4.1, followed by details of the model in Section 4.2, and finally the experiments and results in Section 4.3. A simplified view of the overall pipeline is given in Figure 4.1.

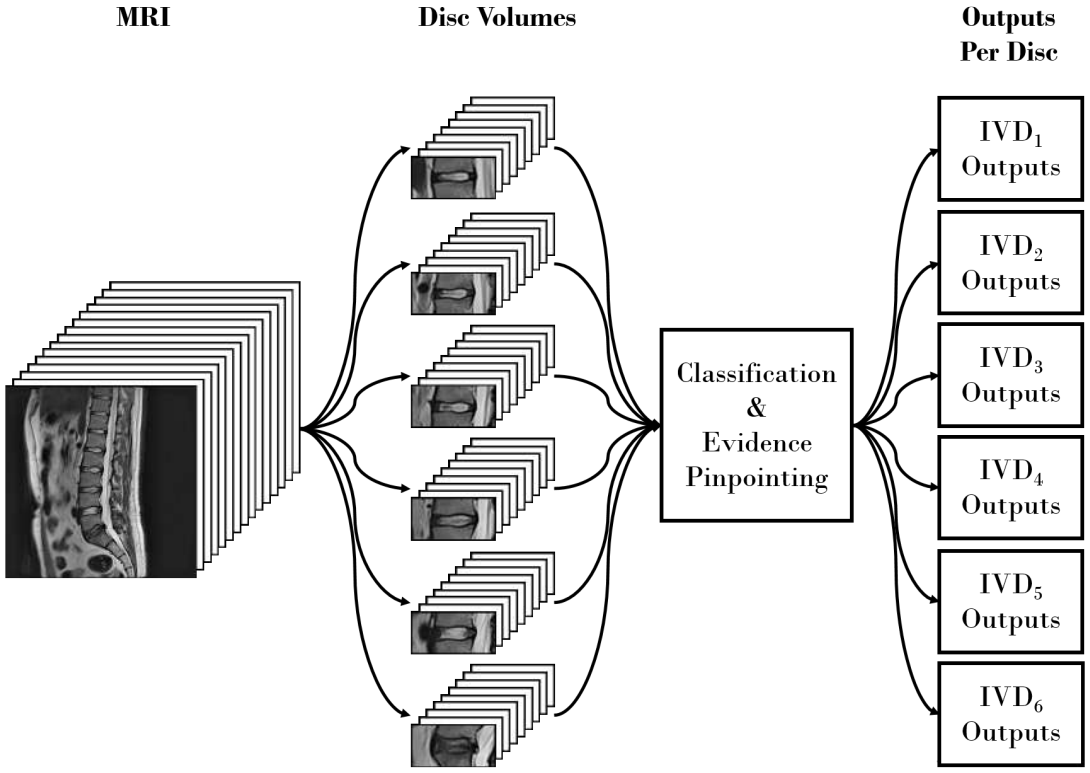


Figure 4.1: ***Automated Radiological Grading Pipeline.*** *Each of the six intervertebral disc volumes is assessed as an input to the CNN. There are eight outputs for each disk consisting of a radiological grading prediction and an associated hotspot mapping. The CNN architecture is given in figure 4.3. The dimension of each input disc volume is* $112 \times 224 \times 9$*, essentially a stack of 9 image slices of the disc, which is also the dimension of the evidence hotspot heatmap in Chapter 5. Description of the intervertebral disc outputs,* $\mathbf{IVD_1}$ *to* $\mathbf{IVD_6}$*, can be seen in Figure 5.1.*

4.1 Classification Overview & Input Volumes

A standard lumbar spinal MR scan typically includes multiple intervertebral discs in its field-of-view. The number of discs varies depending on the pathology of the patient, the clinical question of interest and the protocol standards of the clinical centre, but typically contain the six lumbar discs from the T12-L1 disc to the L5-S1 disc. Our goal is to predict the radiological gradings for the discs and associated vertebrae, and we choose as input to the CNN a volume around each disc that includes the disc and part of the vertebral bodies above and below the disc. The CNN predicts all of the gradings for the disc volumes simultaneously, as well as the heatmap of evidence hotspots (Chapter 5). We have experimented with whole spine images as input and found that there is a significant decrease in terms of classification performance. The network is trained simultaneously to predict all the radiological gradings via a multi-task loss function.

Why Multi-task? Despite their advantages and leading performance, one of the challenges of utilizing CNNs in medical imaging remains: their need for large training datasets. We address this in our work partly through aggressive dataset augmentation but also through the use of multi-tasking where one architecture serves to address multiple problems and hence provides multiple supervisory signals for the common trunk i.e. convolutional layers. We show that multi-tasking can improve classification performance, demonstrating that the features and spatial relationships learnt by the convolutional layers generalize to other related tasks. Since each classification problem is its own unique task, solving them at the same time is akin to multi-tasking. Since clinicians are typically required to assess the state of different anatomical regions within the medical image, either because they are relevant to the clinical question or because they are required to assess visible anatomy for so-called incidental findings, the development of techniques that can predict multiple gradings

simultaneously is desirable. In our application of interest, each disc and vertebra can have grades to describe their state of normality or degradation.

4.1.1 Disc Volume Extraction

From each pair of vertebral bodies detected in Chapter 3, rough estimates of intervertebral disc bounding volumes are obtained using the upper and lower corner points of each vertebra in each slice of a detected volume e.g. a L5-S1 disc volume is made up of the lower corner points of L5 and upper corner points of S1. Then, the disc volumes for classification are defined as follows: the region is rotated within the sagittal slice so that the disc is horizontal and centered; an exploded view of the disc volume can be seen in Figure 4.2. The regions are resized, while maintaining aspect ratio, to be the same dimension 112×224 per slice; this 1:2 ratio is to ensure that the disc region would not include the upper and lower endplates of the adjacent vertebral bodies. Roughly 40% of each vertebrae, upper and lower, appear in each region of interest. The discs are aligned according to their mid-sagittal slices and include up to 9 sagittal slices. Narrow discs with less than 9 slices are zero-padded slice-wise. Each slice of the disc is normalized such that the median intensity of its pair of vertebral bodies is 0.5 to mitigate against bias field effects. The range of the intensity inside the disc volume is set to be between 0 and 1.

4.2 Loss Functions & CNN Architectures

In this section we first describe the multi-task loss functions that are used to train the network, followed by the CNN architectures that we compare. Training and implementation details are then given in Section 4.2.3.

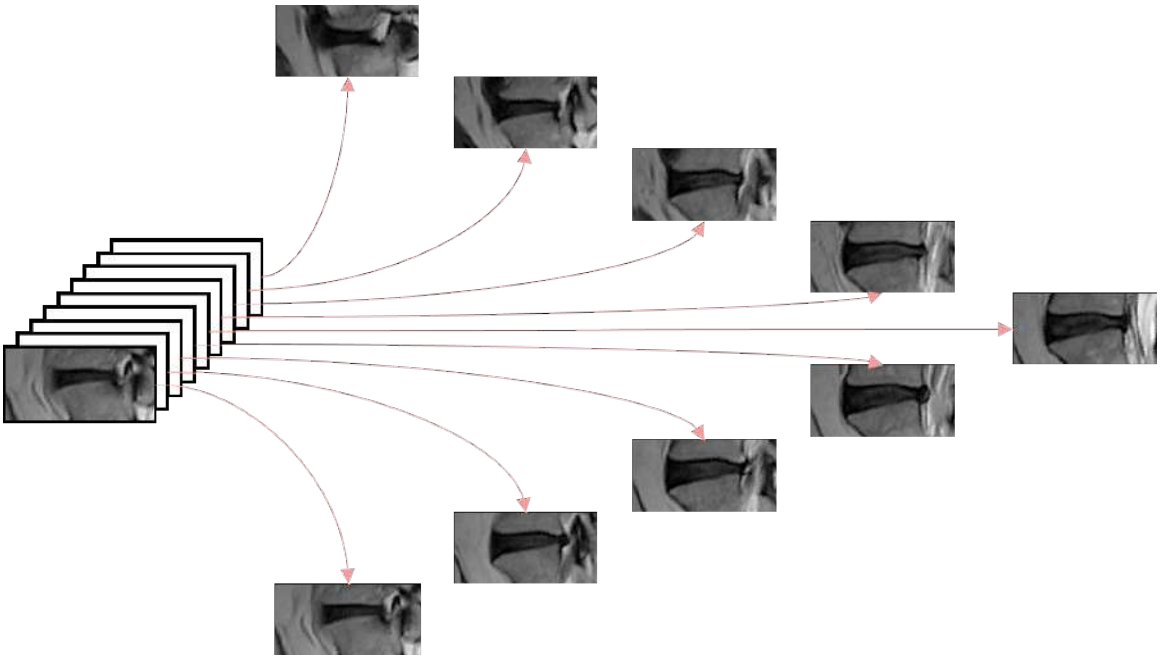


Figure 4.2: ***Disc Volume Example.*** *Shown is an exploded view of an extracted disc volume where each image is an individual slice. The middle slice for each volume is independently detected on a per disc basis and is not the mid-sagittal slice of the MRI volume. The per disc alignment is to mitigate against pathological cases like scoliosis which results in different discs in the same scan to have different mid-sagittal slices.*

4.2.1 Loss Functions

The loss to train the network has to both handle multiple multi-class classifications and also balance the contribution of each class in every task.

4.2.1.1 Multi-task Loss

Since each of the eight radiological gradings (see Chapter 3) are classification tasks, we follow the conventional practice of training a network by minimising the softmax log-losses (cross-entropy) for the grading. Each grading is considered as a separate task, and a separate head is defined on the network for each task (with a common trunk shared between the tasks, see below). For each task, t, where $t \in \{1 \dots T\}$, and input volume, x, the network outputs a vector y of size C_t which corresponds to the number of classes in task t. The loss, $\mathcal{L}_t$, of each task over N training volumes can

be defined as:

$$\mathcal{L}_t = -\sum_{n=1}^{N}\left(y_c(x_n) - \log\sum_{j=1}^{C_t} e^{y_j(x_n)}\right) \tag{4.1}$$

where y_j is the jth component of the **FC8** output, and c is the true class of x_n. Solving the multi-task problem in an end-to-end fashion translates to minimizing the summation of all the losses i.e. all t in Equation 4.1:

$$\mathcal{L} = \sum_t \omega_t \mathcal{L}_t \tag{4.2}$$

where ω_t is the weight of task t. We find setting $\omega_t = 1$ for every t works well on our tasks, but it might be beneficial to fine tune ω_t for different problems. Setting one weight to 1 and the rest to 0 results in a standard training of a single task. At training time, the loss of task t is only calculated for valid labels i.e. missing labels of task t are ignored. This is extremely beneficial as inputs can possess missing labels in one task but not others.

4.2.1.2 Class-balanced Loss

Class imbalance refers to the different number of training examples for each possible outcome of a multi-way classification. Since most of the classification tasks we deal with are unbalanced, a common problem in medical classification tasks, we use a class-balanced loss during training. For each task, we reweight the loss such that the combined losses are balanced. To achieve balance in training, class-specific weights are introduced. These weights are determined to be:

$$\alpha_c = freq(c)^{-1} \tag{4.3}$$

where $freq(c)$ is the class label frequency in the training sets for each task e.g. $freq(1)$ is the frequency of the first class, $c = 1$, in the training set. The loss for each task

can then be expressed as:

$$\mathcal{L}_t = \sum_c \alpha_c \ell_t(c) \tag{4.4}$$

where $\ell_t(c)$ is the component of the loss for class c. This is equivalent to oversampling the minority class but, since our data is multi-labelled, this balance cannot be achieved by a trivial solution such as simply oversampling or undersampling each disc.

4.2.2 CNN Architectures

The base network architecture is a modified version of the 1024 variant of the VGG-M network introduced by Chatfield et al. (2014) with 2D kernels. The input dimension of the disc volume is $112 \times 224 \times 9$ where 9 refers to the number of slices in an extracted disc volume (see Section 4.1 for disc volume details). The **Conv1** kernels of VGG-M are changed to accept 9 input channels instead of the standard 3 for RGB images. We also omit the use of local response normalization after the **Conv1** and **Conv2** layers and the stride of the **Conv2** layer is changed to 1 to ensure no information is lost after the **Conv5** pooling layer. The main configuration we use for parameter sharing is a branch point after **Conv5**. To predict the multiple tasks, each task branches out after the specified branch point i.e. unique **FC6**, **FC7**, and **FC8** layers for each task. The layers succeeding the branch point are identical in terms of the number of weights for each task except for the final **FC8** layer, which has specific output dimensionality according to the number of classes in each task e.g. 5-way softmax for Pfirrmann, and 4-way softmax for disc narrowing classifications. Figure 4.3 shows the network with a branch point after **Conv5**. We also experimented with different branch points.

4.2.2.1 3D kernels

Since MR scans comprise a stack of 2D images forming a 3D volume, we can either use a 2D CNN which at the **Conv1** layer treats individual slice of the whole stack

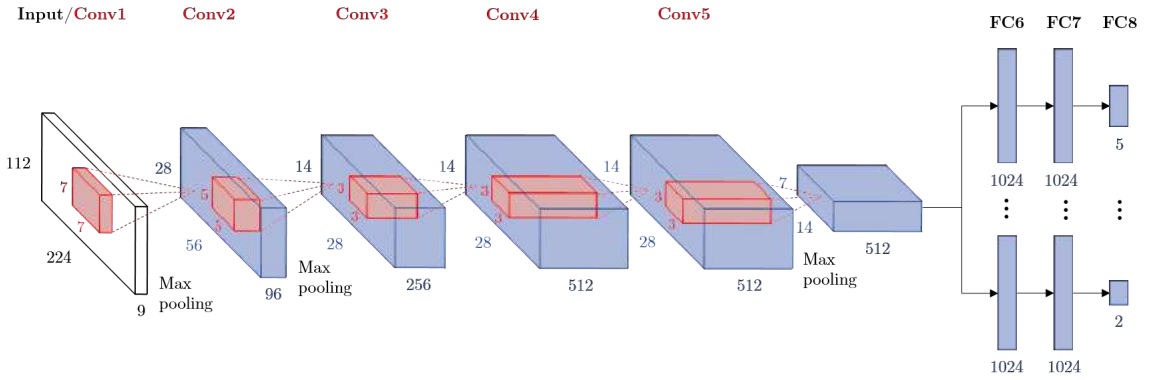


Figure 4.3: ***CNN Architecture – Conv5 (Radiological Gradings Predictions).*** *Multi-task version of the VGG-M architecture by Chatfield et al. (2014) with a branch point after the* ***Conv5*** *layer resulting in a multi-way classification tasks with each task corresponding to a unique radiological grading. The numbers in black refer to the input, the numbers in blue are for the activations after an operation (e.g. convolution) while the numbers in red (for example 7 in* ***Conv1****) refer to the size of the kernels. The max-pooling window is set to be* 2×2 *with a stride of 2. For all the convolutional kernels, the stride is set to be 1 except for* ***Conv1*** *which is set to be 2. For the 3D network, all the convolutional kernels are extended to have an extra dimension of size 3 e.g. for* ***Conv1*** *the kernel would be* $7 \times 7 \times 3$*. In the figure, the output of the upper* ***FC8*** *layer is set to size 5 to predict Pfirrmann grading, while the lower one illustrates a binary output with a size 2 output.*

of 2D images as separate feature channels or a 3D CNN which uses 3D convolutional kernels. Both ideas have merits: using a 2D CNN makes sense since spinal MRIs, and in fact many other MRIs in clinical practice, are often assessed by radiologists on a per slice basis and as such is closer to the condition at which the radiological gradings were annotated; while using a 3D CNN means keeping the integrity of the volumetric information we are assessing. In the 3D case every convolutional kernel except for **Conv4** is extended to have an extra dimension of size 3, going from right-to-left slice-wise in the volume, with no pooling and the stride set to be 1 e.g. in **Conv1** the standard 7×7 kernel is extended to $7 \times 7 \times 3$. The smallest odd kernel size of 3 was used for the last dimension since the input disc is limited to just 9 slices at maximum. To keep the number of parameters comparable, the **Conv4** layer is set to be $3 \times 3 \times 9$ with no padding slice-wise which results in a reduction of the **Conv4** activation dimension. Subsequent layers proceeding **Conv4** are thus exactly the same as those in a 2D CNN. Overall, the number of parameters for the 3D CNN

is roughly 2% more than the 2D CNN (434M vs 426M, eight classification tasks).

4.2.3 Implementation Details

4.2.3.1 Training

Training of all the tasks is done end-to-end simultaneously via stochastic gradient descent with momentum from scratch without any pre-training. The inputs are normalized with per-channel mean subtraction as per common convention for the 2D models but per-volume mean subtraction for the 3D models. The hyperparameters are: mini-batch size 256 (2D) & 128 (3D); momentum 0.9; weight decay 0.0005; initial learning rate 0.001, which is lowered by a factor of 10 as the error plateaus. The weights are initialized according to the method by Glorot and Bengio (2010) and normally reach convergence in about 1000 epochs. The networks were trained via the MatConvNet toolbox (by Vedaldi and Lenc (2014a)) using an NVIDIA Titan X GPU.

4.2.3.2 Data augmentation

We employ several aggressive random on-the-fly augmentation strategies during training. We include the methods suggested by Krizhevsky et al. (2012) for natural images, and also define additional augmentations that are suited to 3D MR scans. The training augmentation strategies are:

1. Rotation with $\theta = -15°$ to $15°$
2. Translation of ± 32 pixels in the x-axis, ± 24 pixels in the y-axis, ± 2 slices in the z-axis
3. Rescaling with a scaling factor between 90% to 110%
4. Intensity variation between -0.1 to 0.1

5. Random slice-wise flip i.e. reflection of the slices across the mid-sagittal

At test time, the final prediction is calculated from the average of 54 predictions:

1. 8 patches, ±16 pixels from the origin, alongside the centre patch,
2. their slice-wise reflections
3. sliding the slice window ±1 across the volume

4.3 Experiments & Results

The dataset is split into a 80:10:10 train:validation:test sets on a per patient basis (not per disc). This results in 1806 patients (10,836 discs) train/val samples, and 203 patients (1,224 discs) test samples. All the results reported here are from two models trained for each experiment by swapping the validation and test sets.

4.3.1 Evaluation Protocols

To evaluate classification performance, we use average per-class accuracy which is suitable for highly unbalanced classifications. For comparison, we provide the average per-class intra-rater agreement calculated from two separate sets of labels by the same radiologist. These labels were obtained for a subset of the dataset (121 patients, 726 discs) assessed by the radiologist at different times (in comparison the full test set consists of 203 patients). The intra-rater agreement serves as a good benchmark of performance since we are essentially limited to the quality of the label i.e. we can only be as good as the radiologist. Note that inter-rater agreement, i.e. agreement between two different radiologists, tends to be consistently worse than the intra-rater agreement e.g. Pfirrmann et al. (2001) reported that **Pfirrmann grading** has an inter-rater kappa range of $\mathbf{0.69 - 0.81}$ while the intra-rater range is $\mathbf{0.84 - 0.90}$; the radiologist in our dataset has an above average intra-rater kappa value of $\mathbf{0.91}$.

To obtain a standard deviation over the results, two models are trained for each experiment by swapping the validation and test sets.

4.3.2 Choosing Branch Point

We look into several variations of the architecture by changing the branch point. Layers immediately after a branch point are duplicated for each of the six tasks e.g. for a network with a branch point after **Conv5**, as seen in Figure 4.3, we have six unique **FC6**, **FC7**, and **FC8** layers, one for each task. The accuracy for each task and the intra-rater agreement is given in Table 4.1. It can be seen that branching immediately after **Conv5** is the best choice, and we use this configuration in subsequent comparisons. However, if a more compact representation of the disc volume is desired, branching at **FC6** also works well (see Figure 4.4). We also experimented with turning off data augmentation during both training and at test time, and found that there is a consistent decrease of 0.5% in performance if test time augmentation is turned off, and the network overfits to the training set when augmentation is turned off during training.

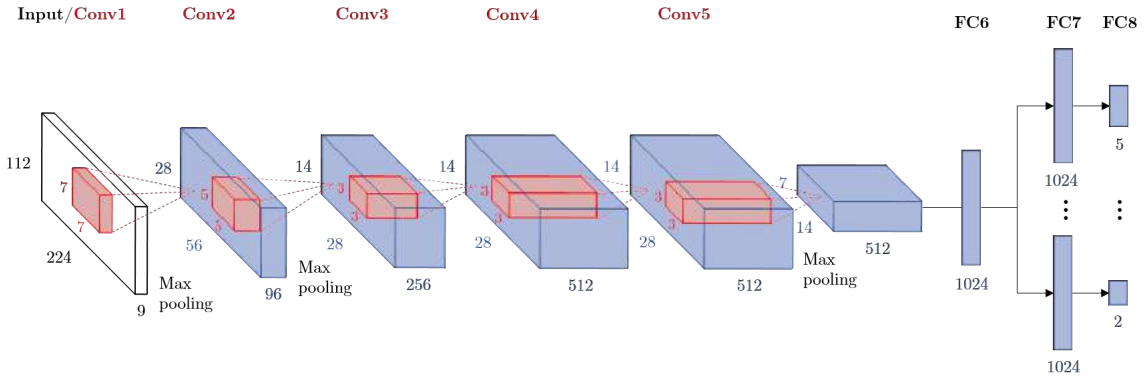


Figure 4.4: ***CNN Architecture – FC6 (Radiological Gradings Predictions).*** *Similar to the network shown in Figure 4.3 but with a different branch point, at* ***FC6*** *instead of* ***Conv5****. One major boon of this branch is the savings in terms of parameters (64M vs 321M for six classification tasks) with relatively low reduction in performance. The branch point can start as early as* ***Conv1*** *and as late as* ***FC7*** *for the VGG-M architecture.*

Task	Intra-rater	Branch Point					
		None	Conv3	Conv4	Conv5	FC6	FC7
Pf	70.4	69.8 ± 0.4	71.2 ± 1.0	70.9 ± 0.1	$\mathbf{71.2 \pm 0.4}$	69.8 ± 0.4	70.9 ± 0.6
DN	72.0	72.3 ± 2.1	74.4 ± 0.4	73.3 ± 1.9	73.9 ± 0.7	73.2 ± 0.1	$\mathbf{74.5 \pm 0.7}$
UED	80.7	79.0 ± 0.5	83.0 ± 1.5	81.7 ± 2.5	84.8 ± 0.6	$\mathbf{85.8 \pm 0.3}$	85.7 ± 0.1
LED	83.3	79.5 ± 1.2	82.8 ± 1.5	84.5 ± 1.2	$\mathbf{87.3 \pm 2.3}$	85.8 ± 2.0	86.4 ± 2.0
UMC	92.5	88.1 ± 0.6	89.1 ± 0.1	89.2 ± 0.2	$\mathbf{90.1 \pm 0.4}$	89.2 ± 0.3	89.2 ± 0.5
LMC	91.4	87.3 ± 0.3	88.2 ± 0.7	88.9 ± 0.2	$\mathbf{89.0 \pm 0.4}$	88.5 ± 0.8	88.2 ± 0.7

Table 4.1: ***Choosing Branch Point Experiment.*** *The performance (mean ± std %) with various different branch points on the test set.* ***Pf*** *=* ***Pfirrmann grading****,* ***DN*** *=* ***disc narrowing****,* ***UED*** *=* ***upper endplate defects****,* ***LED*** *=* ***lower endplate defects****,* ***UMC*** *=* ***upper marrow changes****, and* ***LMC*** *=* ***lower marrow changes****. Branch point labelled "****None****" refers to six individual networks each trained for the six individual tasks. It can be seen that multi-task training results in a better performance across the multiple tasks and that there is a sweet spot for choosing a branch point at* ***Conv5*** *in our case. "****Intra-rater****" is the intra-rater agreement. Since we are using a predefined architecture which was proposed for a harder classification task, we theorized that learning multiple tasks constraint the training of the network, effectively acting as regularizers.*

4.3.3 Multi-tasking

We investigate the relationship between the addition of tasks and its impact on the performance of the model. For this experiment we used the 2D network that share layers up till **Conv5**, and each individual task afterwards has its own **FC6**, **FC7**, and **FC8** layers. The accuracy for each task and the intra-rater agreement is given in Table 4.2. It can be seen that going from a CNN trained on a single task, in this case **lower endplate defects**, to one trained on multiple tasks improves the performance considerably (**79.5**% → **86.4**%). This is also true one other tasks, although not shown in Table 4.2, i.e. **Pfirrmann grading** (**69.8**% → **71.2**%), **disc narrowing** (**72.3**% → **73.9**%), **upper endplate defects** (**79.0**% → **85.7**%), **upper marrow changes** (**88.1**% → **89.2**%), and **lower marrow changes** (**87.3**% → **88.2**%). Overall, all the learned tasks have better results if the tasks are jointly learned and the inclusion of more tasks generally results in a better performance. One possible explanation of how adding more tasks might be beneficial is the correlation between certain tasks. A clear example is endplate defect; an appearance of an upper endplate

defect is normally also accompanied by a lower endplate defect. The appearance of an endplate defect is theorized to come from a disc degeneration which results in defects to both the lower and upper vertebral endplates. This prior of the correlation is also true for some other radiological scores. Another possible explanation of the benefit of adding more tasks, at least in our application, is the fact that different tasks operate on different regions in the input i.e. **Pfirrmann grading** and **disc narrowing** for the disc, **endplate defects** for the vertebral endplates, and **marrow changes** for the vertebral bodies.

4.3.4 2D vs 3D Architectures

The classification performance of the 2D and 3D models is compared in Table 4.3. It appears that most gradings, except for **Pfirrmann grading** (**71.9**% → **71.5**%) and **disc narrowing** (**75.9**% → **75.0**%), are better or equal in performance with the 3D CNN. These exceptions might be due to the fact that both **Pfirrmann grading** and **disc narrowing** are normally graded by radiologists purely on only the mid-sagittal slice of each disc. The most significant difference in terms of performance going to 3D can be seen in the prediction of **spondylolisthesis** (**92.9**% → **95.2**%), slippage of the vertebral bodies, which is the most global grading (in terms of the disc volume fed into the CNN) and benefits most with the addition of the 3D kernels. Overall, the 3D network is on average **0.6**% (**85.7**% → **86.3**%) better in terms of per task accuracy.

We investigated alternative 3D networks i.e. those without the reduction of slice-wise dimension at **Conv4** but these proved to be only marginally better than our baseline 2D network. We suspect with the addition of more training data the 3D model may improve further, as in the case for Tran et al. (2015) where the presence of a large-scale dataset made 3D CNNs shine. We also experimented with deeper networks (VGG-16, ResNet, etc.) using 2D kernels, but obtained little to no improvements in

Task	Intra -rater	Lower Endplate Defects	+ Upper Endplate Defects	+ Lower Marrow Changes	+ Upper Marrow Changes	+ Pfirrmann	+ Disc Narrowing
Lower Endplate Defects	83.3	79.5 ± 1.2	84.3 ± 0.7	85.4 ± 0.9	85.2 ± 0.1	86.0 ± 2.3	**86.4 ± 2.0**
Upper Endplate Defects	80.7	-	84.8 ± 1.9	84.7 ± 3.0	85.0 ± 1.9	85.6 ± 0.5	**85.7 ± 0.1**
Lower Marrow Changes	91.4	-	-	**88.6 ± 0.2**	88.4 ± 2.4	87.9 ± 1.2	88.2 ± 2.0
Upper Marrow Changes	92.5	-	-	-	89.5 ± 0.1	**89.7 ± 0.2**	89.2 ± 0.5
Pfirrmann	70.4	-	-	-	-	68.8 ± 0.7	**71.2 ± 0.4**
Disc Narrowing	72.0	-	-	-	-	-	**73.9 ± 0.7**

Table 4.2: ***Multi-tasking Experiment.*** *Performance under variations of the number of tasks for the 2D network. The network was first trained on the* ***lower endplate defects*** *and from left to right the network is reinitialized and trained again from scratch with the addition of more classification tasks. The rightmost column shows the results with a network trained with six of the eight tasks. The tasks are added after the* ***Conv5*** *branch point. Overall, there is a positive correlation between the amount of shared tasks and the performance of the network.*

Tasks	Intra-rater	Models	
		2D	3D
Pfirrmann	70.4	**71.9 ± 1.5**	71.5 ± 1.0
Disc Narrowing	72.0	**75.9 ± 0.4**	75.0 ± 2.3
Upper Endplate Defects	80.7	83.7 ± 1.8	**85.2 ± 2.1**
Lower Endplate Defects	83.3	86.9 ± 1.3	**87.5 ± 0.4**
Upper Marrow Changes	92.5	89.9 ± 2.2	**91.0 ± 1.3**
Lower Marrow Changes	91.4	90.1 ± 1.7	**90.3 ± 2.1**
Spondylolisthesis	89.6	92.9 ± 0.7	**95.2 ± 0.0**
Central Canal Stenosis	79.7	94.3 ± 0.0	**94.3 ± 0.9**
Average	82.5	85.7 ± 0.9	**86.3 ± 0.3**

Table 4.3: ***2D vs 3D.*** *The performance (%) of various models on the test set. "**Intra-rater**" is the intra-rater agreement. "**2D**" and "**3D**" are the results from the 2D and 3D CNNs, two models each with swapped validation and test sets.*

performance. Again, we suspect that this lack of improvement is due to insufficient training data so that we are not able to take advantage of the additional capacity of the deeper networks. We also experimented with a 10 fold-cross validation split of the dataset using the 3D CNN, ten folds to keep the consistency of the 90:10 train:validation split, but due to time constraint only manages to run four out of the ten folds. We find that the performance on the test sets of the four folds us similar swapping the validation:test sets. Overall, we find that 3D CNNs work better for our use case.

4.3.5 Adding Disc Level Supervision

Since all the disc volumes extracted are used as independent samples without any information regarding the disc level, we also experimented on adding more disc level supervision into the network. The intuition was that the disc level might improve performance as discs of differing levels have different grading distributions: it has been observed that pathological cases appear significantly more on lower level discs (L4-L5 or L5-S1) when compared to upper level discs. We specified the lumbar disc level (which one out of the six) as a one-hot encoding vector, and added intermediate

fully-connected layers prior to concatenation with the **FC6** activations of each of the eight gradings (see Figure 4.5). However, we did not see any significant improvement in adding this supervision.

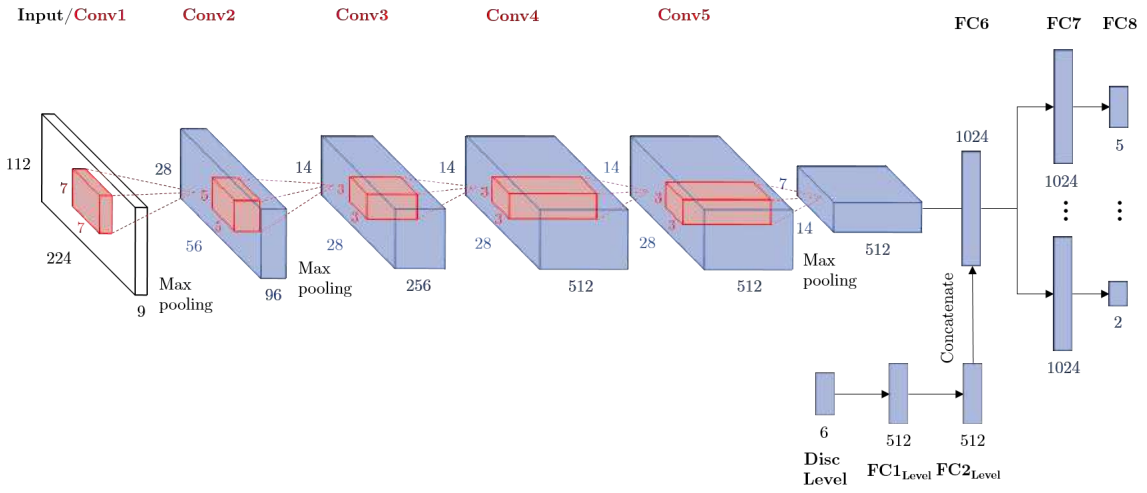


Figure 4.5: ***CNN Architecture – FC6 with Disc Level Encoding (Radiological Gradings Predictions).*** *Similar to the network shown in Figure 4.5 but with the addition of disc level information added as a one-hot encoding vector. The main idea was to let the network itself learn the prior of the disc levels and gradings.*

4.3.6 Comparison to Other Methods

We evaluate our performance on **Pfirrman grading** and **disc narrowing** classifications using the test set and evaluation protocol of Lootus (2015). It is important to note that for Lootus (2015), Pfirrmann grading is measured in terms of accuracy to ± 1 of the radiologist grade i.e. a prediction is deemed to be correct if the predicted grading is within 1 grade away from the ground truth grading, and disc narrowing grading is simplified to a binary classification of normal/abnormal discs where grade 1 is considered to be normal and grades 2 to 4 are abnormals. For the comparison, we adapt our gradings to match those of Lootus (2015). We surpass their performance by $+\mathbf{8.7}/+\mathbf{8.2}\%$ ($\mathbf{87.4}\% \rightarrow \mathbf{96.1}/\mathbf{95.6}\%$) for Pfirrmann grading, and $+\mathbf{4.1}/+\mathbf{6.3}\%$ ($\mathbf{83.7}\% \rightarrow \mathbf{87.8}/\mathbf{90.0}\%$) for disc narrowing; compared against our 2D/3D models.

We also compare our **central canal stenosis** performance against Zhang et al. (2017) at different disc levels where we obtain a comparable performance when comparing their results on their data against our results on our test set: $+$**7.5%** (**87.2%** $\rightarrow$ **94.7**) for L3-L4, $+$**0.8%** (**85.1%** $\rightarrow$ **85.9**) for L4-L5, and $+$**6.2%** (**87.5%** $\rightarrow$ **93.7**) for L5-S1). Note that the method described by Zhang et al. (2017) uses axial images, the standard for stensosis classification, while we use sagittal.

4.3.7 Limitations

We only use T2-weighted sagittal scans in our experiments while certain gradings are normally assessed with the addition of other modalities e.g. axial scans are commonly used to assess **central canal stenosis**, and Modic changes grading typically requires both T1-weighted and T2-weighted scans. We compensate for the classification of Modic changes by only looking at Modic changes specific to T2-weighted scans. This is why we call our grading, **marrow changes**. We are able to achieve reasonable performance in **central canal stenosis** with only sagittal scans.

4.4 Summary

In this chapter, we have shown that multi-task training is extremely beneficial especially in medical vision problems where it is common to have multiple labels for a single modality. Unsurprisingly, training with multiple tasks in general improves over the performance of training on individual tasks. We also shown that the branch point for the multi-task CNN directly affect the performance; **Conv5** works best. Finally, we have shown that a 3D CNN is slightly better than a 2D one for our task of predicting radiological gradings of intervertebral disc volumes. Moving on, in Chapter 5 we look at extracting localization of pathologies or in a way evidence of the predictions using CNNs trained only on classification described in this chapter.

Chapter 5

Evidence Hotspots

In the previous chapter, we trained CNNs on radiological grading classification given a disc volume input. In this chapter, we discuss how to use those CNNs trained for classification to produce heatmaps of pathological voxels in the disc. The heatmaps or the localization of these pathologies are achieved implicitly through training the network for the classification tasks, and no other labels are needed apart from the disc-level classification labels. We believe that the integration of automated quantitative scores into clinical practice can be aided if the system can highlight the regions within the image that lead to the prediction. However, the cost of obtaining such image mark-up, e.g. segmentation masks, can be prohibitive. Moreover, it is often difficult for a clinician to identify precisely the pixels or voxels that resulted in their opinion. Often, it is the overall appearance of an area that leads to the conclusion. To this end, we leverage the ability of CNNs to learn to localize important contributing image regions when trained for classification tasks. Given an input intervertebral disc volume, we show that a trained CNN, that predicts eight different radiological gradings, is able to produce eight heatmaps, each one showing possible pathological evidence unique to each grading, hence evidence hotspots. These hotspots not only give credence to the predictions but is also a good visual companion piece to just raw classifications.

This chapter can be broken down into three main sections; the first one being the details of several different methods in acquiring saliency map or heatmap of prediction in Section 5.1, followed by details of the test time augmentation scheme in Section 5.2, and finally the experiments and results in Section 5.3. A simplified view of the overall pipeline is given in Figure 5.1.

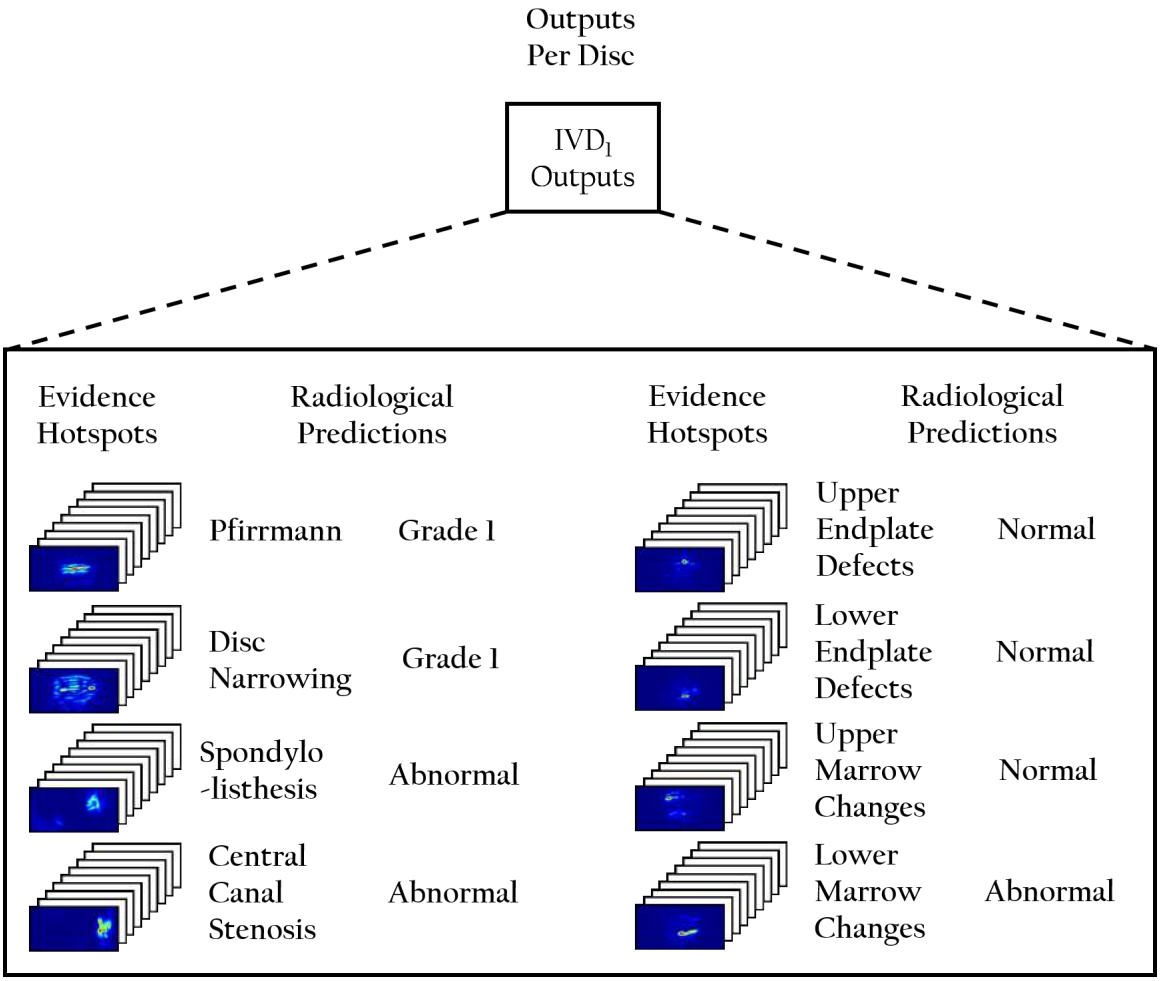


Figure 5.1: ***Output of the CNN.*** *For each of the six intervertebral disc output,* $\mathbf{IVD_1}$ *to* $\mathbf{IVD_6}$*, to the pipeline in Figure 4.1, there are eight radiological grading predictions and the associated evidence hotspot heatmaps. The predictions are numerical values of the gradings while the heatmaps show probable voxel responsible for the prediction (hotter/redder = more probable). The dimension of each evidence hotspot heatmap is* $112 \times 224 \times 9$ *which is also the dimension of the input volume of the intervertebral disc consumed by the CNN described in Chapter 4.*

5.1 Visualizing Evidence Hotspots

We show here in this section, that a network trained for a fine-grained classification task can produce a heatmap which pinpoints the region in the image space responsible for the prediction. This map lights up pathological areas, 'hotspots' of the prediction, specific to the trained task in the image. Only the class labels attached to the disc volume are needed in training without any other extra supervisory information. We term the heatmap produced 'evidence hotspots'. In the following sub-sections, we consider three different saliency extraction methods to produce the evidence hotspots: (1) backpropagation, (2) guided backpropagation, and (3) excitation backpropagation. Another comparable method is the class activation mapping by Zhou et al. (2016), but since it can only be produced by networks with a global average pooling layer, we could not directly compare against this method since the networks we trained in Section 4 do not possess such a layer.

5.1.1 Saliency by Backpropagation

Simonyan et al. (2014) proposed a method that ranks the influence of an image, x, according to its influence on a specific class score. The method proceeds by linearizing the relationship between a specific output class score and the input x as prediction of a specific class, y. For CNNs, we can approximate the highly non-linear function of y to be

$$y \approx w^{\mathbf{T}} x + b \tag{5.1}$$

where w and b are the weight and bias of the model. So ranking the influence of the input, x, can be posed as ranking the magnitude of the weight, w, that influences the output y. The weight can be obtained as

$$w = \frac{\partial y}{\partial x} \tag{5.2}$$

which can be found via backpropagation. However, unlike Simonyan et al. (2014), where the input x is a 3 channel RGB image ($z = 3$), our input consists of 9 channels ($z = 9$), each a greyscale image of an MRI slice in the disc volume. Furthermore, instead of producing a single channel saliency map calculated from the maximum magnitude of w,

$$\mathcal{M} = max_z|w| \tag{5.3}$$

as in the case for Simonyan et al. (2014), the saliency map is also 9-dimensional, such that

$$\mathcal{M}_z = |w_z| \tag{5.4}$$

where $z \in \{1 \dots 9\}$ since each input channel corresponds to an actual slice of the disc volume. The main visual difference from our maps to the maps shown by Simonyan et al. (2014) is that our salient regions are more localized and more specific to the area that is the cause for the classification. We suspect that this might be because our classification tasks are more fine-grained, and because our input images are visually very similar.

5.1.2 Saliency by Guided Backpropagation

Guided backpropagation proposed by Springenberg et al. (2015) is a slightly modified, compared to Simonyan et al. (2014), saliency extraction method also based on backpropagation. The proposed method changes the gradient, δ, of ReLUs during backpropagation. As we are only interested in finding the derivative of the class score with respect to the input, the gradient, δ, here refers to the derivative with respect to the input (or activations in the case of gradients of intermediate layers), x, and not the weights of the CNN. During standard backpropagation, a ReLU unit,

$$y = max(0, x) \tag{5.5}$$

backpropagates the gradient of the layer next to the ReLU but for only the positive portion of the input

$$\delta_{i-1} = \delta_i \cdot [x > 0] \tag{5.6}$$

where δ is the gradient and i denotes the layer of the gradient i.e. δ_i is the gradient of the layer succeeding the ReLU and δ_{i-1} is the gradient after backpropagating through ReLU (gradient after backward pass through ReLU). Springenberg et al. (2015) further restricts the gradient of the ReLU unit by only propagating positive gradients with positive inputs

$$\delta_{i-1} = \delta_i \cdot [x > 0] \cdot [\delta_i > 0] \tag{5.7}$$

This is in a way a combination of normal backpropagation used by Simonyan et al. (2014) and the 'deconvnet' method proposed by Zeiler and Fergus (2014) which has a ReLU gradient that depended only on the positivity of the gradient of the next layer

$$\delta_{i-1} = \delta_i \cdot [\delta_i > 0] \tag{5.8}$$

5.1.3 Saliency by Excitation Backpropagation

Zhang et al. (2016) proposed a method based on a probabilistic Winner-Take-All process. Similarly to guided backpropagation, this method changes the backpropagation of the gradient by looking at both the gradient δ_i, and the input, x. Unlike guided backpropagation however, only gradients of convolutional and the average pooling layers are changed. The methods starts by thresholding the convolutional weights

$$w^+ = max(0, w) \tag{5.9}$$

and running a forward pass producing an intermediate output y^+. The gradient δ_i is then normalized by a factor of y^+ via element-wise division:

$$\hat{\delta}_i = \delta_i \oslash y^+ \tag{5.10}$$

Then the backpropagated gradient δ_{i-1} is calculated from the normalized $\hat{\delta}_i$ via a backward pass, which is then normalized via element-wise multiplication with the input:

$$\hat{\delta}_{i-1} = \delta_{i-1} \odot x \tag{5.11}$$

Zhang et al. (2016) also suggested an extension to their method called contrastive excitation backpropagation which was shown to be better and we follow this in our implementation. Contrastive excitation saliency, Δ_c, can be found by backpropagating the difference between the gradients of the penultimate layer:

$$\Delta_c = \delta_i^+ - \delta_i^- \tag{5.12}$$

where δ_i^+ is the gradient calculated via positive weights w^+ and δ_i^- is the gradient calculated via the negation of the positive weights $-w^+$. For our implementation, we backpropagate only up till **Conv1** which gives us visually better results. As in Zhang et al. (2016), we find that backpropagating back to the input space does not give good saliency maps.

5.2 Test Time Augmentation

Instead of just one standard run of backpropagation, we find that aggressive test time augmentations is key to producing more salient localized hotspots. The final heatmap is computed from the average of multiple saliency maps produced from randomly augmented images using our training augmentation scheme; each resulting saliency

map is transformed back to the original image space with the reverse of the applied augmentation. We use around 150 augmentations (each requiring a backpropagation pass) which takes around 90 seconds to complete for a single disc volume. It can be seen in Figure 5.2 that test time augmentation helps improve the hotspots essentially by smoothing the heatmaps.

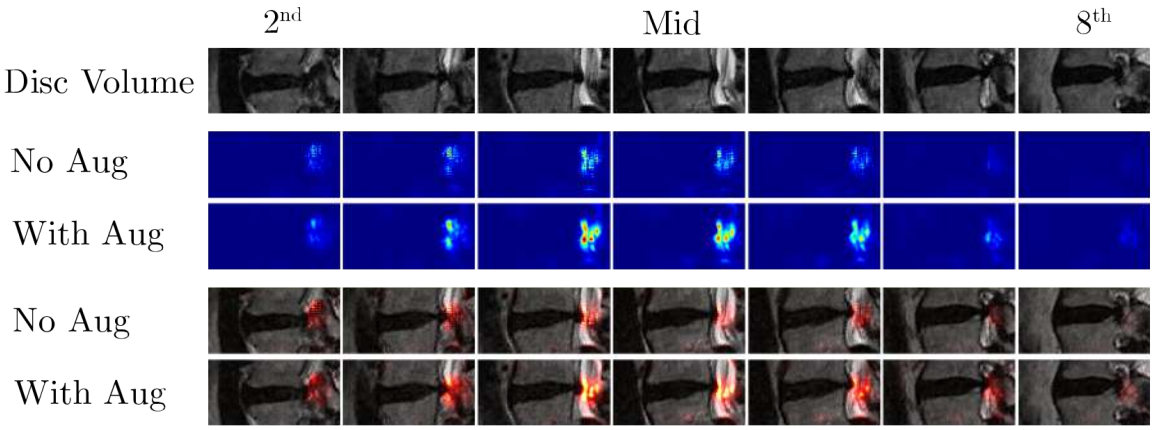


Figure 5.2: ***Test Time Augmentation.*** *Example of* ***central canal stenosis*** *example taken from the test set alongside its hotspot via excitation backpropagation with and without test time augmentation. The top row shows images of the slices of a disc volume (slice 2 to slice 8) followed by two rows of the heatmaps of the hotspots, and two rows of the heatmaps overlaid on the image slices. Note the improvement in hotspot visibility brought by the test time augmentation.*

5.3 Experiments & Results

5.3.1 Qualitative Results

We use the 3D CNN trained in Chapter 4 to predict the different gradings to visualize the hotspots. We find no significant difference in terms of hotspots quality between 2D and 3D CNNs for standard and guided backpropagations. Since the 2D CNN essentially merges the slices of the disc volume after **Conv1**, it is necessary to backpropagate all the way back to the input space to get the hotspot for each channel/slice in the disc volume. This in turn results in poorer heatmaps when using

a 2D CNN and excitation backpropagation (which works poorly when propagated all the way back to the input layer). This is different in the 3D case, where we can backpropagate to any layer in the network (**Conv1** is best for our dataset).

Figure 5.3 and Figure 5.4 compare the three methods we experimented on, in each case using test time augmentation. Qualitatively, excitation backpropagation produces the best results in our dataset with cleaner hotspots and better localizations. For example, in Figure 5.3, the upper endplate defect was correctly highlighted only by backpropagation and excitation backpropagation. Similarly, in Figure 5.4, the lower marrow change was correctly highlighted only by the guided and excitation backpropagations.

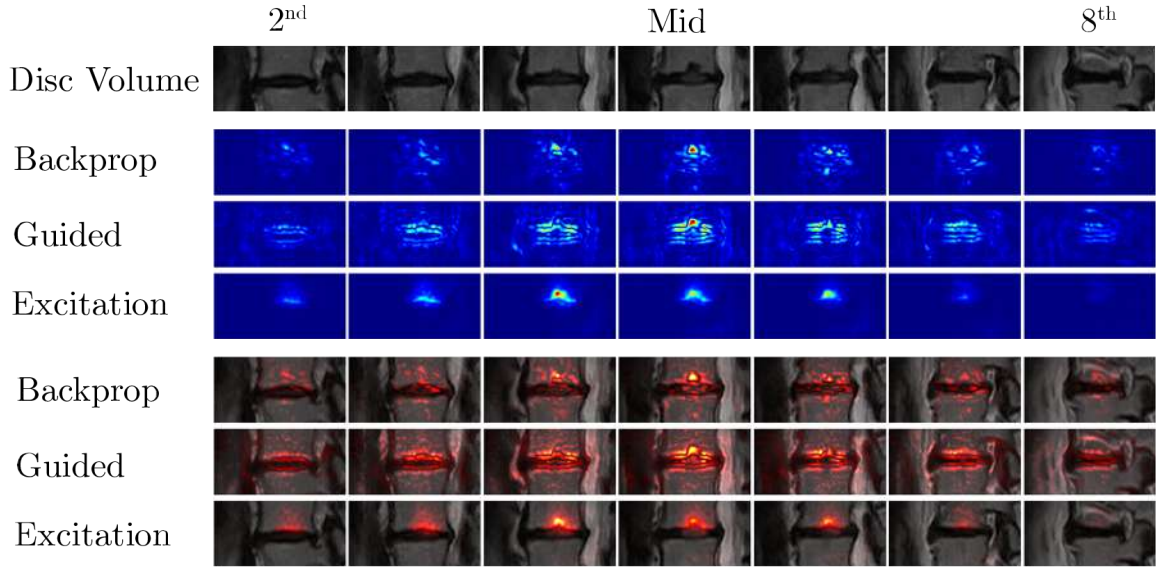


Figure 5.3: ***Hotspot Comparison 1.*** *Example of upper endplate defects example taken from the test set. Topmost set of images are slices of a disc volume (slice 2 to slice 8) followed by: (i) the heatmaps of the hotspots, and (ii) the heatmaps overlaid on top of the disc volume. We show the three different methods of producing hotspots with test time augmentation. The excitation method gives the best visibility of the three.*

Figure 5.5 shows hotspots (via excitation backpropagation) on different tasks or gradings on randomly selected discs in the test set. It can be seen that the hotspots of endplate defects and marrow changes appear to highlight the respective abnormalities in the endplate regions of the vertebral bodies. Interestingly, since spondylolisthesis

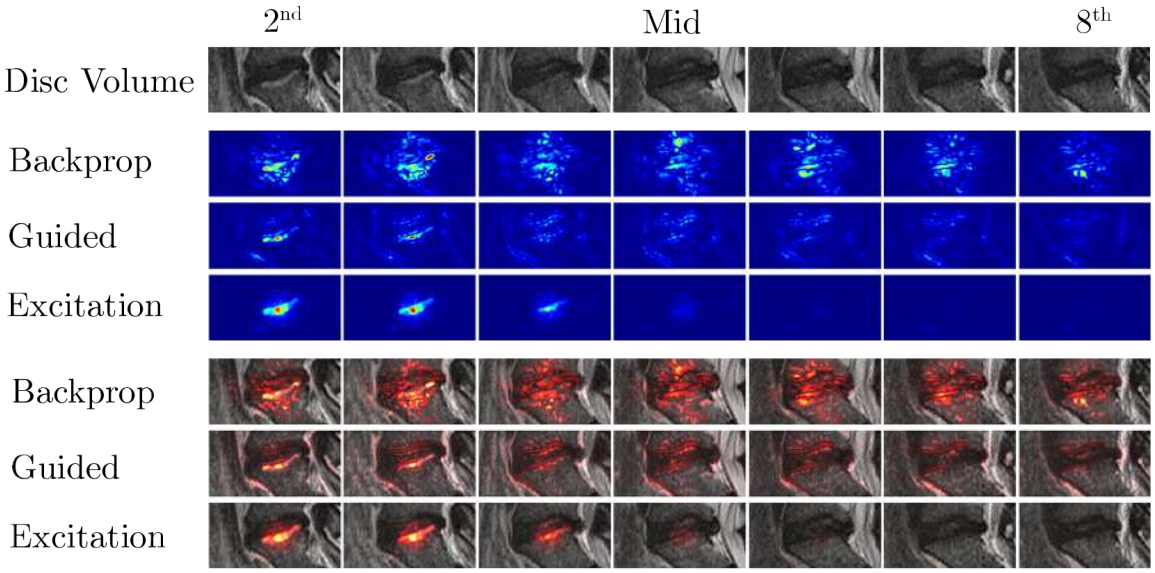


Figure 5.4: ***Hotspot Comparison 2.*** *Example of lower marrow change example taken from the test set. Topmost set of images are slices of a disc volume (slice 2 to slice 8) followed by: (i) the heatmaps of the hotspots, and (ii) the heatmaps overlaid on top of the disc volume. We show the three different methods of producing hotspots with test time augmentation. The excitation method gives the best visibility of the three.*

is the measure of the vertebral slip, the hotspots highlight: (i) both vertebral bodies, and (ii) their misalignment near the posterior. Similarly, in the case of central canal stenosis the hotspots highlight the specific region of only the pinched area near the posterior of the disc and not the whole canal.

Since excitation backpropagation is qualitatively the best method to produce evidence hotspots, we produce several examples of the hotspots for each grading learnt in Chapter 4: **Pfirrmann grading** in Figure 5.6, **disc narrowing** in Figure 5.7, **upper endplate defects** in Figure 5.8, **lower endplate defects** in Figure 5.9, **upper marrow changes** in Figure 5.10, **lower marrow changes** in Figure 5.11, **spondylolisthesis** in Figure 5.12, and **central canal stenosis** in Figure 5.13.

5.3.2 Quantitative Results

To validate the evidence hotspots produced by the 3D CNN we compare them against 'ground truth' bounding boxes annotated by a spinal surgeon, Professor Jeremy Fair-

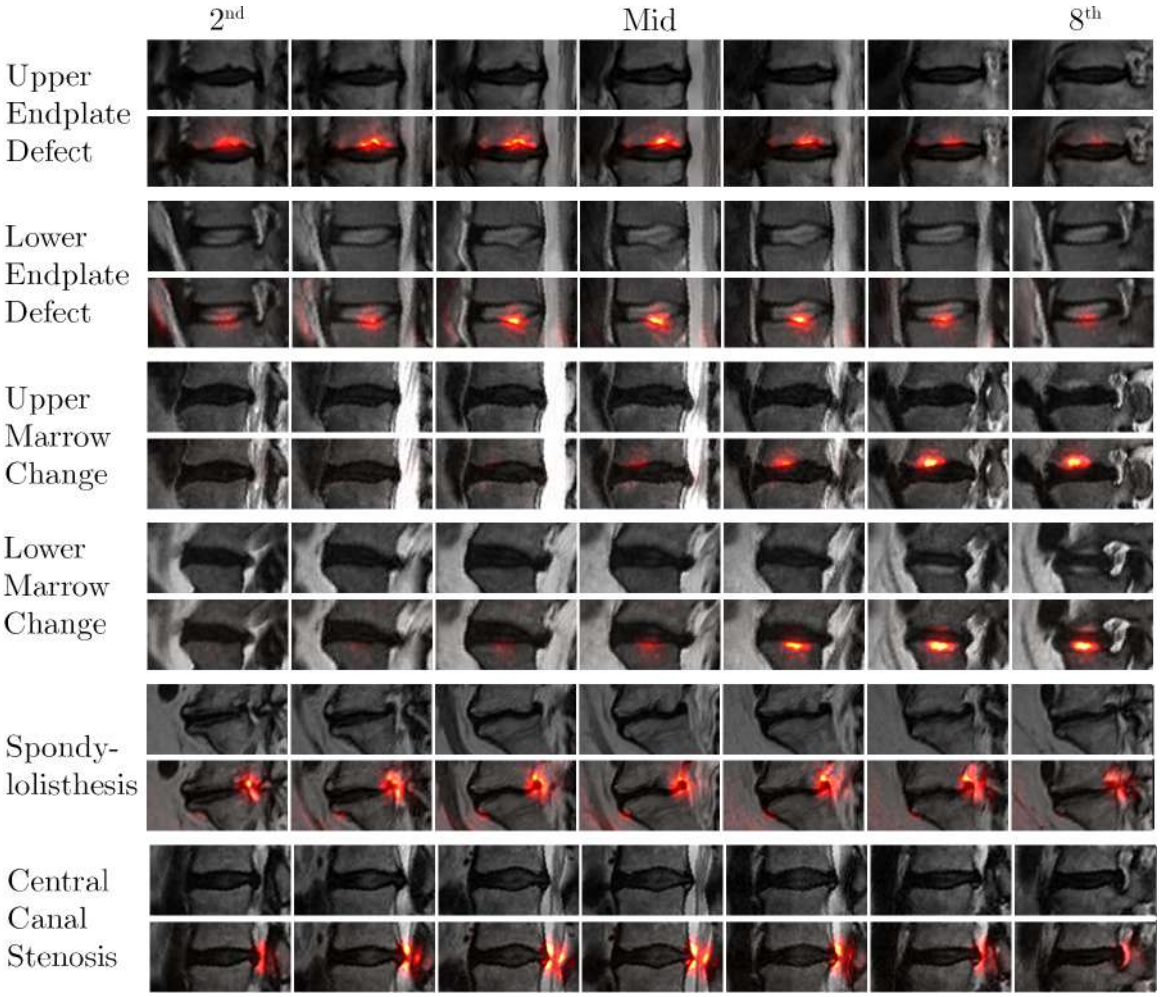


Figure 5.5: ***Evidence Hotspots of Different Gradings.*** *Examples of disc volumes (upper in each pair) and their corresponding evidence hotspots (lower in each pair). The leftmost and rightmost columns of images are the second and eighth slice for each disc, out of the full volume of 9 slices. Going from top to bottom are examples for each of the binary gradings: (i)* ***upper endplate defects****, (ii)* ***lower endplate defects****, (iii)* ***upper marrow changes****, (iv)* ***lower marrow change****, (v)* ***spondylolisthesis****, and (vi)* ***central canal stenosis****. Only pathological/abnormal examples are shown for each radiological grading/classification task, i.e. endplate defects appearing as protrusions of the discs into the vertebral bodies, and marrow changes appearing as localized discolourations of the vertebral bodies near the vertebral endplates. Note that these hotspots localize extremely well to the assigned tasks e.g. in the lower endplate defects example the hotspots appear only in the lower endplate even though there are defects on the upper endplate. These examples are randomly selected on different patients in the test set.*

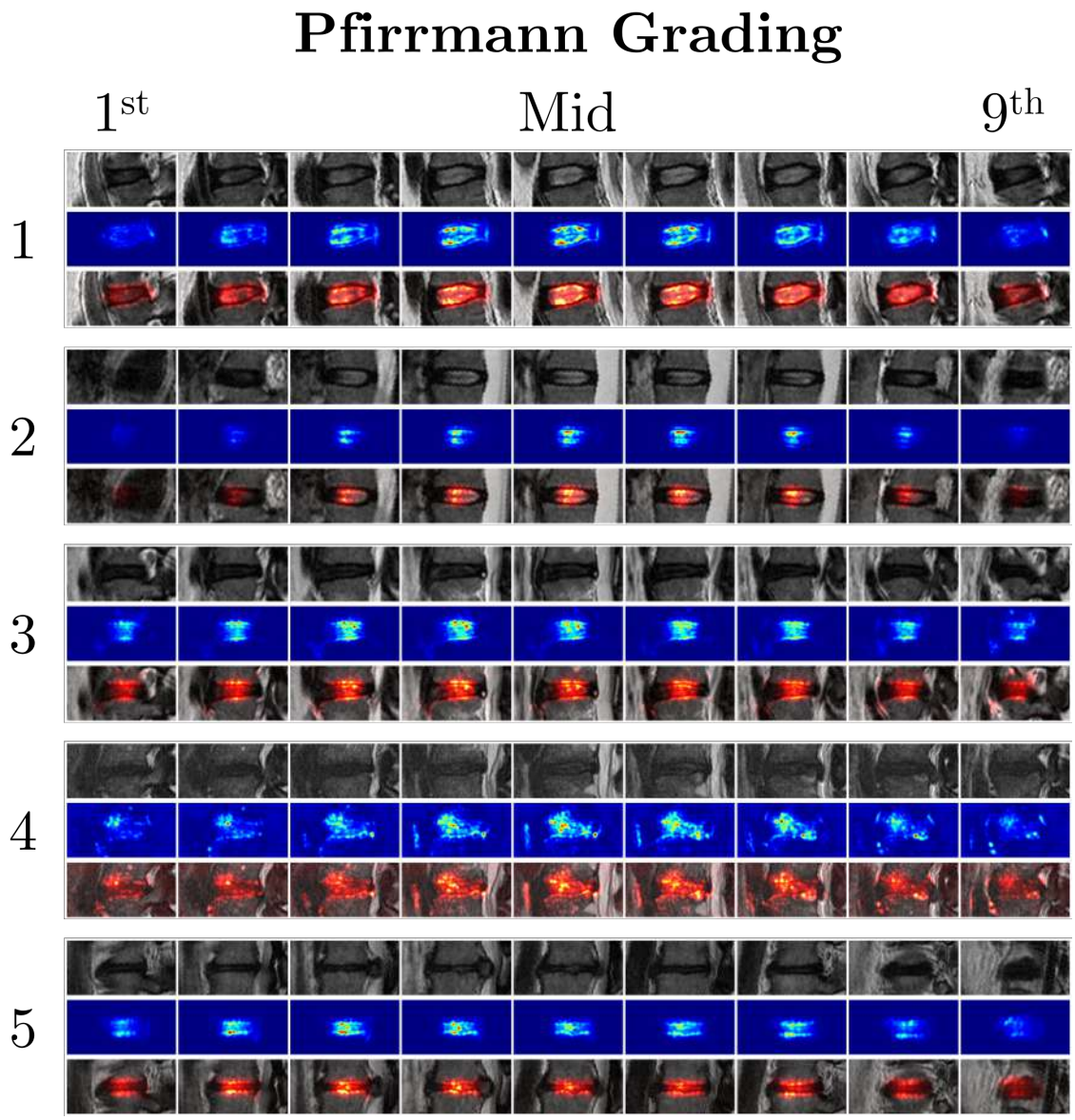


Figure 5.6: ***Hotspots – Pfirrmann Grading.*** *Evidence hotspots of* ***Pfirrmann grading*** *from Pfirrmann Grade 1 to Pfirrmann Grade 5. Notice that the hotspots tend to appear brighter near the mid-sagittal slices which is consistent to the definition of* ***Pfirrmann grading****.*

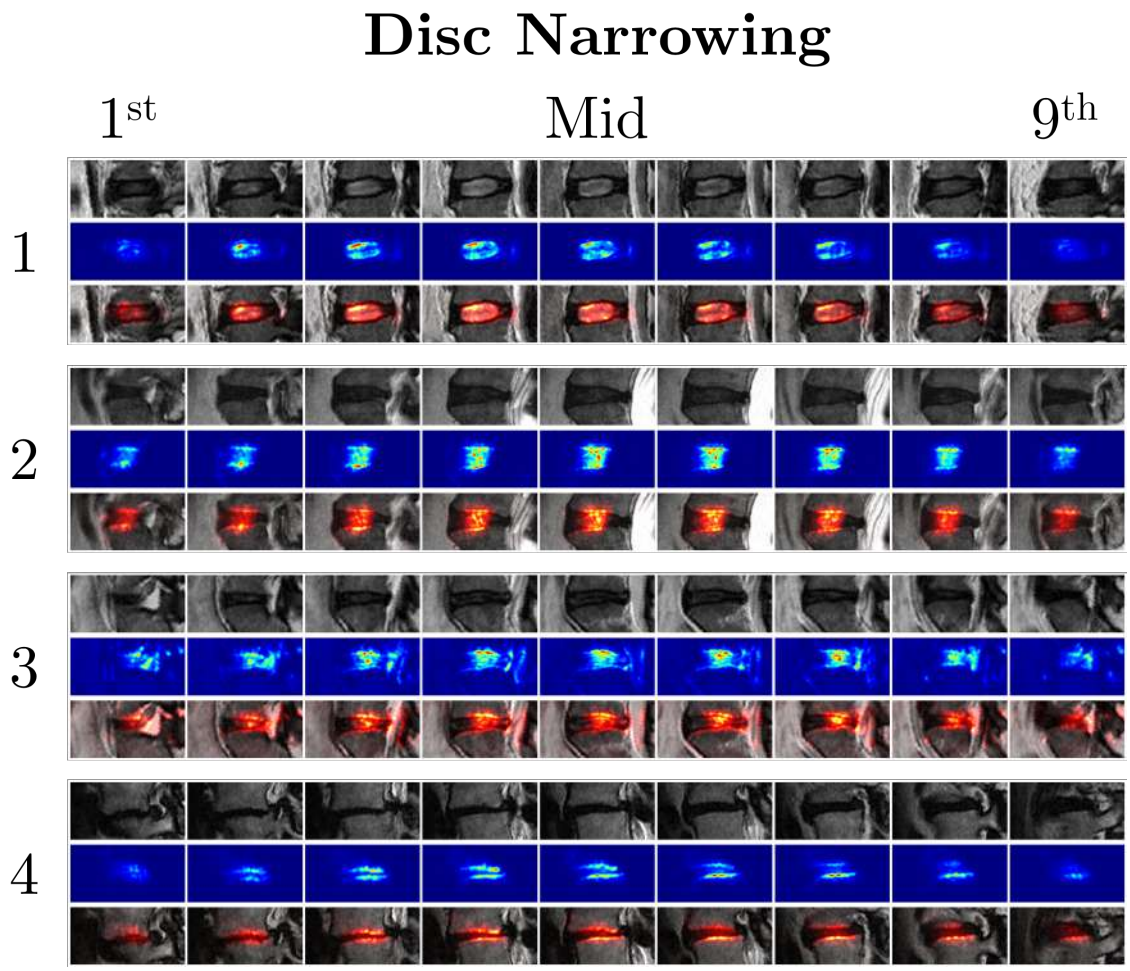


Figure 5.7: ***Hotspots – Disc Narrowing.*** *Evidence hotspots of* ***disc narrowing*** *examples graded from 1 to 4 in order of severity, with 1 for normal discs and 4 for extremely narrow discs. Similar to* ***Pfirrmann grading****,* ***disc narrowing*** *was also graded by the radiologist based only on the mid-sagittal slice on each disc. Hence, evidence hotspots appear brighter on the mid-sagittal slices in the example.*

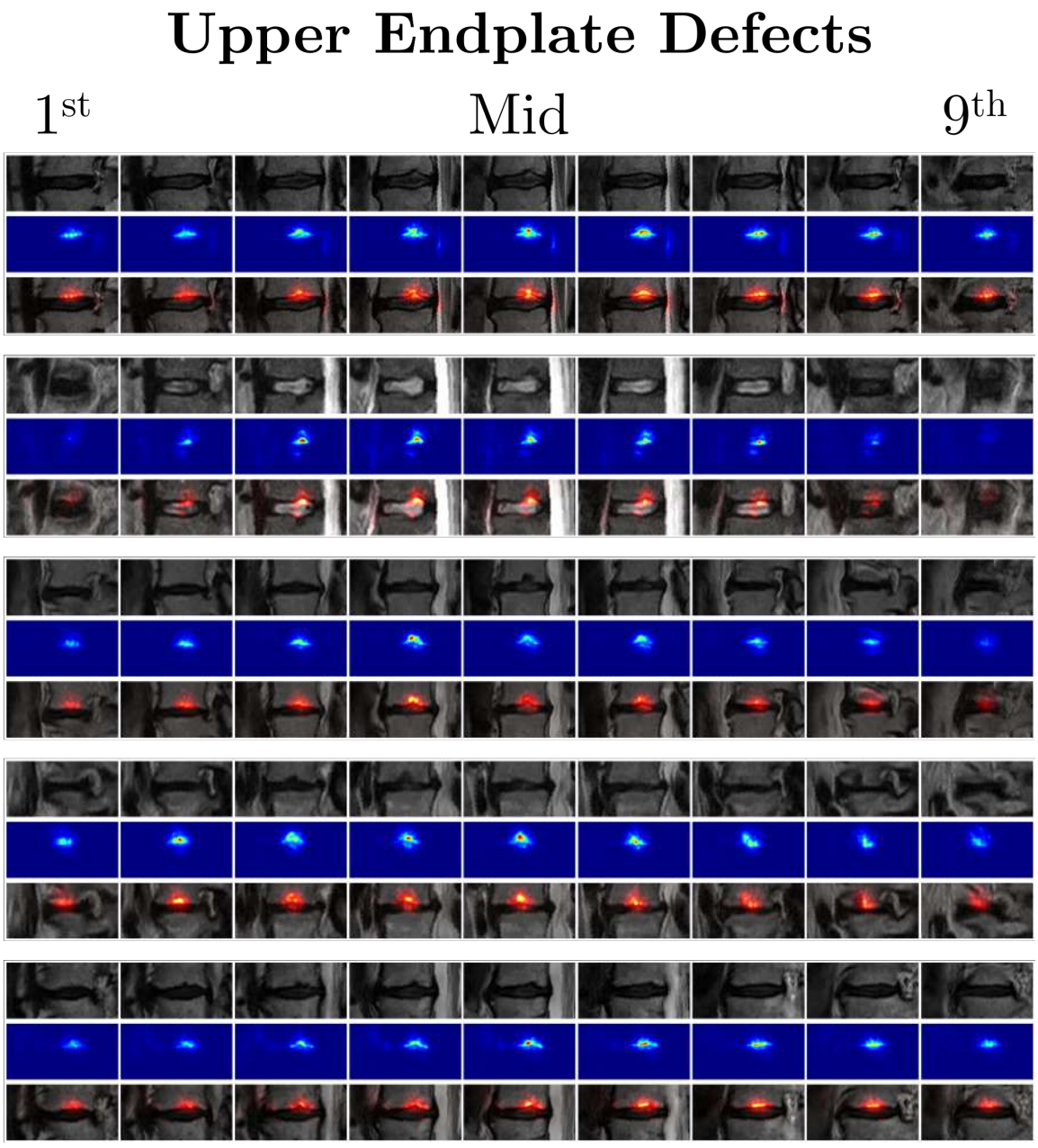


Figure 5.8: ***Hotspots – Upper Endplate Defects.*** *Examples evidence hotspots for disc volumes with* ***upper endplate defects****. Shown here are just the hotspots for just the pathological samples.*

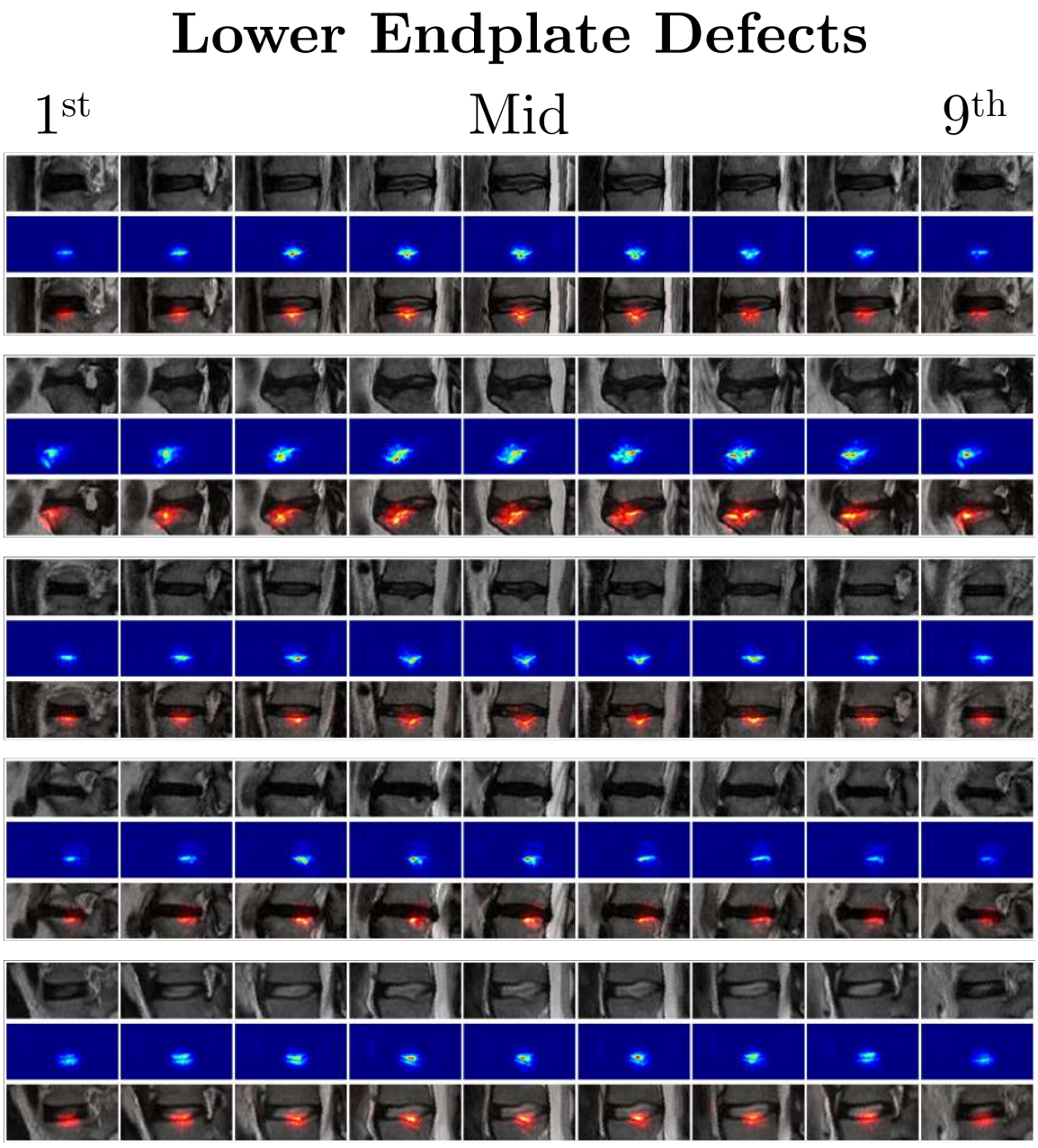


Figure 5.9: ***Hotspots – Lower Endplate Defects.*** *Examples evidence hotspots for disc volumes with* ***lower endplate defects****. Shown here are just the hotspots for just the pathological samples.*

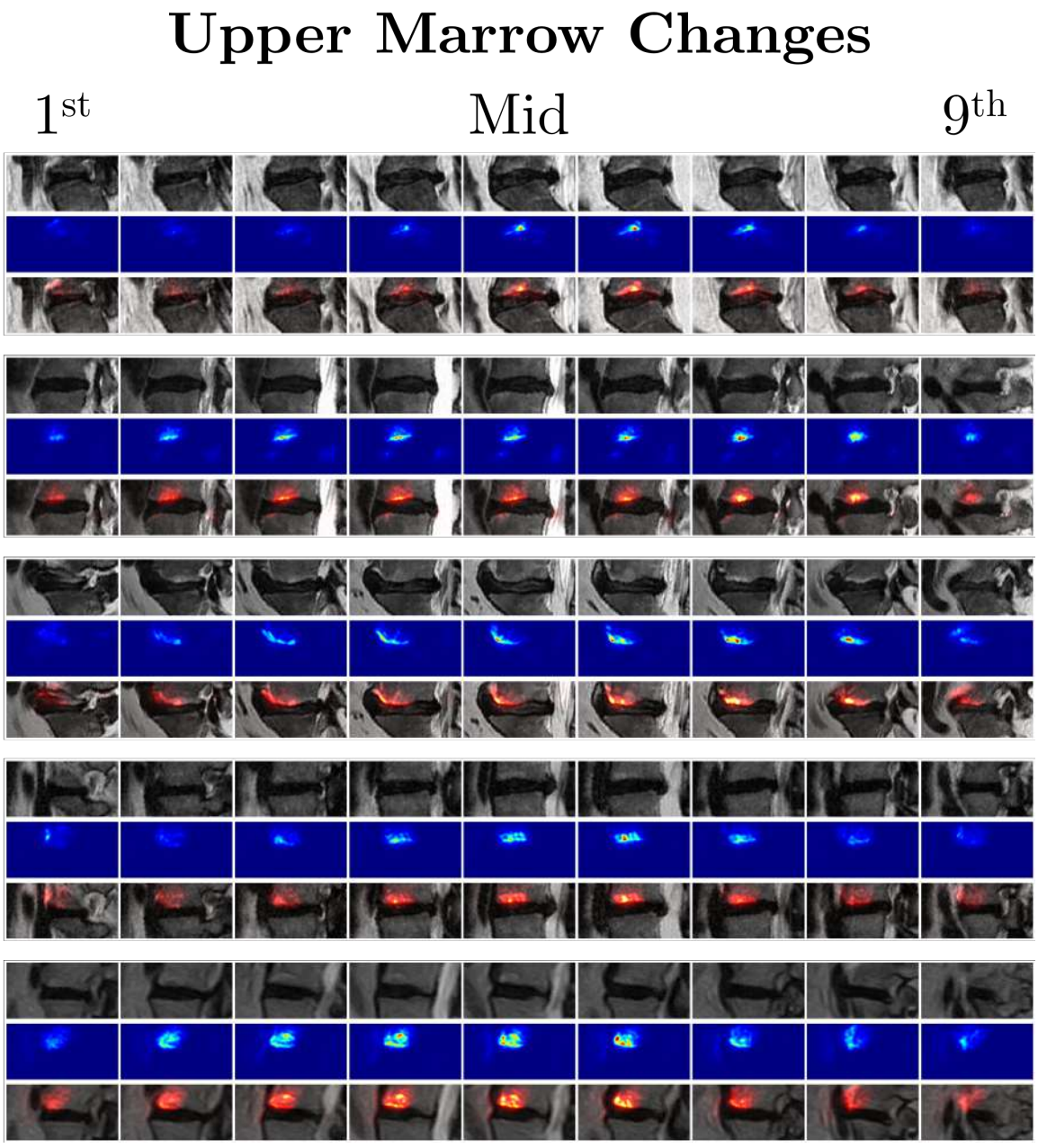


Figure 5.10: ***Hotspots – Upper Marrow Changes.*** *Examples evidence hotspots for disc volumes with* ***upper marrow changes****. Shown here are just the hotspots for just the pathological samples.*

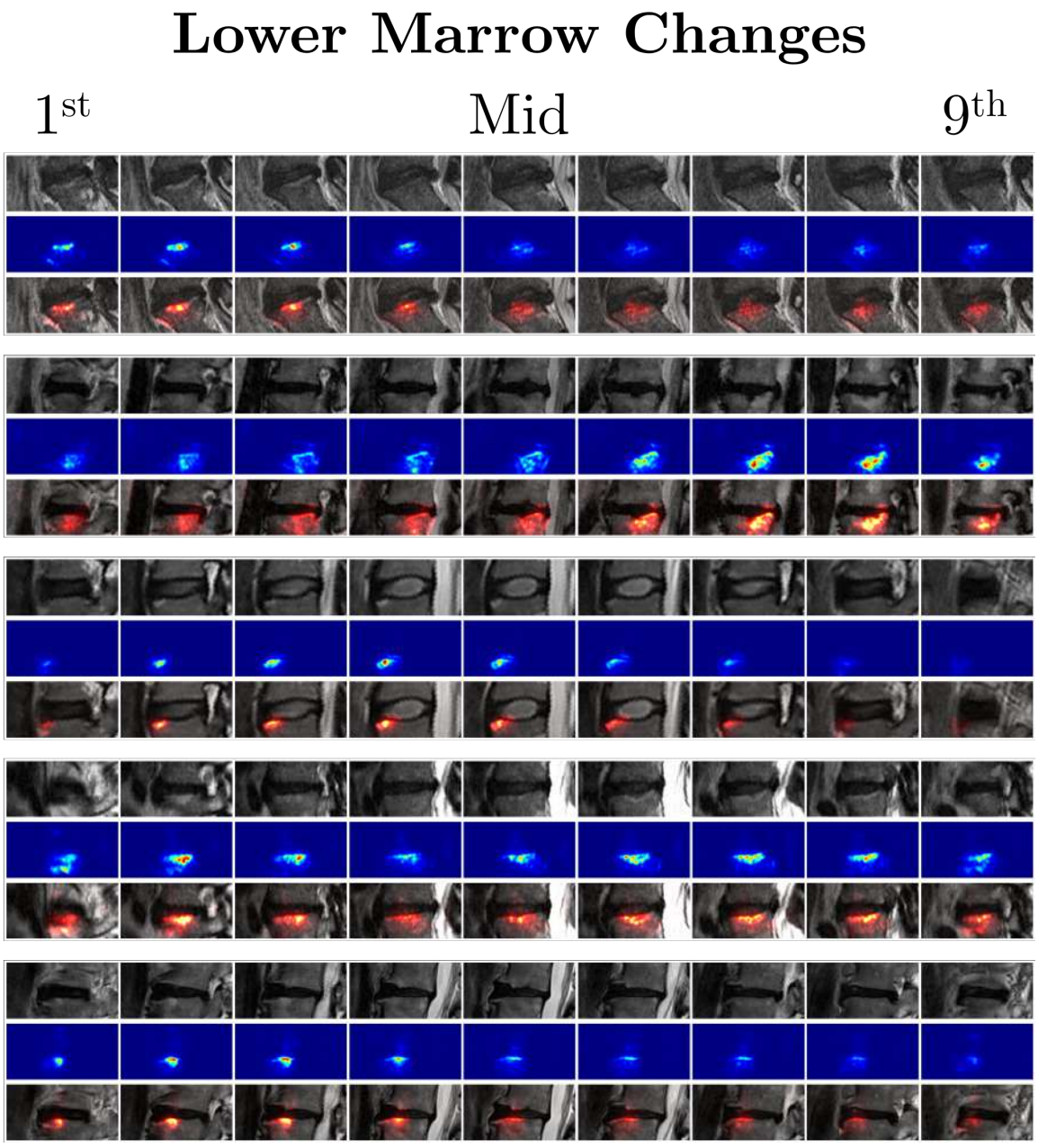


Figure 5.11: ***Hotspots – Lower Marrow Changes.*** *Examples evidence hotspots for disc volumes with* ***lower marrow changes****. Shown here are just the hotspots for just the pathological samples.*

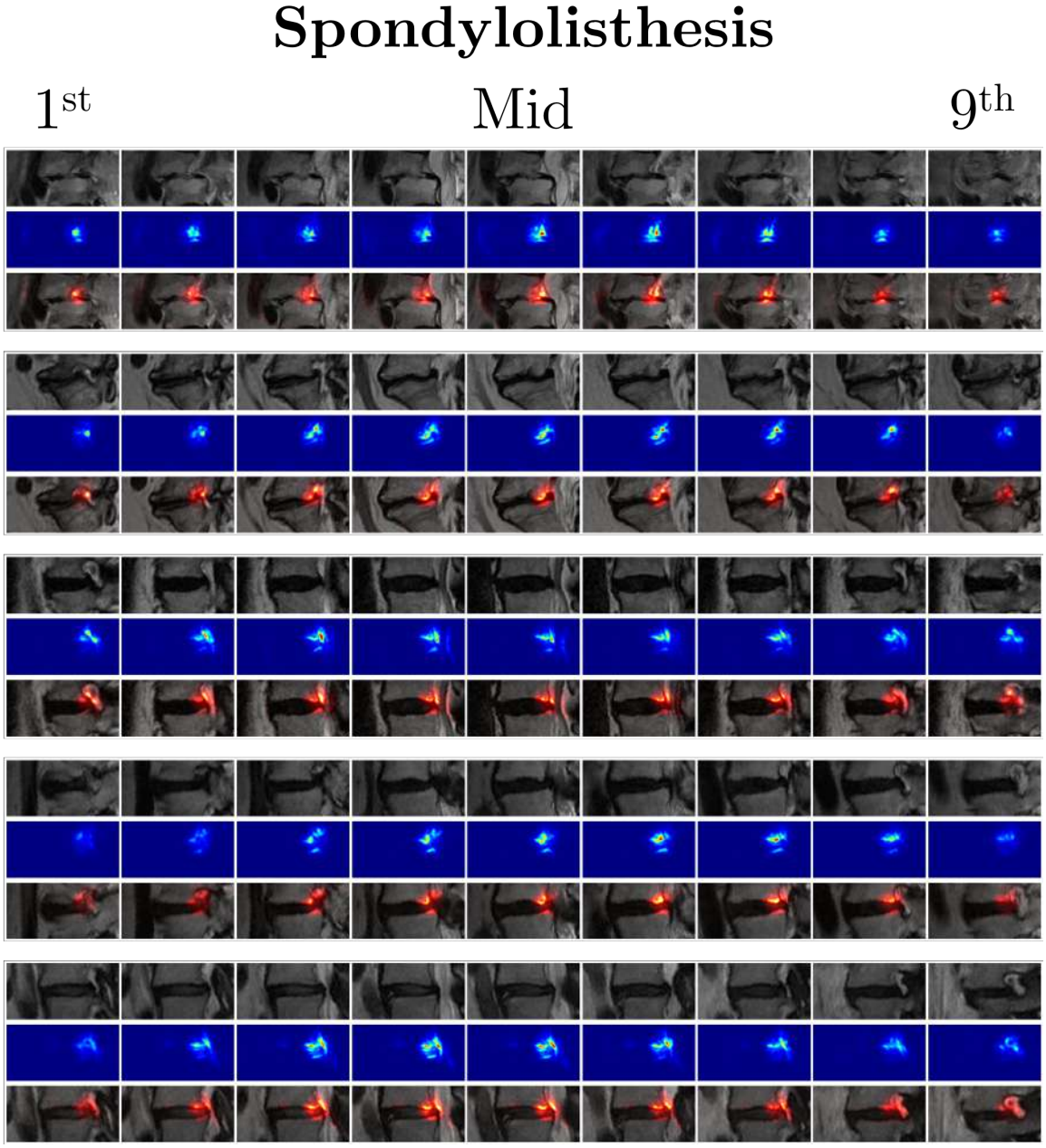


Figure 5.12: ***Hotspots – Spondylolisthesis.*** *Examples evidence hotspots for disc volumes with* ***spondylolisthesis****. Shown here are just the hotspots for just the pathological samples.*

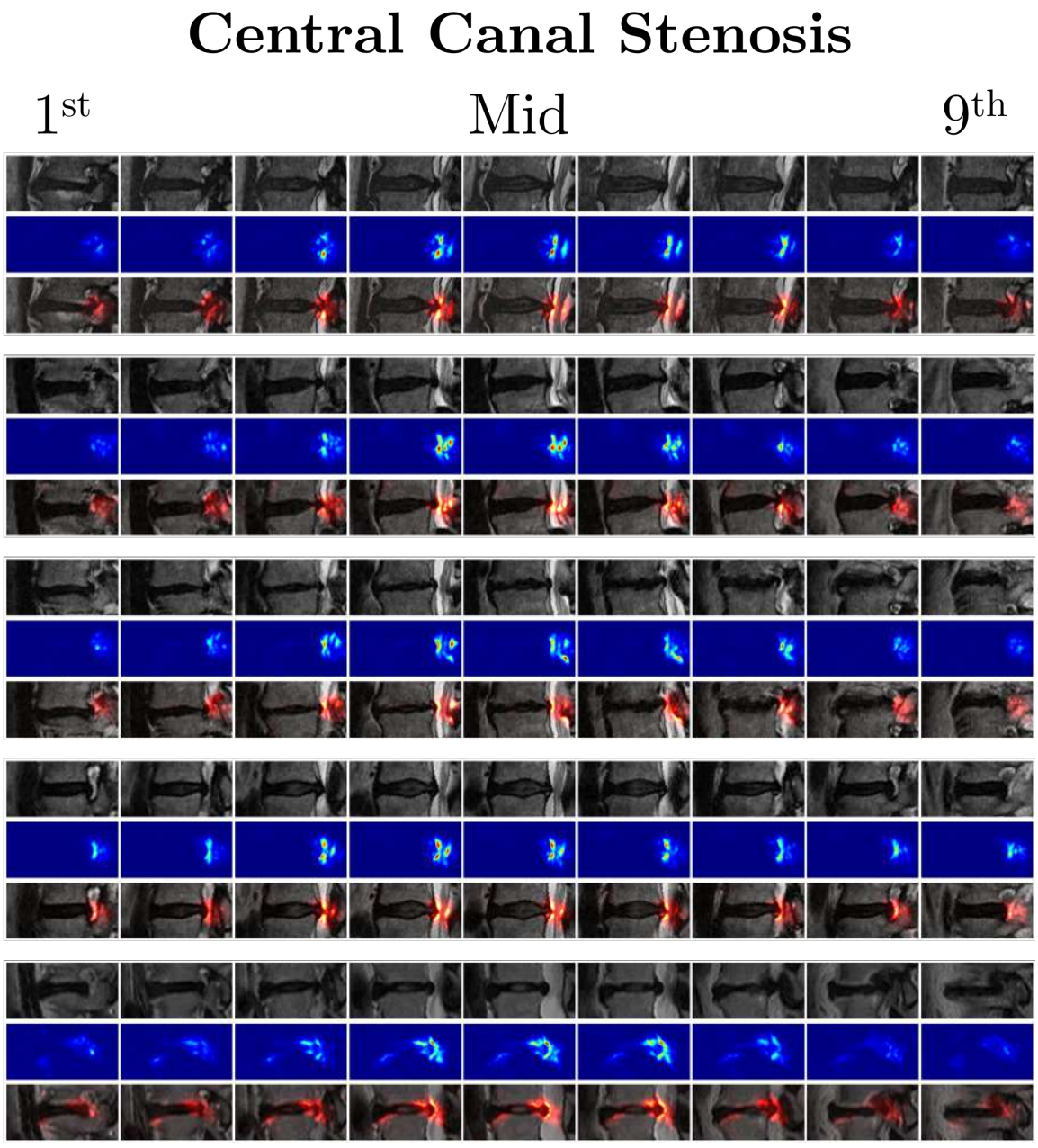


Figure 5.13: ***Hotspots – Central Canal Stenosis.*** *Examples evidence hotspots for disc volumes with* ***central canal stenosis****. Shown here are just the hotspots for just the pathological samples.*

bank. These comparisons are done on 312 disc volumes in the test set that have been graded with one or more of: (i) **marrow changes**, (ii) **endplate defects**, or (iii) **central canal stenosis**. These discs were read by a spinal surgeon, who has access to both the disc level information and the radiological gradings reported by the radiologist. The task was to annotate the region that best supported the grading with a tight bounding box in each slice of the scan. Figure 5.14 shows the annotation tool used to annotate the bounding boxes.

Figure 5.14: ***Annotation Tool.*** *Annotation tool used to annotation the bounding boxes of the pathology. Annotation is done on a per slice basis, for all 9 slices.*

To measure the localization performance of the hotspots, we calculate the Jaccard similarity coefficient, or the Interesection over Union (IoU), between the ground truth bounding boxes and the hotspots. To do so, for a given input disc volume, we produce a binary mask by thresholding based on the percentile of the intensity distribution of the hotspot and fit a minimum bounding box for each mask. Examples of discs with the hotspots, predicted bounding boxes and 'ground truth' bounding boxes can be seen in Figure 5.15 while the effects of thresholding on the sizes of the fitted bounding boxes can be seen in Figure 5.16.

We find that thresholding around the $90^{th} - 99^{th}$ percentile consistently works

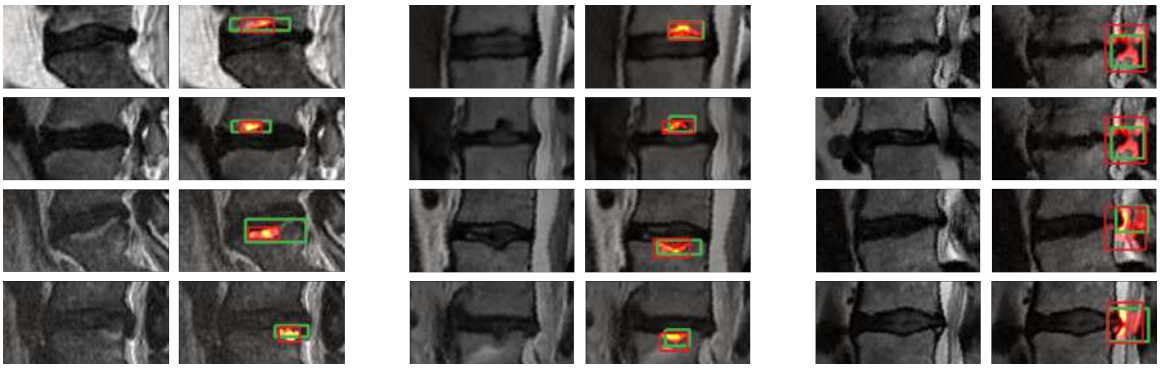


Figure 5.15: ***Bounding Boxes Examples.*** *Ground truth in green and predicted in red.* ***Left: marrow changes*** *examples.* ***Middle: endplate defects*** *examples.* ***Right: central canal stenosis*** *examples.*

Figure 5.16: ***Effects of Thresholding.*** *The threshold increases going from the left to right i.e. from the* 95^{th} *to the* 99^{th} *percentile. The ground truth bounding boxes are green in colour and the predicted bounding boxes are red.*

better than other percentiles. Since the magnitude of the hotspots is correlated with the certainty of the grading prediction, it is not surprising to see that higher thresholds have better localization performance. However, since the ground truth is constrained to regular boxes instead of delineated segmentations of the pathological regions, there is a trade-off in performance after a certain threshold unique to each grading. Table 5.1 shows the Jaccard/IoU of the three gradings.

Marrow Changes	0.242
Endplate Defects	0.149
Central Canal Stenosis	0.405

Table 5.1: ***Quantitative Performance.*** *Jaccard/IOU of the fitted bounding boxes against ground truth bounding boxes.*

As a baseline, we also calculate the Jaccard/IOU of predicting the whole of the disc volume as the detected bounding boxes (each box per slice) for each grading. Since sizes of the bounding boxes vary depending on the grading, **endplate defects** tend to be smaller than **central canal stenosis** for example, comparing them to

this baseline is a good indicator of the localization performance. Comparing each grading: **marrow changes** (**0.023** $\rightarrow$ **0.242**), **endplate defects** (**0.013** $\rightarrow$ **0.149**), and **central canal stenosis** (**0.053** $\rightarrow$ **0.405**). Overall, we see that the performance of the bounding boxes from the evidence hotspots for each grading, produced using the 3D CNN trained only for classification, is around 10 times better than the naive baseline. The Jaccard measure used however is still lower than expected and we suspect this is due to the strict nature of our annotated ground truth against our predictions; see Figure 5.17. Another example with **central canal stenosis** can be seen in Figure 5.18.

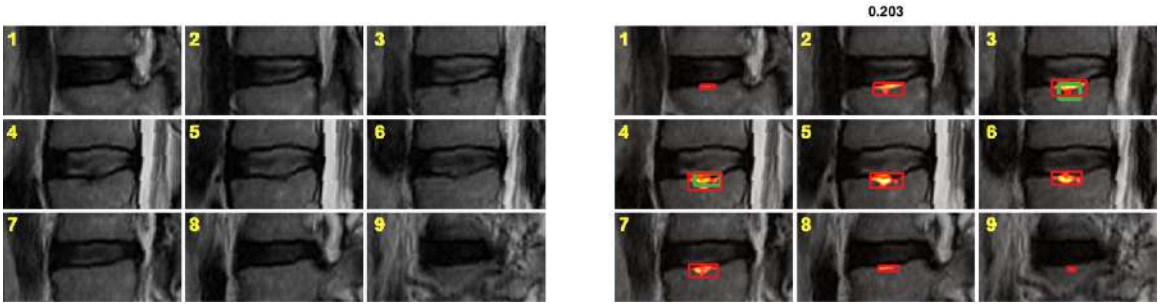


Figure 5.17: ***Slice-by-slice Quantitative Performance 1.*** *Shown is an example disc volume with a* ***lower endplate defect*** *in the* ***Genodisc*** *test set, where each slice is marked with its slice number, 1 to 9. The overall IoU for this volume is 0.203, with the predicted bounding boxes in red and the ground truth bounding boxes in green. The IoU per slice is as follows:* $1^{st}slice = 0$, $2^{nd}slice = 0$, $3^{rd}slice = 0.452$, $4^{th}slice = 0.557$, $5^{th}slice = 0$, $6^{th}slice = 0$, $7^{th}slice = 0$, $8^{th}slice = 0$, $9^{th}slice = 0$.

5.4 Summary

In this chapter, we have shown that we can produce high quality visualizations of pathology or evidence hotspots of the predictions from CNNs trained only on weak supervision i.e. class labels. We have also shown, at least in our use case, that excitation backpropagation produces the best saliency maps and from said heatmaps we can produce bounding boxes of the pathologies. So far, in predicting the radiological gradings and producing the evidence hotspots, we have only operated on T2-weighted

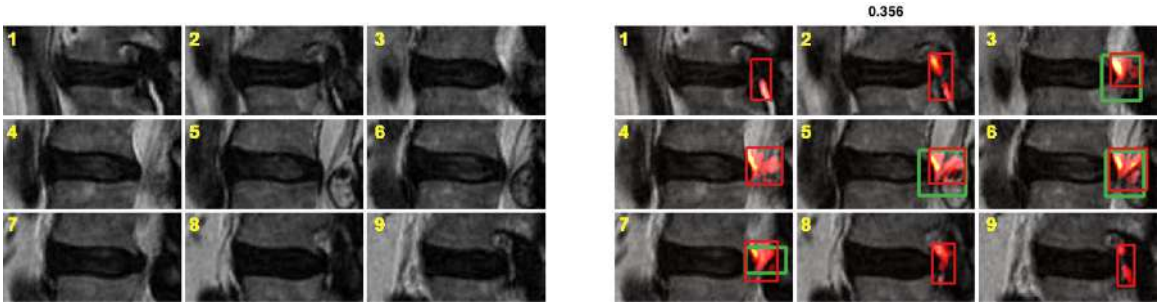


Figure 5.18: ***Slice-by-slice Quantitative Performance 2.*** *Shown is an example disc volume with a* ***central canal stenosis*** *in the* ***Genodisc*** *test set, where each slice is marked with its slice number, 1 to 9. The overall IoU for this volume is 0.356, with the predicted bounding boxes in red and the ground truth bounding boxes in green. The IoU per slice is as follows:* $1^{st}slice = 0$, $2^{nd}slice = 0$, $3^{rd}slice = 0.503$, $4^{th}slice = 0.0$, $5^{th}slice = 0.549$, $6^{th}slice = 0.701$, $7^{th}slice = 0.551$, $8^{th}slice = 0$, $9^{th}slice = 0$.

sagittal inputs but in Chapter 6 we look at predicting gradings from other modalities.

Chapter 6

Predicting Radiological Gradings using Other Planes, Sequences & Raw MRIs

Previously, we have only used T2-weighted sagittal scans for predicting radiological gradings which meant we might lose performance on certain gradings that depended on (i) both axial and sagittal scans for assessment e.g. **anterior disc bulging**, **posterior disc bulging**, **disc herniation** or (ii) both T1-weighted and T2-weighted scans e.g. **Modic/marrow changes**. In this chapter, we discuss how to learn to predict from multiple inputs, either inputs of different planes but the same sequence, T2-weighted axial and sagittal scans, or inputs of different sequences but the same plane, T1-weighted and T2-weighted sagittal scans. We also look at the possibility of directly predicting gradings from the raw MR scan without first extracting the disc volumes. As in Chapter 4, the method is trained, validated and tested on the **Genodisc** dataset (see Chapter 3).

In more detail, in Section 6.1 we introduce a novel disc volume extraction method in the axial plane that is built upon the method for detecting disc volume in sagittal

scans discussed in Chapter 3. We then train a model that jointly learn to predict radiological gradings from both axial and sagittal disc volumes. In Section 6.2 we experimented with adding an extra T1-weighted sagittal scan sequence in addition to the T2-weighted scan. Finally in Section 6.3, we look at predicting the gradings from only raw MR volume by directly classifying multiple disc gradings simultaneously.

6.1 Adding Axial Scans

A standard series of lumbar scans in the clinic normally consists of both axial and sagittal scans and typically show all the lumbar discs: T12-L1 to L5-S1 discs. Gradings attached to these discs are normally done by reading either the sagittal scan or axial scan, and sometimes both. The main issue we have with adding axial information is that instead of consistently having 6 discs per patient, the axial scans can be quite inconsistent in terms of field-of-view resulting in a variation of number of discs, ranging from 3 to 5 discs per patient. This inconsistency is due to the fact that axial scans are normally taken for just the lower lumbar discs.

Why Add Axial? Radiologists in the clinic normally have access to both sagittal and axial scans and there exists several gradings which require both for proper reading. This is due to the fact that certain features are much easier to see in the axial view e.g. the region containing both the posterior of the disc and dural sac. Since both scans are equally important we propose a multi-stream training scheme that jointly learn several gradings which require both axial and sagittal scans in **Genodisc** (see Chapter 3) namely: **anterior disc bulging**, **posterior disc bulging**, and **disc herniation**. Like before, we train a single model to predict the gradings simultaneously for each disc but now each disc is represented as two volumes, one extracted from the sagittal scan and one extracted from the axial scan.

6.1.1 Disc Volume Extraction – Axial

We use the detected disc volumes from sagittal scans described in Section 4.1.1 to find the intersection between the axial planes and the disc. Since the detected disc volumes possess level labels, for each disc we now know the corresponding axial slices. In an ideal case, as in Figure 6.1, where the axial slices are scanned roughly parallel to the orientation of the disc we can just use the intersections between the sagittal disc volumes and the axial planes as the disc volume detections of the axial slices.

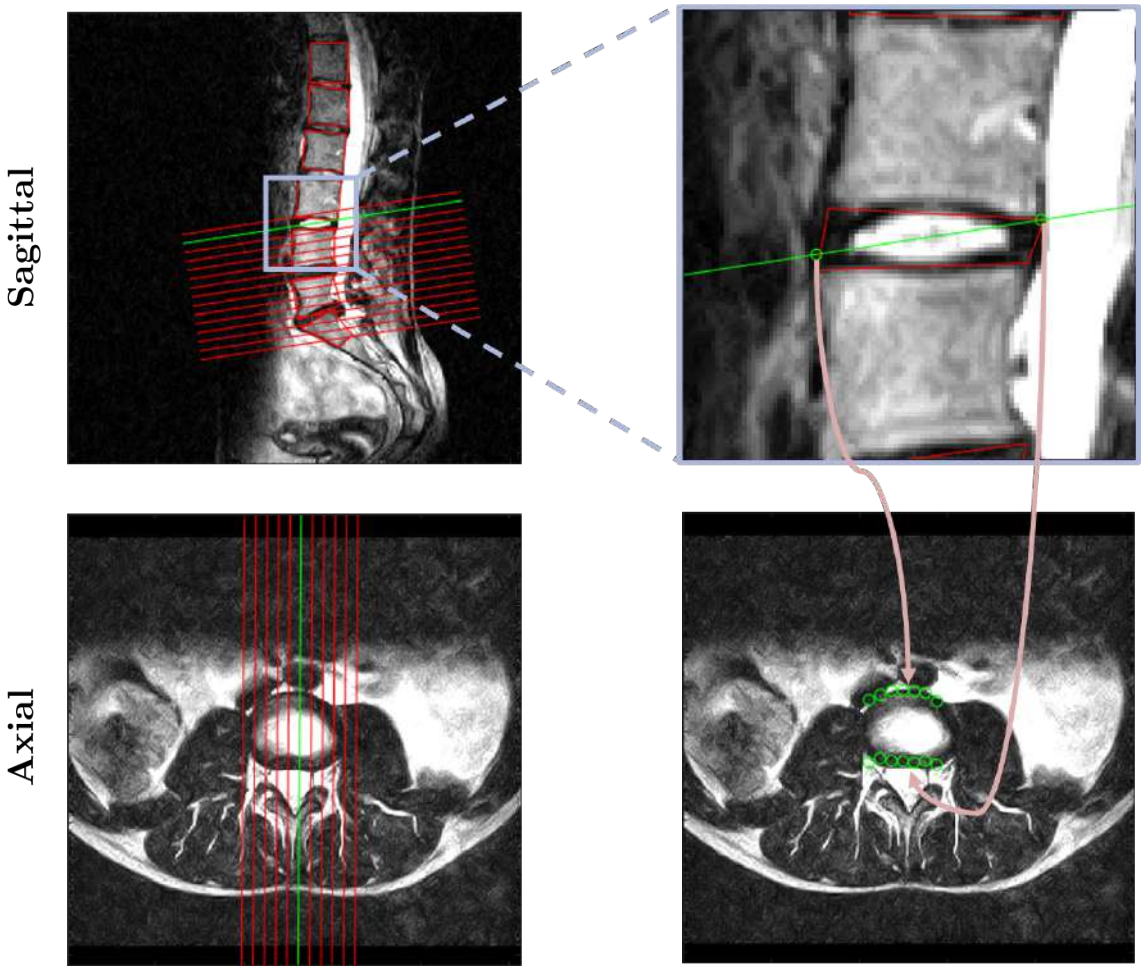


Figure 6.1: ***Ideal Axial Orientation.*** *Good alignment of axial slices to the disc.* ***Top left*** *shows the axial slices overlaid on top of the sagittal scan, the axial slice in green is the mid-axial slice with respect to the disc which can be seen in the* ***bottom left****.* ***Top right*** *shows a close-up of the disc and the intersections between the disc and the axial plane; left side is the anterior part of the disc and the right side is the posterior side of the disc.* ***Bottom right*** *shows the transferred detected points overlaid on top of the axial slice.*

This is however quite inconsistent as this is extremely dependant on the orientation of the axial slices. We find the best solution is to first project the posterior and anterior lines of the detected disc volumes where the posterior line is made up of the top and bottom anterior corners of the disc and similarly for the anterior line. Then, the intersection of the plane and the anterior/posterior lines are transferred to give us a projection of the detection in axial view, see Figure 6.2. We find detecting the axial bounding boxes through this method is extremely robust to partial volume effect of the axial scans.

Now that we have a rough estimate where a disc is in an axial scan, we can now define an axial-specific disc volume. To do that we center the disc both on the detected mid-sagittal point and the mid-axial slice. A volume is set to be only 3 slices per disc which roughly translates to the first slice containing the upper endplate of the disc, the middle slice shows the mid-axial view of the disc, and the last slice cointaining the lower endplate of the disc. The volumes are resized, while maintaining aspect ratio, to be the 224×224 in dimension per slice; we find this works well for us since each volume includes the disc and the central canal as shown in Figure 6.3. Similar to the sagittal disc volume, the range of the intensity inside the axial disc volume is set to be between 0 and 1. For each disc now we have two disc volumes, one sagittal and one axial; exploded views of the disc volumes can be seen in Figure 6.4.

6.1.2 CNN Architecture

The network architecture we use is a modified version of the network used in Section 4. Since the two input streams, axial and sagittal, are of different modalities, we combine after the **FC6** layer. The input dimension of the sagittal disc volume is $112 \times 224 \times 9$ like before and the input dimension of the axial disc volume is $224 \times 224 \times 3$ where 9 and 3 are the number of slices. We use **2D** kernels for the axial network and **3D** kernels for the sagittal network. We experimented with two different methods to

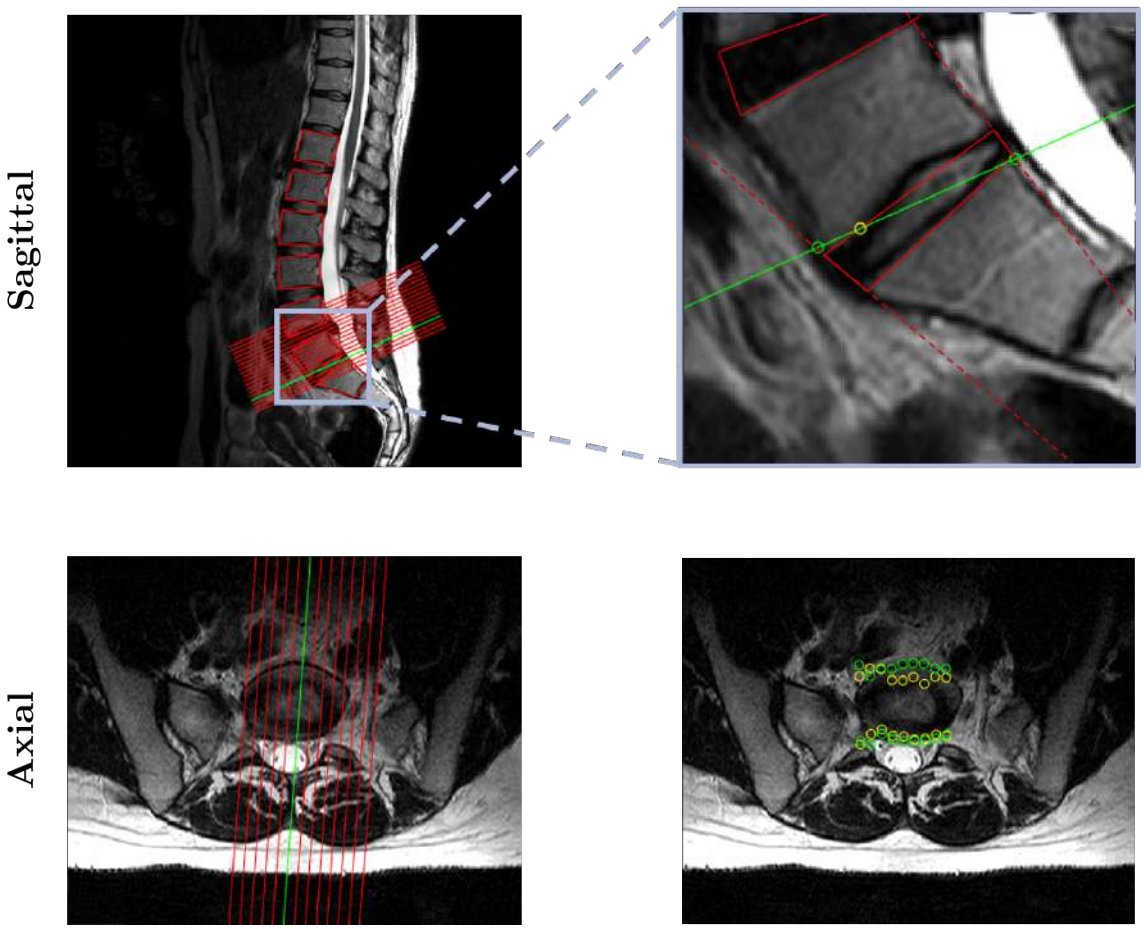


Figure 6.2: ***Top Left:*** *Mid-sagittal slice with axial slices overlaid on top, where the green line is the mid-axial slice.* ***Bottom Left:*** *Mid-axial slice with sagittal slices overlaid on top, where the green line is the mid-sagittal slice.* ***Top Right:*** *A sagittal close-up of the disc: (i) green circles mark the intersections between the anterior/posterior lines and the axial planes, while (ii) the yellow circle is the intersection of the plane with the detected disc.* ***Bottom Right:*** *Mid-axial slice with detected points overlaid on top. In this case, since the the axial slices are not completely parallel to the disc orientation, transferring the detection via the intersection with the anterior/posterior lines produces better results (green circles compared to yellow circles).*

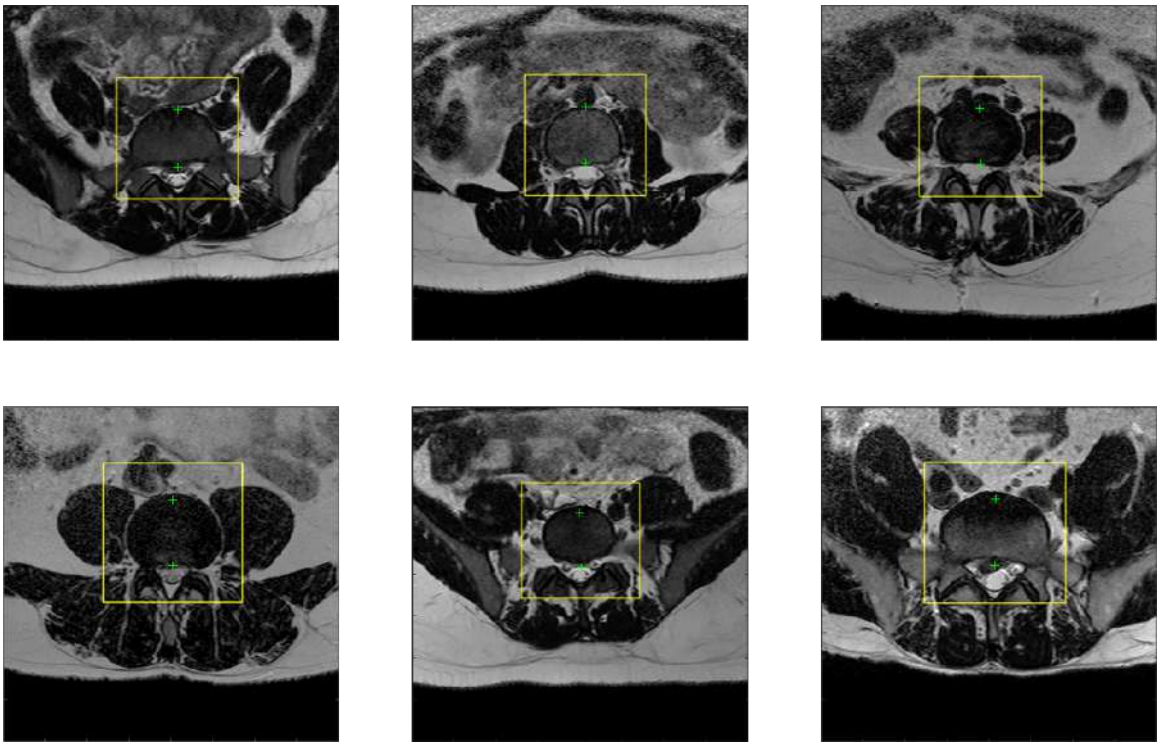


Figure 6.3: ***Example Bounding Boxes – Axial Scans.*** *The green markers are the intersections between anterior/posterior lines and the sagittal disc corner points. The bounding boxes, shown in yellow, are set to be twice the height and width of the anterior-to-posterior distance of each disc. The detected bounding boxes are quite robust to partial volume effect especially those caused by non-parallel axial scans.*

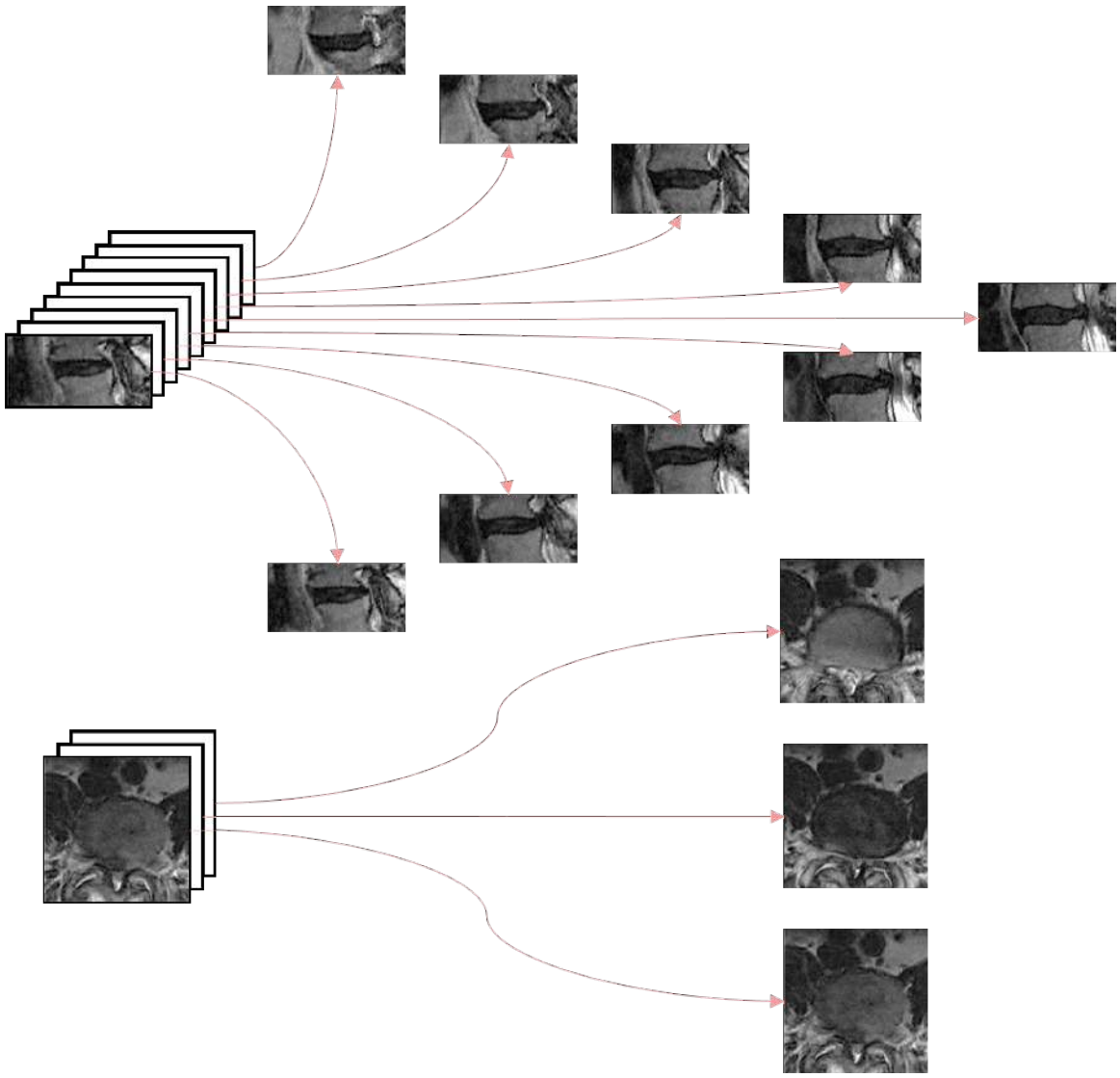


Figure 6.4: ***Disc Volume Example.*** *Shown are exploded views of extracted disc volumes, from axial and sagittal scans, where each image is an individual slice.* ***Top:*** *A sagittal disc volume.* ***Bottom.*** *An axial disc volume.*

combine after the **FC6** layer: (i) concatenation, and (ii) summation. Each of the three gradings (**anterior disc bulging**, **posterior disc bulging**, and **disc herniation**) are predicted via their very own unique **FC7**, and **FC8** layers. We use the same multi-task and class balanced losses in Section 4. Figure 6.5 shows the configuration of the network. We also experimented with predicting the gradings from only one of the streams.

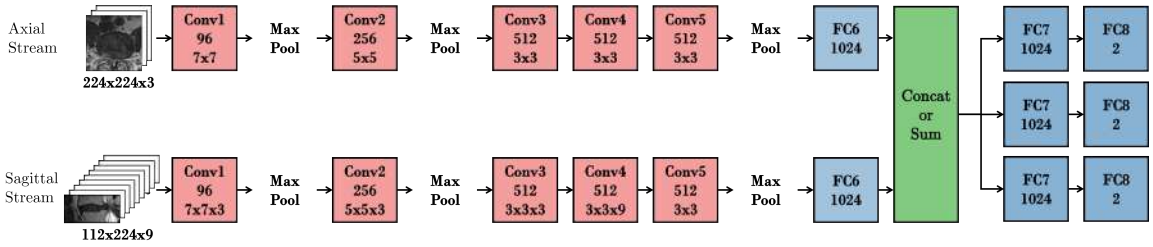


Figure 6.5: ***Multi-stream Network – Axial & Sagittal.*** *Multi-stream network that accepts both the sagittal and axial disc volumes as separate inputs. The network either concatenate or sum the* ***FC6*** *layers of the two different streams and proceeds to predict the three different grading via three separate* ***FC7*** *and* ***FC8*** *layers. For the sagittal stream, we use 3D kernels for* ***Conv1*** *to* ***Conv4*** *e.g. the kernel for* ***Conv1*** $7 \times 7 \times 3$ *instead of* 7×7.

6.1.3 Training

We train the CNN via stochastic gradient descent with momentum from scratch without any pre-training. The hyperparameters are: mini-batch size 128; momentum 0.9; weight decay 0.0005; initial learning rate 0.001, which is lowered by a factor of 10 as the error plateaus. We opted for two independent augmentation methods, one for each input stream. The training augmentation of the sagittal stream is identical to that mentioned in Chapter 4 while the augmentation for the axial stream is as follows:

1. Rotation with $\theta = -15°$ to $15°$
2. Translation of ± 24 pixels in the x-axis, and ± 24 pixels in the y-axis
3. Rescaling with a scaling factor between 90% to 110%

4. Intensity variation between -0.1 to 0.1

5. Randomly drop out one input stream

No test-time augmentation was used.

6.1.4 Results

The **Genodisc** dataset is split into a 80:10:10 train:validation:test sets on a per patient basis (not per disc). We only use discs that have both axial and sagittal scans. This results in a total of 6,883 discs: 5,443 discs in train, 692 in validation, and 648 in the test set. As in Chapter 4, we compare the average per-class accuracy against the average per-class intra-rater agreement of the radiologist that read the **Genodisc** dataset.

We experimented with several methods to merge the input streams and compare them to just learning the streams independently, the results of which can be seen in Table 6.1. Surprisingly, learning only with sagittal inputs, **Sag** in Table 6.1, is sufficient enough to achieve performance nearing the radiologist. This is however only on binary tasks of predicting the presence or absence of **anterior disc bulging**, **posterior disc bulging**, and **disc herniation**. Another interesting observation is that the model trained only on axials, **Axi** in Table 6.1, achieves significantly lower performance compared to **Sag**. This might be due to the fact that the axial scans in **Genodisc** were not in any way standardized; some appear parallel to the disc orientation and some do not. In comparison, sagittal scans do not experience this issue. Nevertheless, we tried naively averaging the direct predictions of both **Sag** and **Axi**, **Sag + Axi (Avg)** in Table 6.1, by average pooling the softmax after both **FC8** layers. It appears, doing this we consistently get lower performance across all the different tasks, compared to **Sag**. Only by learning both inputs simultaneously, we see an improvement in performance on **disc herniation**: **87.1%** $\rightarrow$ **87.3%** for

Sag + Axi (Cat), **87.1% → 88.3%** for for **Sag + Axi (Sum)**. We also see a slight jump for **anterior disc bulging**: **88.8% → 88.9%** for for **Sag + Axi (Sum)**. On average, we get a 0.3% jump in performance when predicting from both axial and sagittal disc volumes.

Tasks	Intra-rater	Models				
		Sag	Axi	Sag + Axi (Avg)	Sag + Axi (Cat)	Sag + Axi (Sum)
ADB	85.6	88.8 ± 1.7	78.5 ± 2.4	88.2 ± 2.6	88.6 ± 1.8	$\mathbf{88.9 \pm 1.6}$
PDB	82.6	$\mathbf{84.7 \pm 1.3}$	76.1 ± 3.0	83.2 ± 2.1	83.9 ± 1.1	83.9 ± 1.7
DH	91.2	87.1 ± 3.5	79.8 ± 0.4	86.7 ± 0.9	87.7 ± 1.2	$\mathbf{88.3 \pm 0.9}$
Average	86.5	86.8 ± 2.1	78.1 ± 1.9	86.0 ± 1.9	86.7 ± 1.4	$\mathbf{87.1 \pm 1.4}$

Table 6.1: ***Axial + Sagittal Results.*** *The performance (%) of various models; two models are trained each with swapped validation and test sets. "**ADB**" is **anterior disc bulging**, "**PDB**" is **posterior disc bulging**, and "**DH**" is **disc herniation**. "**Intra-rater**" is the intra-rater agreement. "**Sag**" shows the results from training a model just with sagittal scans while "**Axi**" shows the results from an axial only model. "**Sag + Axi**" are models that uses both the axial and sagittal input streams for classification. "**Sag + Axi (Avg)**" refers to directly averaging the predictions of the "**Sag**" and "**Axi**" models. "**Sag + Axi (Cat)**" is the model with concatenated **FC6** layers while "**Sag + Axi (Sum)**" sums the two **FC6** layers.*

We can also produce the associated hotspots for both the two input streams, axial and sagittal, for all the radiological gradings: **anterior disc bulging** in Figure 6.6, **posterior disc bulging** in Figure 6.7, and **disc herniation** in Figure 6.8.

6.2 Adding T1-weighted Scans

So far, we have only focused on T2-weighted scans as they are the most common sequence used for lumbar scans. However, the **Genodisc** dataset does not only contain T2-weighted scans but also T1-weighted scans. In fact, 98.6% of the subjects in the **Genodisc** dataset possess both scans. A T1-weighted scan often possesses the same field-of-view as its T2-weighted counterpart as can be seen in Figure 6.9. They also

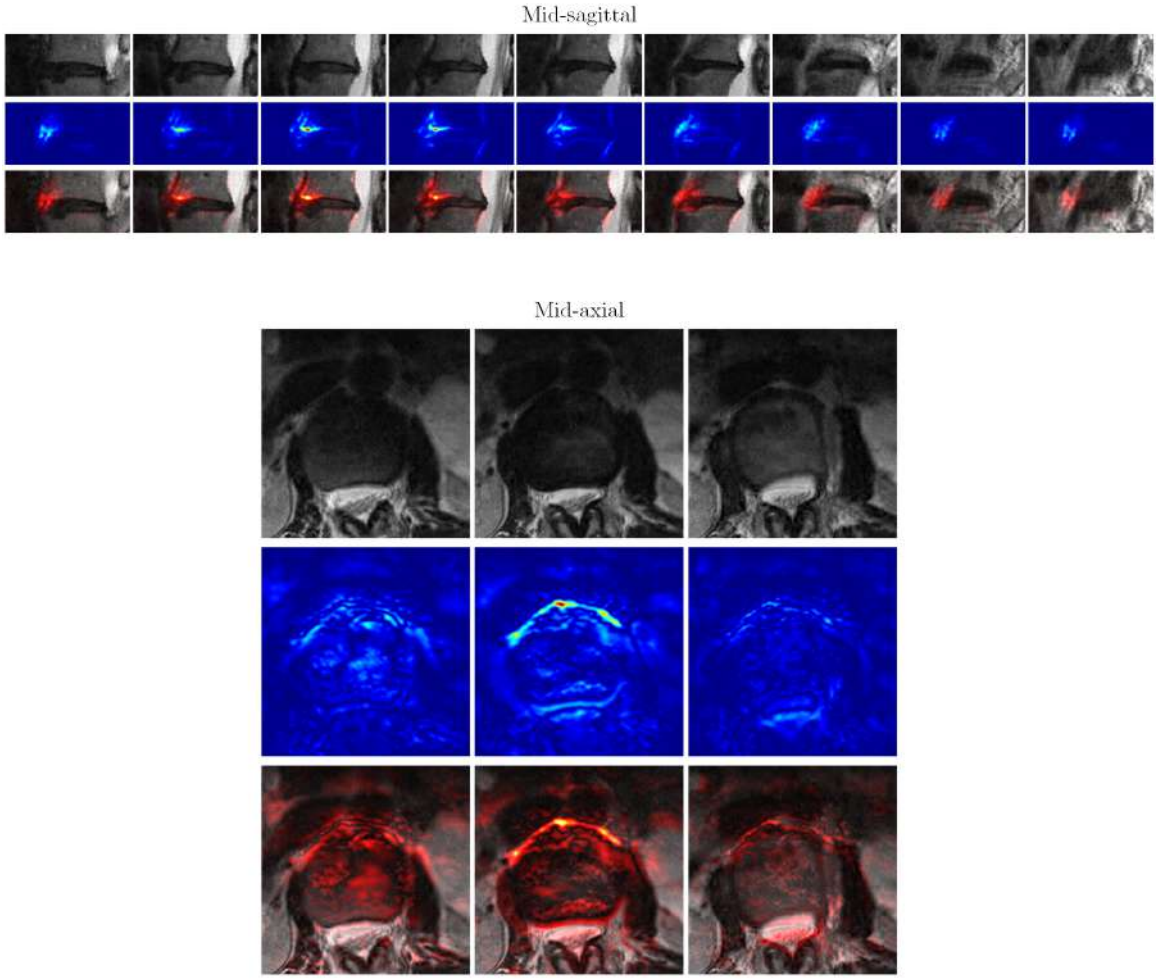


Figure 6.6: ***Hotspots – Anterior Disc Bulging.*** *Example of* ***anterior disc bulging*** *in the test set. The* ***top*** *images are the sagittal disc while the* ***bottom*** *images are from the axial volume. The hotspots highlight the anterior part of the disc i.e. the left side of the sagittal images and the top part of the axial images.*

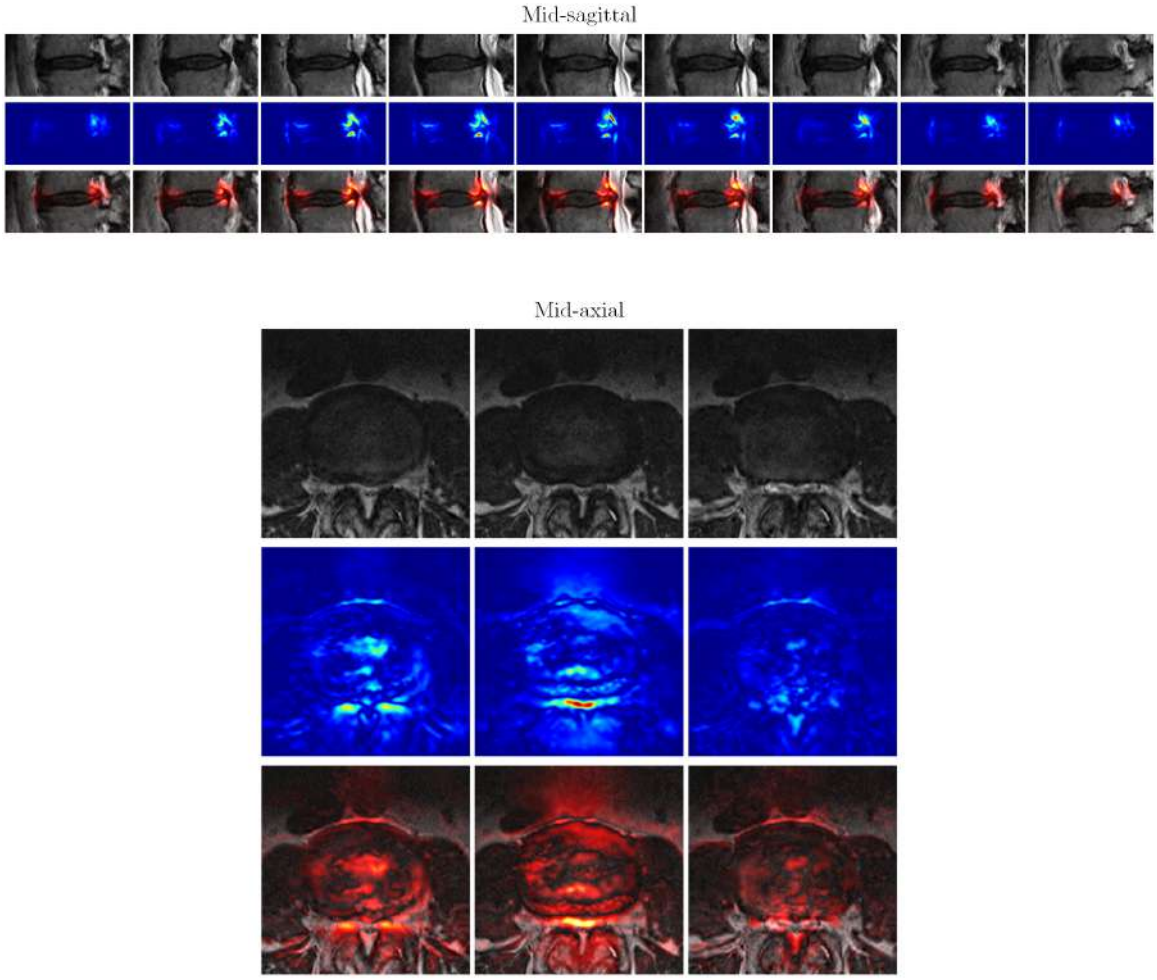


Figure 6.7: ***Hotspots – Posterior Disc Bulging.*** *Example of* ***posterior disc bulging*** *in the test set. The* ***top*** *images are the sagittal disc while the* ***bottom*** *images are from the axial volume. The hotspots highlight the posterior part of the disc i.e. the right side of the sagittal images and the bottom part of the axial images.*

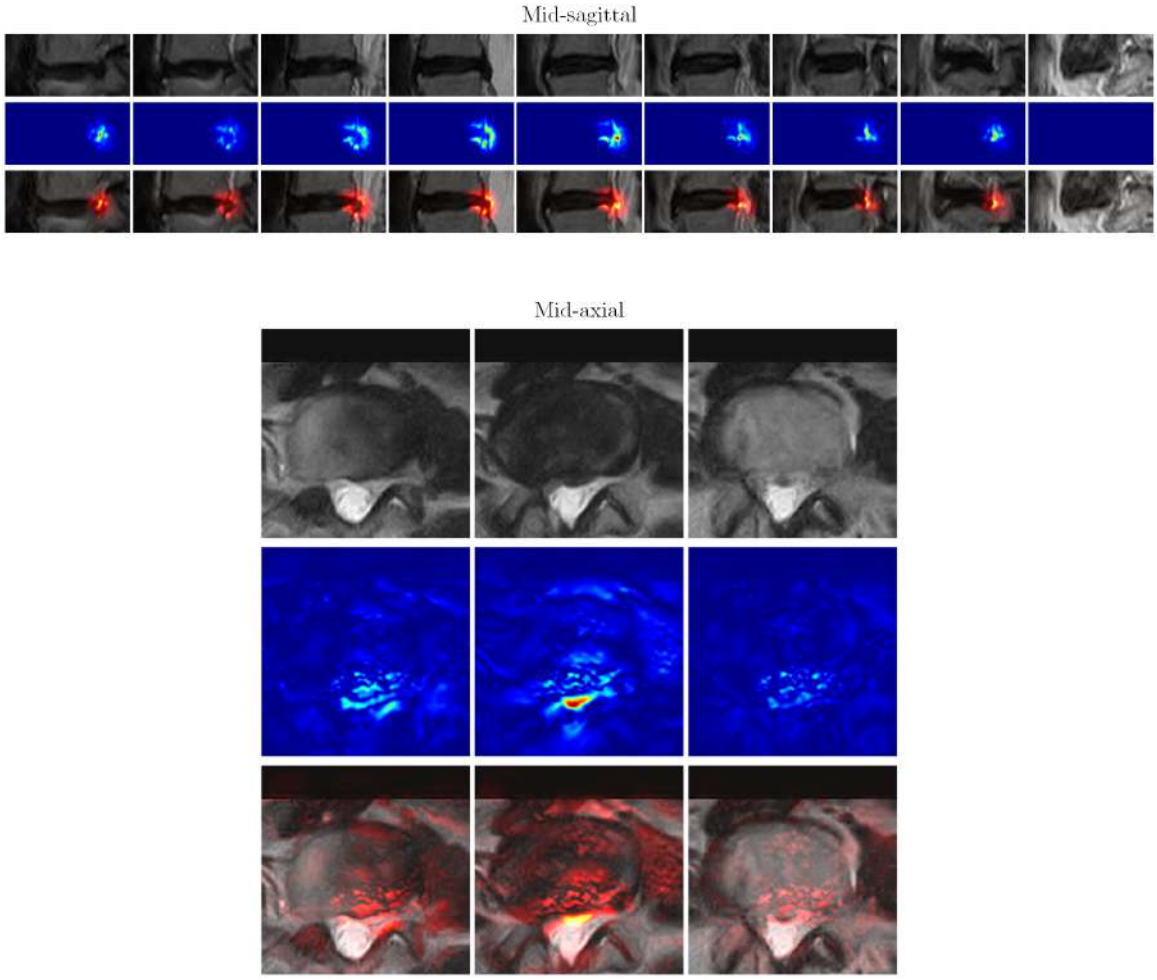


Figure 6.8: ***Hotspots – Disc Herniation.*** *Example of* ***disc herniation*** *in the test set. The* ***top*** *images are the sagittal disc while the* ***bottom*** *images are from the axial volume. The hotspots highlight the herniated region in the disc both in sagittal (right side of the images) and axial views (bottom side of the images).*

tend to be similar in resolution to the T2-weighted scans. As these scans are normally conducted in a single session, there is on average only a slight shift in positions of the same subject in the two sequences.

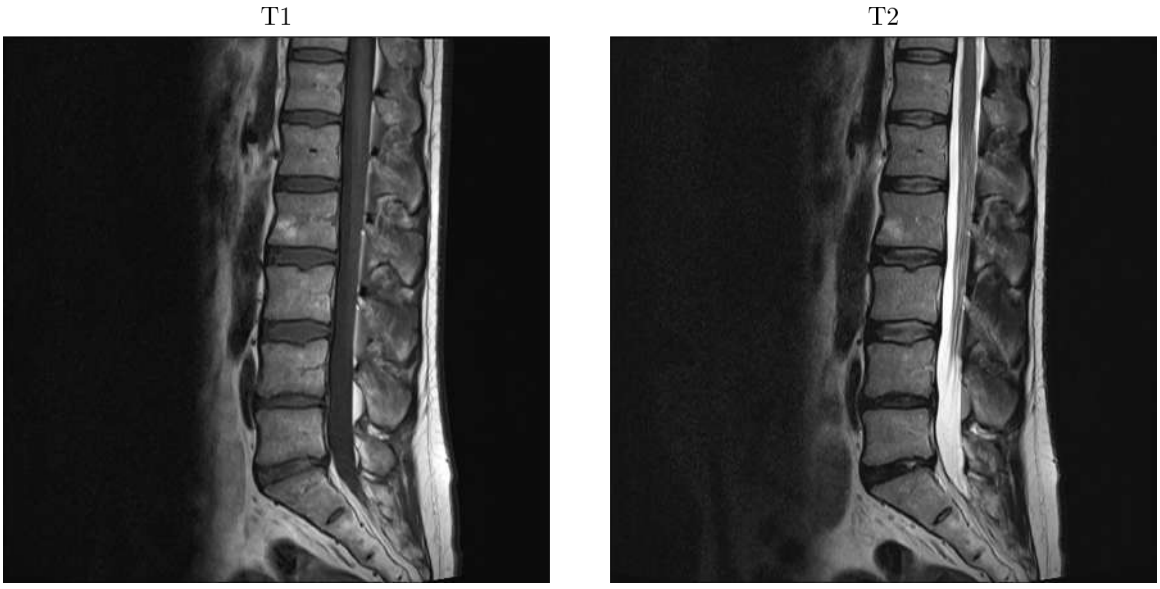


Figure 6.9: ***Same Spine, Different Sequences. Left:*** *T1-weighted sagittal scan.* ***Right:*** *T2-weighted sagittal scan. T2-weighted scans are normally used for gradings of disc degeneration as the disc more muted in T1-weighted scans. However, clearer distinction of some features like the dural sac indicate that T1-weighted scans might be useful for other degenerative measures of the spine.*

Why Add T1-weighted Scans? Cheung et al. (2014) found that in terms of radiological gradings, T1-weighted scans is as easily readable as T2-weighted scans which is corroborated by the findings of Moffit et al. (1988) with one more exception in that they found that certain features e.g. the dural sac can be seen more clearly in T1-weighted scans and vice versa for other features in T2-weighted scans. This suggests that having both sequences would be beneficial in some gradings. One other clear example are marrow changes of the intervertebral discs, which in the original definition by Modic et al. (1988) require both sequences. Modic et al. (1988) also noted that there are cases of discs with marrow changes that may appear abnormal in one sequence, hyper/hypo-intense regions, but remain normal in the other se-

quence. This suggests that marrow changes are better classified with both sequences. We use the same multi-stream training to predict radiological gradings, the same gradings in Chapter 4, namely: **Pfirrmann grading**, **disc narrowing**, **upper endplate defects**, **lower endplate defects**, **upper marrow changes**, **lower marrow changes**, **spondylolisthesis**, and **central canal stenosis**.

6.2.1 Disc Volume Extraction – Sagittal T1

We use the same disc volume extraction method, used on T2-weighted scans described in Section 4.1.1, on the the T1-weighted scans. Examples of the extracted disc volumes can be seen in Figure 6.10.

6.2.2 CNN Architecture

There are two input streams, one each for the T1-weighted and T2-weighted disc volumes. As with the multi-stream axial and sagittal network, we experimented with two different methods to combine after the **FC6** layer can be seen in Figure 6.11: (i) concatenation, and (ii) summation. The input dimension of both disc volumes is $112 \times 224 \times 9$ like before. We share the weights first six layers: **Conv1** to **FC6**. Since the inputs, T1 and T2-weighted discs, are almost homogeneous in appearance, we also experiment with fully sharing the weights of the stream and summing the **FC8** outputs. For this network we branch out from **Conv5** similar to the best network in Chapter 4. We use the same multi-task and class balanced losses and the training regime described in Chapter 4.

6.2.3 Results

The **Genodisc** dataset is split into a 80:10:10 train:validation:test sets on a per patient basis (not per disc). We only use discs that have both T1 and T2 sequences.

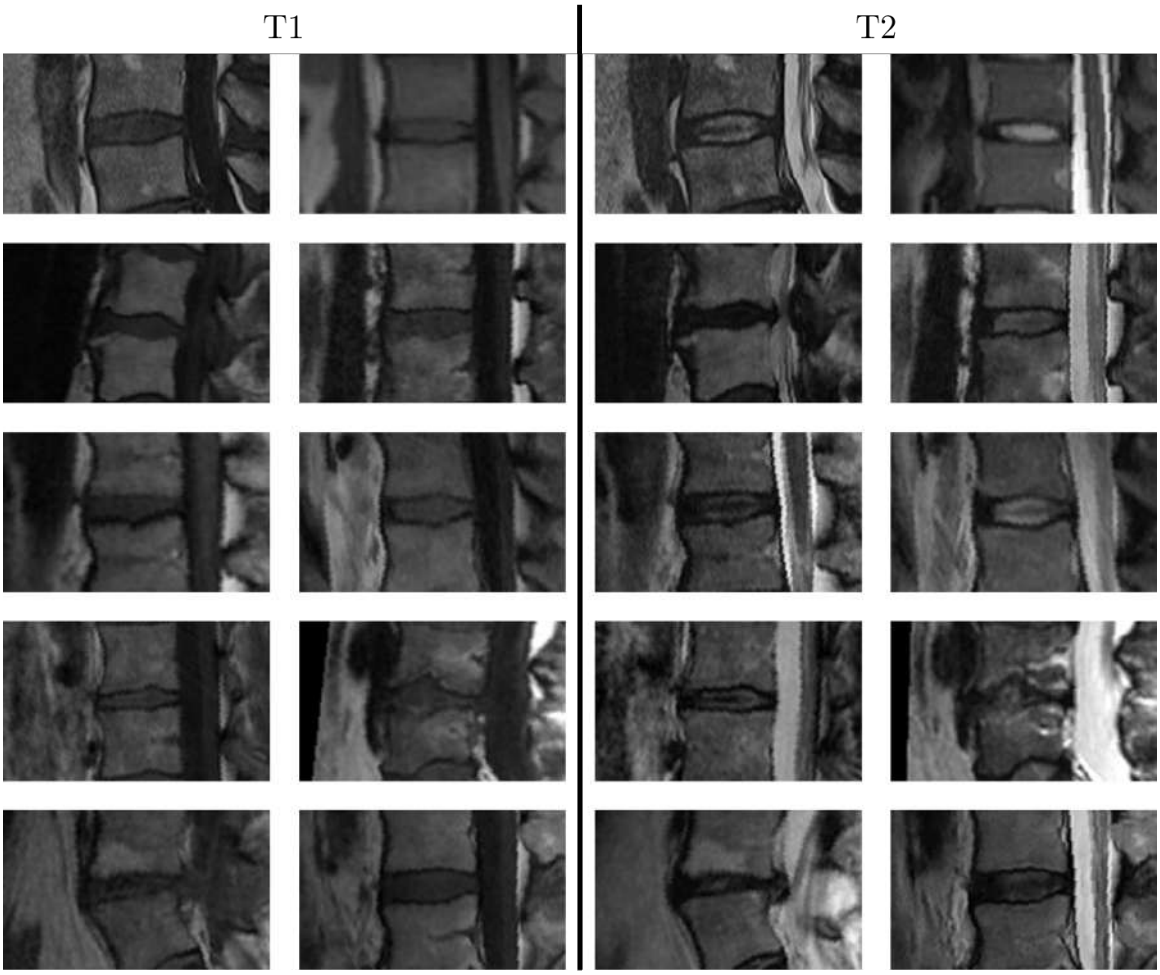


Figure 6.10: ***Same Disc, Different Sequences. Left:*** *Examples of disc volumes extracted from T1-weighted images.* ***Right:*** *T2-weighted counterparts of the same discs.*

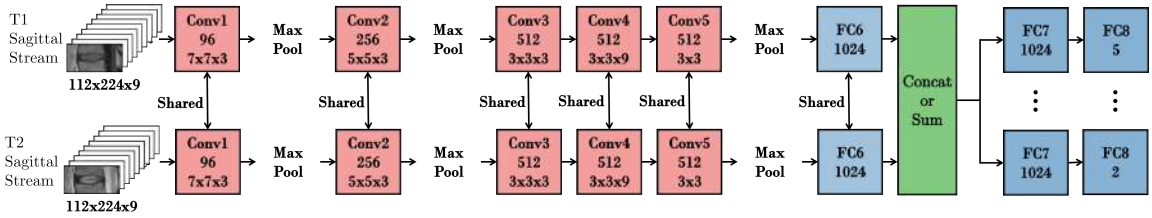


Figure 6.11: ***Multi-stream Network – T1 & T2.*** *Multi-stream network that accepts both the T1-weighted and T2-weighted disc volumes as separate inputs. The network either concatenate or sum the* ***FC6*** *layers of the two different streams and proceeds to predict the eight different grading via eight separate* ***FC7*** *and* ***FC8*** *layers. We use 3D kernels for* ***Conv1*** *to* ***Conv4****.*

This results in a total of 11,946 discs: 9,525 discs in train, 1,198 in validation, and 1,223 in the test set. As in Chapter 4, we compare the average per-class accuracy against the average per-class intra-rater agreement of the radiologist that read the **Genodisc** dataset. As with the axial experiments, we look at several methods to merge the input streams and compare them to the de facto method of learning from just the T2-weighted stream, the results of which can be seen in Table 6.2.

Tasks	Intra-rater	Models			
		T2 Only –	T1 + T2 FC6 Concat	T1 + T2 FC6 Sum	T1 + T2 FC8 Sum
Pf	70.4	**71.5 ± 1.0**	67.8 ± 1.1	69.3 ± 0.0	68.8 ± 3.2
DN	72.0	**75.0 ± 2.3**	73.4 ± 2.4	72.8 ± 1.8	73.5 ± 3.0
UED	80.7	85.2 ± 2.1	84.3 ± 0.6	**87.6 ± 2.0**	85.8 ± 2.0
LED	83.3	87.5 ± 0.4	85.7 ± 0.8	87.8 ± 1.7	**89.1 ± 2.2**
UMC	92.5	91.0 ± 1.3	91.0 ± 3.3	91.2 ± 1.6	**91.6 ± 2.0**
LMC	91.4	**90.3 ± 2.1**	89.0 ± 1.4	90.0 ± 1.6	90.2 ± 2.9
Spon	89.6	95.2 ± 0.0	95.1 ± 2.2	96.9 ± 1.0	**97.6 ± 0.2**
CCS	79.7	94.3 ± 0.3	93.1 ± 0.6	93.0 ± 1.2	**95.4 ± 0.5**
Average	86.5	86.3 ± 0.3	84.9 ± 0.7	86.1 ± 0.5	**86.5 ± 0.1**

Table 6.2: ***T1 + T2 Results.*** *The performance (%) of various models.* ***Pf*** *=* ***Pfirrmann grading****,* ***DN*** *=* ***disc narrowing****,* ***UED*** *=* ***upper endplate defects****,* ***LED*** *=* ***lower endplate defects****,* ***UMC*** *=* ***upper marrow changes****,* ***LMC*** *=* ***lower marrow changes****,* ***Spon*** *=* ***spondylolisthesis****, and* ***CCS*** *=* ***central canal stenosis****. "****T2 Only****" is the "****3D****" model in Table 4.3 with a branch at* ***Conv5****. The only other network which branches out after* ***Conv5*** *for the multi-tasking is the "****FC8 Sum****" network. Merging streams with summing the intermediate FC layers,* ***FC6*** *and* ***FC8*** *in our case seems to work best.*

Surprisingly, learning only with concatenated intermediate layer, **FC6** in our case, works slightly worse than the standard method of just learning with T2-weighted disc volumes. Another interesting observation is that the models with merged input streams all suffer from a decrement in performance for **Pfirrmann grading** classification: **71.5%** → **67.8%**, **71.5%** → **69.3%**, and **71.5%** → **68.8%**. This is not surprising since Pfirrmann gradings are solely classified through the T2-weighted scans. T1-weighted scans also typically show discs in a slightly muted fashion, which might hinder the classification task since we share all the convolutional models i.e.

independent weights might work better at least for **Pfirrmann grading**. In comparison however, we see improvements in performance for a majority of the radiological gradings, when comparing **T2 Only** to **FC8 Sum** in Table 6.2: **upper endplate defects 85.2% → 85.8%**, **lower endplate defects 87.5% → 89.1%**, **upper marrow changes 91.0% → 91.6%**, **spondylolisthesis 95.2% → 97.6%**, and **central canal stenosis 94.3% → 95.4%**. On average, we get a 0.2% jump in performance when predicting from both T1 and T2-weighted scans; more if we ignore **Pfirrmann grading** which is a T2-only dependant grading, with a 0.7% boost.

Comparison to Other Methods. Abbati et al. (2017) proposed two methods on predicting **lumbar spinal stenosis**: (i) based on quantitative and qualitative features marked up by radiologists, and (ii) directly classifying from radiological images. We compare our results in predicting the **Genodisc** definition of **central canal stenosis** against their results. They achieved an AUC of 85.4%±3.3% from classifiers trained on the quantitative and qualitative features (human annotated features) and an AUC of $70.6\% \pm 6.7\%$ when classifying from the MRIs while we achieve an AUC of $98.8\% \pm 0.2\%$.

6.3 Direct Classification From Raw Volume

All of the classifications of radiological gradings discussed so far relied on the extremely tight detections of intervertebral disc bounding volumes. This is true for any sequence, T1-weighted or T2-weighted, and any plane of view, sagittal or axial. Here, in this section we briefly discuss the possibility of directly classifying multiple disc gradings of a scan via raw unprocessed MR volume. We resize all the volumes to be $512 \times 512 \times 9$ in dimension with no slice re-sampling in the z-dimension.

6.3.1 CNN Architecture

We use the same architecture used in Chapter 4, specifically the 3D version of the architecture in Figure 4.3. We adapt a curriculum learning approach by Bengio et al. (2009): first we learn to predict radiological gradings from the disc volumes and we then reuse this network for the actual task of predicting disc grading from the raw volume. Training from scratch consistently results in slightly lower performance in our case. Now instead of learning gradings on a per disc basis we now need to predict the grading of all the six lumbar discs simultaneously from the input data i.e. raw MRI volume. We use the same multi-task and class balanced losses and the identical training regime described in Chapter 4 but since each loss of the disc is now unique we have six times the number of losses to minimize (a new loss for each disc: T12-L1 to L5-S1), 48 in total.

6.3.2 Results

The **Genodisc** dataset is split into a 80:10:10 train:validation:test with 2009 subjects, where each subject possesses one input MRI volume. Results can be seen in Table 6.3.

The performance of the model that ingests raw MR volumes for prediction is slightly worse for all the eight different gradings when compared to a model trained on disc volumes. On average, we lose 12.4% in average per-class accuracy for all the gradings. Similar observations can also be made when we look at evidence hotspots for example the ones in Figure 6.12. Overall, probably due to the relatively low amount of data, directly predicting radiological gradings of intervertebral discs from raw MRI without any pre-processing is still not up to par with predicting the gradings from extracted disc volumes.

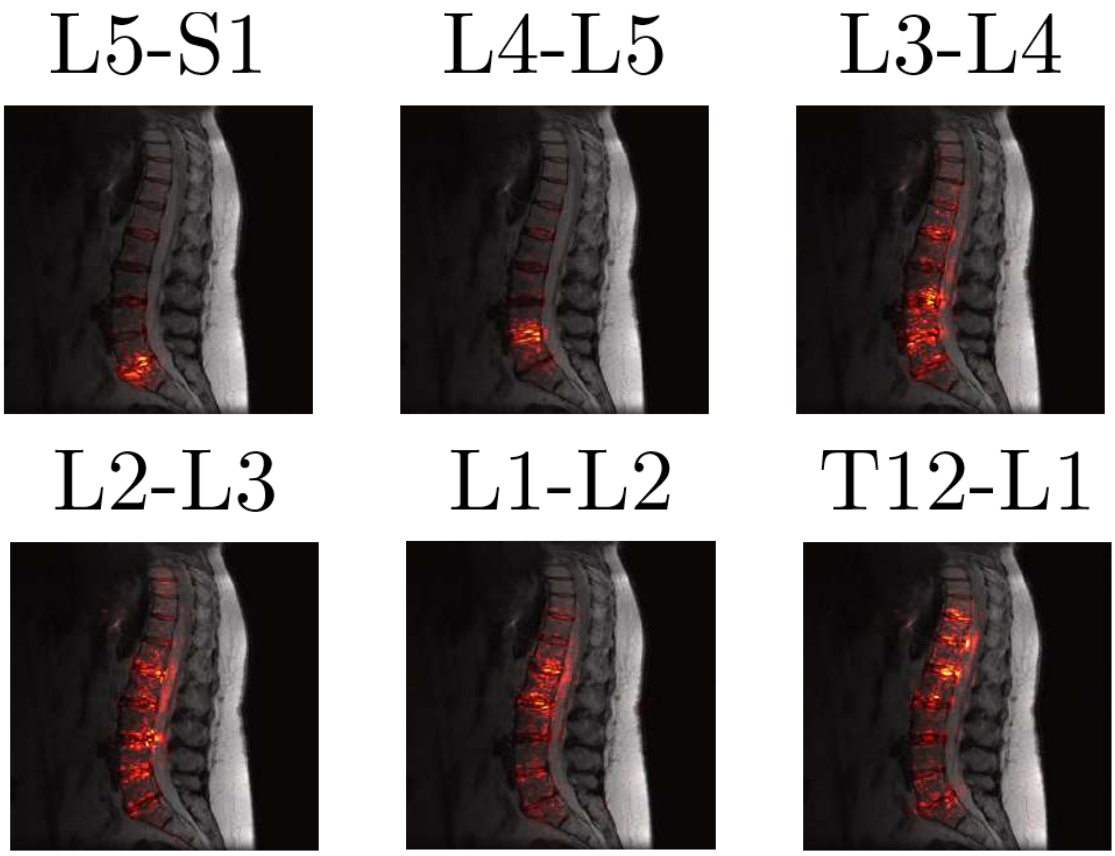


Figure 6.12: ***Hotspots – Pfirrmann Grading From Raw MRI. Pfirrmann grading*** *hotspots for each the six discs. Since each disc is now its own predictive task, we can essentially backpropagate from each disc prediction. In this case, we produce* ***Pfirrmann grading*** *hotspots for each of the six discs: L5-S1, L4-L5, L3-L4, L2-L3, L1-L2, and T12-L1. The first two hotspots strongly correlate with two lowest discs but the other four do no show specific localization of the discs. However, the network still focuses on the discs of the whole MRI which suggests that only the discs are important, at least for* ***Pfirrman grading*** *classification.*

Tasks	Intra-rater	Models	
		IVD	Raw MRI
Pfirrmann	70.4	**71.5 ± 1.0**	57.1
Disc Narrowing	72.0	**75.0 ± 2.3**	59.2
Upper Endplate Defects	80.7	**85.2 ± 2.1**	77.1
Lower Endplate Defects	83.3	**87.5 ± 0.4**	75.3
Upper Marrow Changes	92.5	**91.0 ± 1.3**	82.7
Lower Marrow Changes	91.4	**90.3 ± 2.1**	75.0
Spondylolisthesis	89.6	**95.2 ± 0.0**	80.1
Central Canal Stenosis	79.7	**94.3 ± 0.3**	84.8
Average	86.5	**86.3 ± 0.3**	73.9

Table 6.3: ***Disc Volume vs Raw MRI.*** *The performance (%) of two models; "**IVD Only**" is the "**3D**" model in Table 4.3 while "**Raw Volume**" is the model trained on raw MRIs.*

6.4 Summary

In this chapter, we have shown that adding axial and T1-weighted scans help improve some radiological gradings, particularly **anterior disc bulging** and **disc herniation** in the case of axial, and most of the gradings in the case of T1-weighted sagittal scans. Unsurprisingly, we show that training from raw MR volumes is not the best course of action and that focusing of predicting from disc volumes is best, at least in our case. Finally, we also show that evidence hotspots can also be produced from networks trained on: (i) axial scans, and (ii) full sagittal MR scans. So far, we have focused on radiological gradings available in the **Genodisc** dataset to train our models but these gradings might not be as readily available in other datasets. Thus, in Chapter 7 we look at a novel way to gain performance using only the data and essentially free labels.

Chapter 7

Self-Supervision

Methods in the previous chapters have relied upon on the relatively well-defined **Genodisc** dataset (Chapter 3), with each subject possessing only one scan and for each scan a radiologist had graded or annotated, in a very consistent manner several gradings for each of the six lumbar discs. This is of course very different in the clinic. A significant proportion of patients scanned in clinical settings have follow-up scans, not just one, and often possess incomplete and non-standardized radiological reports. In this chapter we show that such cases with longitudinal scans alone can be used as a form of "free" self-supervision for training a deep network. We demonstrate this self-supervised learning for the case of T2-weighted sagittal lumbar MRIs. We train a Siamese CNN to distinguish between pairs of images that contain the same patient scanned at different points in time, and pairs of images of entirely different patients. We also illustrate that additional data-dependent self-supervision tasks can be included by specifying an auxiliary task of predicting vertebral body levels, and including both types of self-supervision in a multi-task training scheme. The performance of this pre-trained network is then assessed on a grading classification task on the **TwinsUK** dataset i.e. with the end goal of learning the disc degeneration radiological gradings attached to the intervertebral discs. Overall in this chapter,

we experimented on two different datasets: (i) **TwinsUK**, and (ii) **OSCLMRIC**. We also compare to pre-training a CNN on **Genodisc** to predict eight radiological gradings, the same one trained in Chapter 4.

This chapter can be broken down into three main sections; the first one being the discussion of self-supervision in Section 7.1, followed by details of the input in Section 7.2, details of the losses and the architecture used for training in Section 7.3, and finally the experiments and results in Section 7.4.

7.1 Why Self-Supervision?

A prerequisite for the utilization of machine learning methods in medical image understanding problems is the collection of suitably curated and annotated clinical datasets for training and testing. Due to the expense of collecting large medical datasets with the associated ground-truth, it is important to develop new techniques to maximise the use of available data and minimize the effort required to collect new cases. In this chapter, we propose a self-supervision approach that can be used to pre-train a CNN using the embedded information that is readily available with standard clinical image data. Many patients are scanned multiple times in a so-called longitudinal manner, for instance to assess changes in disease state or to monitor therapy. We define a pre-training scheme using only the information about which scans belong to the same patient (see Figure 7.1). Note, we do not need to know the identity of the patient; only which images belong to the same patient. This information is readily available in formats such as DICOM (Digital Imaging and Communications in Medicine) that typically include a rich set of meta-data such as patient identity, date of birth and imaging protocol (and DICOM anonymization software typically assigns the same 'fake-id' to images of the same patient).

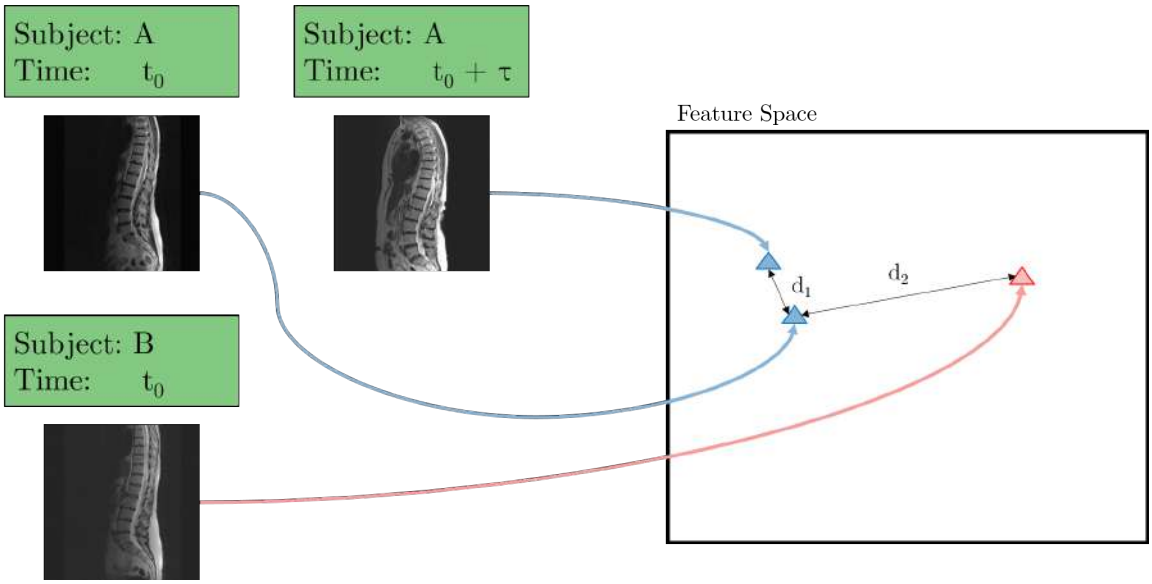


Figure 7.1: ***Free to Learn.*** *For each pair of subject or subjects, we learn the appropriate discriminative features so that the distances of the MRI in the feature space are correct. In this case we learn that the same subjects at different time points, marked as cyan triangles in the feature space, are close and such the distance, d_1, between them must be small and correspondingly the distance, d_2, between different subjects, red and cyan triangles in the feature space, must be large. Ideally, the learned features would learn that $d_1 < d_2$. This problem can also be posed as a triplet but we opted for the easier pairwise learning approach. Note that, instead of directly learning distances between MRIs, we choose to learn distance between vertebrae and discs.*

7.2 Input Volumes

Instead of directly learning individuals using the whole MRI, we instead learn individual bodies in the spine i.e. vertebrae and the discs. This section first describes the details of the input volumes extraction: the vertebral bodies (VBs) that will be used for the self-supervised training; and the intervertebral discs (IVDs) that will also be used for the self-supervised training and the supervised classification experiments. We then describe the loss functions and network architecture.

7.2.1 Extracting Vertebral Bodies and Intervertebral Discs

For each T2-weighted sagittal MRI we automatically detect bounding volumes of the (T12 to S1) VBs alongside the level labels using the pipeline outlined in Chapter 3 and the same disc volume extraction method in Chapter 4 to extract both VB and IVD volumes. We extract the corresponding IVD volumes (T12-L1 to L5-S1, where T, L, and S refer to the thoracic, lumbar, and sacral vertebrae) from the pairs of vertebrae e.g. a L5-S1 IVD is the disc between the L5 and S1 vertebrae. Figure 7.2 shows the input and outputs of the extraction pipeline.

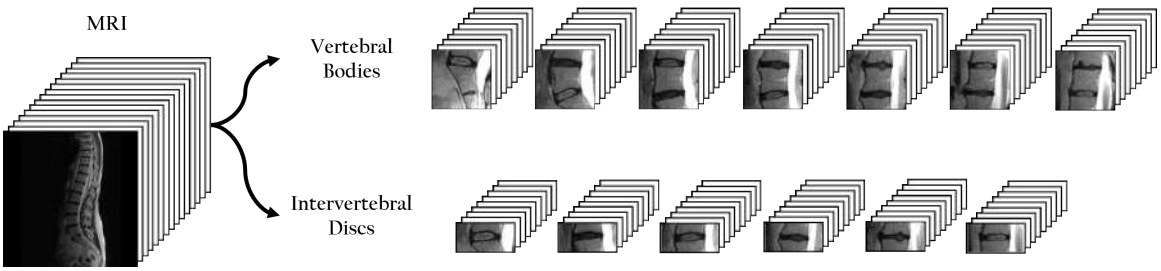


Figure 7.2: ***Extracting VB and IVD Volumes.*** *For each MRI, we extract 7 VB (T12 to S1) and 6 IVD (T12-L1 to L5-S1) volumes. The dimensions of the volumes are: (i)* $224 \times 224 \times 9$ *for the VBs, and (ii)* $112 \times 224 \times 9$ *for the IVDs. The whole volume is centred at the detected middle slice of the volume of the VB/IVD.*

The slices of both the VB and IVD volumes are mid-sagittally aligned (to prevent misalignment that can occur from scoliosis and other disorders) and zero-padded slice-wise if the number of slices is below the predefined 9 channels. The volumes

are rescaled according to the width of the VB or IVD and normalized such that the median intensity of the VB above and below the current VB or IVD is 0.5.

7.3 Loss Functions & CNN Architectures

In this section we first describe the loses that are used to train the network, followed by the CNN architectures that we use. Training and implementation details are then given in Section 7.3.5.

7.3.1 Loss Functions

The losses to train the CNN are made up of two distinct losses: (i) a contrastive loss on whether an input pair is of the same person (i.e. longitudinal) or not, and (ii) a classification loss on predicting the level of vertebral bodies.

7.3.2 Self-Supervision via Contrastive Loss

The longitudinal information of the scans is used to train a Siamese network such that the embeddings for scans of the same person are close, whereas scans of different people are not. The input is a pair of VBs or IVDs of the same level; an S1 VB is only compared against an S1 VB and vice versa. We use the contrastive loss defined by Chopra et al. (2005):

$$\mathcal{L}_C = \sum_{n=1}^{N} (y) d_n^2 + (1-y) \max(0, m - d_n)^2 \tag{7.1}$$

where

$$d_n = \|a_n - b_n\|_2 \tag{7.2}$$

and a_n and b_n are the 1024-dimensional **FC7** (embedding) vectors for the first and second VB/IVD in an input pair, and m is a predefined margin. Positive, $y =$

1, VB/IVD pairs are those that were obtained from a single unique subject (same VB/IVD scanned at different points in time) and negative, $y = 0$, pairs are VBs from different individuals (see Fig. 7.3). The dimension of each VB is $224 \times 224 \times 9$ while the dimension of each IVD is $112 \times 224 \times 9$.

We compare two different ways to train the Siamese network: (i) training only on VB pairs, and (ii) training on VB pairs and IVD pairs simultaneously. We start with the VBs instead of the IVDs due to the fact that vertebrae tend to be more constant in shape and appearance over time. Figure 7.4 shows examples of both VB and IVD at different points in time. In other medical tasks the pair of VBs can easily be changed to other anatomies for example comparing lungs in chest X-rays. However, since our end goal is to reuse the weights of the Siamese networks for a disc classification task, we also explore training a network on the IVD pairs.

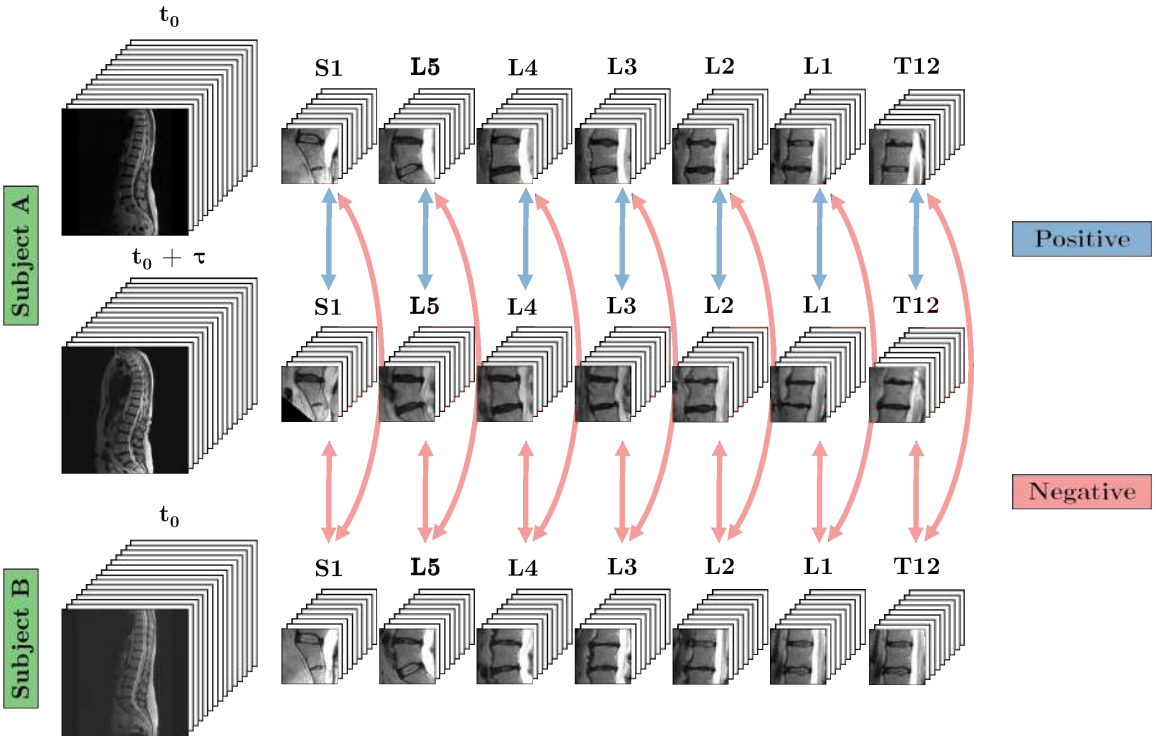


Figure 7.3: ***Generating the Positive/Negative Pairs.*** *The arrows mark a pair of VBs, where blue arrows highlight positive pairs while negative are highlighted in red. Each pair is generated from two scans.* $\mathbf{t_0}$ *refers to the time of the initial scan and* $\boldsymbol{\tau}$ *is the time between the baseline and the follow-up scans, typically 10 years in our dataset.*

Figure 7.4: ***VB and IVD Across Time.*** *The VBs appear unchanged but over time the IVDs lose intensity of its nucleus and in some cases experience some loss in height. Furthermore, in the IVD example in the top right, we can observe a vertebral slip, or spondylolisthesis, which does not change the appearance of the VBs themselves but significantly changes the IVD.*

7.3.3 Auxiliary Loss – Predicting VB & IVD Levels

In addition to the contrastive loss, we also employ an auxiliary loss to give complementary supervision. Since each VB pair is made up of VBs of the same level, we train a classifier on top of the **FC7** layer, i.e. the discriminative layer, to predict the seven levels of the VB (from T12 to S1) with the following softmax log loss:

$$\mathcal{L}_{VB} = -\sum_{n=1}^{N} \alpha_c \left(y_c(x_n) - \log \sum_{j=1}^{7} e^{y_j(x_n)} \right) \tag{7.3}$$

where y_j is the jth component of the **FC8** output, c is the true class of x_n, and α_c is the class-balanced weight as described in Chapter 4. Similarly, we predict the six IVD levels (T12-L1 to L5-S1) with a similar log loss:

$$\mathcal{L}_{IVD} = -\sum_{n=1}^{N} \alpha_c \left(y_c(x_n) - \log \sum_{j=1}^{6} e^{y_j(x_n)} \right) \tag{7.4}$$

The overall loss can then be described as a combination of the contrastive and softmax log losses:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{VB} \tag{7.5}$$

for the case with only the VB pair Siamese network and

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{VB} + \mathcal{L}_{IVD} \tag{7.6}$$

when both IVD and VB pairs are used.

7.3.4 CNN Architectures

The base architecture trained to distinguish VB pairs is based of the VGG-M network by Chatfield et al. (2014) with 3D kernels similar to the one described in Chapter 4.

7.3.4.1 VB Self-Supervision

The input to the Siamese CNN is the pairs of VBs, with dimension $224 \times 224 \times 9$, or IVDs, with dimension $112 \times 224 \times 9$. Similar to the network in Chapter 4, we use 3D kernels from **Conv1** to **Conv4** layers followed by a 2D **Conv5**. To transform the tensor to be compatible with 2D kernels, the **Conv4** kernel is set to be $3 \times 3 \times 9$ with no padding, resulting in a reduction of the slice-wise dimension after **Conv4**. We use 2×2 max pooling. The output dimension of the **FC8** layer depends on the number of classes i.e. seven for the self-supervisory auxiliary task of predicting the seven VB levels (see Figure 7.5).

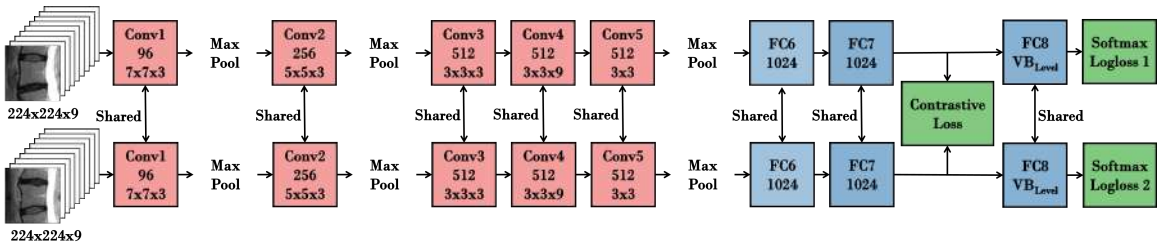


Figure 7.5: ***CNN Architecture – VB Only Siamese.*** *The Siamese network architecture is trained to distinguish between VB pairs from the same subject (same VB scanned a decade apart) or VB pairs from different subjects. There is also an additional loss (shown as VB_{Level}) to predict the level of the VB.*

7.3.4.2 VB + IVD Self-Supervision

Similar to the VB pair supervision, the network takes in the pairs of VBs while also in taking two extra input streams for pairs of IVD, with dimension $112 \times 224 \times 9$. The configurations of the weights are identical to before; instead of two streams sharing one set of weights, now we have four streams with one shared set of weights from **Conv1** to **Conv5**. The outputs are also similar but with the addition of IVD level softmax logloss (see Figure 7.6). The main idea behind training both VB and IVD pairs in tandem is so that the network learns not only the easier task of comparing VB pairs that roughly do no change in physiology in time (see Figure 7.4) but also

learns the physiology of the IVD which produces a more useful model for transfer learning to the IVD radiological grading classification task later on.

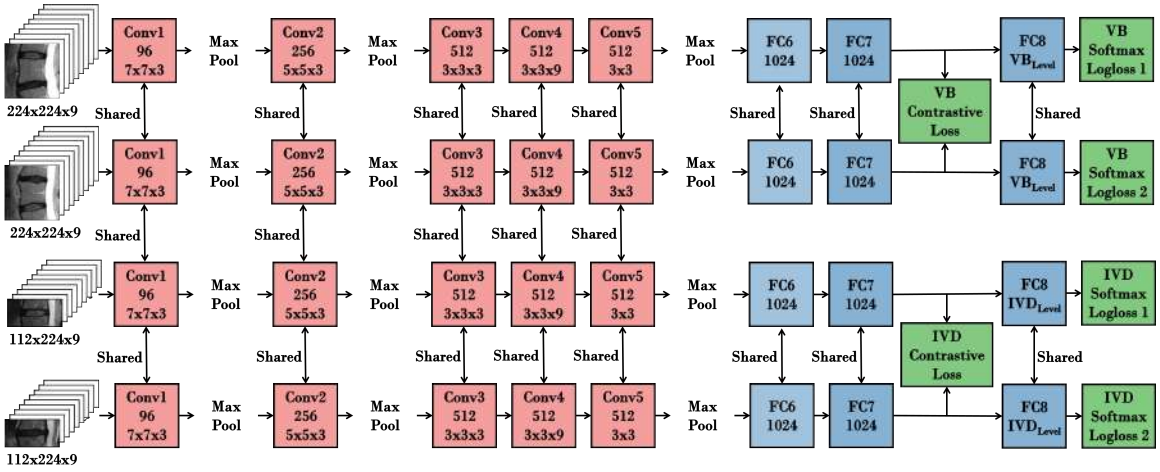


Figure 7.6: ***CNN Architecture – VB & IVD Siamese.*** *The Siamese network architecture is trained to distinguish between VB pairs and IVD pairs from the same subject or VB pairs and IVD pairs from different subjects. There is also an additional loss (shown as VB_{Level} and IVD_{Level}) to predict both the levels of the VB and the IVD.* ***Conv1*** *to* ***Conv5*** *are shared for all four streams while* ***FC6*** *to* ***FC8*** *are shared for each respective input stream, VBs or IVDs.*

7.3.4.3 IVD Radiological Grading Classification

The input to this CNN is an IVD volume, with dimension $112 \times 224 \times 9$. The network is identical to previous networks, albeit only one stream of the multi-stream Siamese network, resulting in interchangeable weights between this network and any of the networks for self-supervision. The end goal is to classify the **disc degeneration** radiological grading in the **TwinsUK** dataset as defined in Chapter 3.

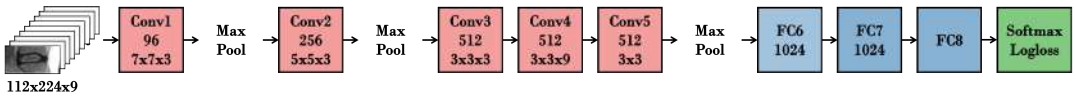


Figure 7.7: ***CNN Architecture – IVD Radiological Grading Classification.*** *The architecture of the classification CNN. The convolutional weights are transferable from the Siamese network in Figure 7.5 or Figure 7.6.*

7.3.5 Implementation Details

7.3.5.1 Data Augmentation

The augmentation strategies are identical to that described in Chapter 4 for the classification CNN and IVD Siamese CNN, while we use slightly different augmentations to train the VB Siamese network. The augmentations for the VB CNN are:

1. Rotation with $\theta = -15°$ to $15°$
2. Translation of ± 48 pixels in the x-axis, ± 24 pixels in the y-axis, ± 2 slices in the z-axis
3. Rescaling with a scaling factor between 90% to 110%
4. Intensity variation between -0.2 to 0.2
5. Random slice-wise flip i.e. reflection of the slices across the mid-sagittal

7.3.5.2 Training Details

Our implementation is based on the MatConvNet toolbox by Vedaldi and Lenc (2014b). The networks were trained using NVIDIA Titan X GPUs. The hyperparameters are: batch size 128 for classification and 32 for the Siamese network; momentum 0.9; weight decay 0.0005; learning rate $1e^{-3}$ (classification) and $1e^{-5}$ (self-supervision) and lowered by a factor of 10 as the validation error plateaus, which is also our stopping criterion, normally around 2000 epochs for the classification network and 1000 epochs for the Siamese network.

7.4 Experiments & Results

7.4.1 Self-Supervision

The dataset is split by subject 80:10:10 into train, validation, and test sets. For **TwinsUK** with 423 subjects with multiple scans this results in a 339:42:42 split, while for **OSCLMRIC** the 530 subjects was split into different sets of 424:53:53 per subject. Unlike **TwinsUK** which has only one or two follow-up scans per subject at a regular interval, **OSCLMRIC** possess multiple follow-up scans at irregular scans per subject. Note, for **TwinsUK** a pair of twins will only be in one set i.e. one subject part of a twin pair can't be in training and the other in test. We also do not compare twin pairs to train the network. Though the number of subjects are similar, the number of positive samples, $y = 1$, are higher in **OSCLMRIC** since the average amount of scans per subject are higher than **TwinsUK**; 6522 IVD & 7609 VB positive samples in **OSCLMRIC**, 2070 IVD & 2415 VB positive samples in **TwinsUK**.

With the trained network, each input VB or IVD can be represented as a 1024-dimensional **FC7** vector. For each pair of VBs or IVDs i.e. two **FC7** vectors, we can calculate the L_2 distance, d (Equation 7.2), between two samples, which is the same distance used during training. Figure 7.8 and Figure 7.10 show histograms of the distances for all the VB pairs (both positive and negative) in the test set using the network trained on just VBs. Similarly, Figure 7.9 and Figure 7.11 show similar histograms for both VB and IVD pairs and the corresponding ROC of classification with network that was trained on VB and IVD pairs simultaneously. In general, VB pairs that are from the same subject have lower distances compared to pairs from different subjects. We also obtain a very good performance of accuracy on the auxiliary task of predicting VB level; **97.8%** for the **TwinsUK** trained VB network. Table 7.1 shows the AUCs for the VBs and IVDs pairs positive/negative classifications. The

performance improvement in learning both VB and IVD pairs together is easier to see for the case of **TwinsUK**, (**94.8**% → **96.4**%), which has a much lower amount of data than **OSCLMRIC**. Nevertheless, it can be seen that the performance of VB classification is slightly better in both datasets, probably due to VBs remaining unchanged over the years as we hypothesized in Figure 7.3.

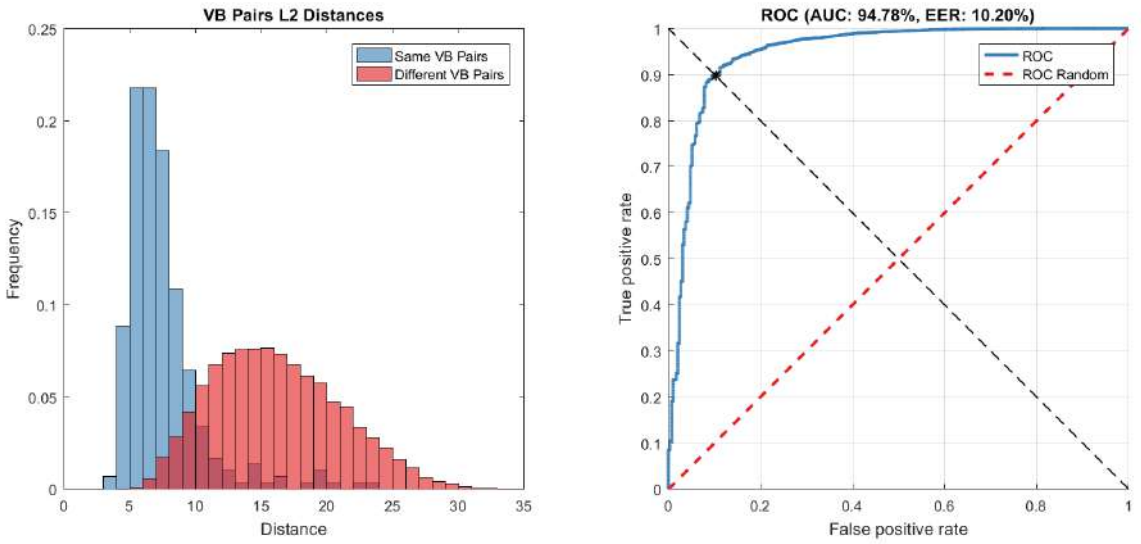


Figure 7.8: ***Results – Siamese-A.*** *Siamese-A in Table 7.1 trained only on VB pairs from the* ***TwinsUK*** *dataset. The counterpart to this CNN, Siamese-C can been seen in Figure 7.10 trained on the* ***OSCLMRIC*** *dataset.* ***Left:*** *Histogram of VB pairs distances in the* ***TwinsUK*** *test set. Positive pairs in blue, negative in red.* ***Right:*** *The ROC of the classification of positive/negative VB pairs.*

Model	Dataset	Trained on		$\mathbf{AUC_{VB}}$	$\mathbf{AUC_{IVD}}$
		VB	IVD		
Siamese-A	**TwinsUK**	✓		94.8	-
Siamese-B	**TwinsUK**	✓	✓	96.4	95.5
Siamese-C	**OSCLMRIC**	✓		**99.7**	-
Siamese-D	**OSCLMRIC**	✓	✓	99.5	**99.2**

Table 7.1: ***Self-Supervision Results.*** $\mathbf{AUC_{VB}}$ *is the AUC for VB classification and* $\mathbf{AUC_{IVD}}$ *is the AUC for IVD classification.*

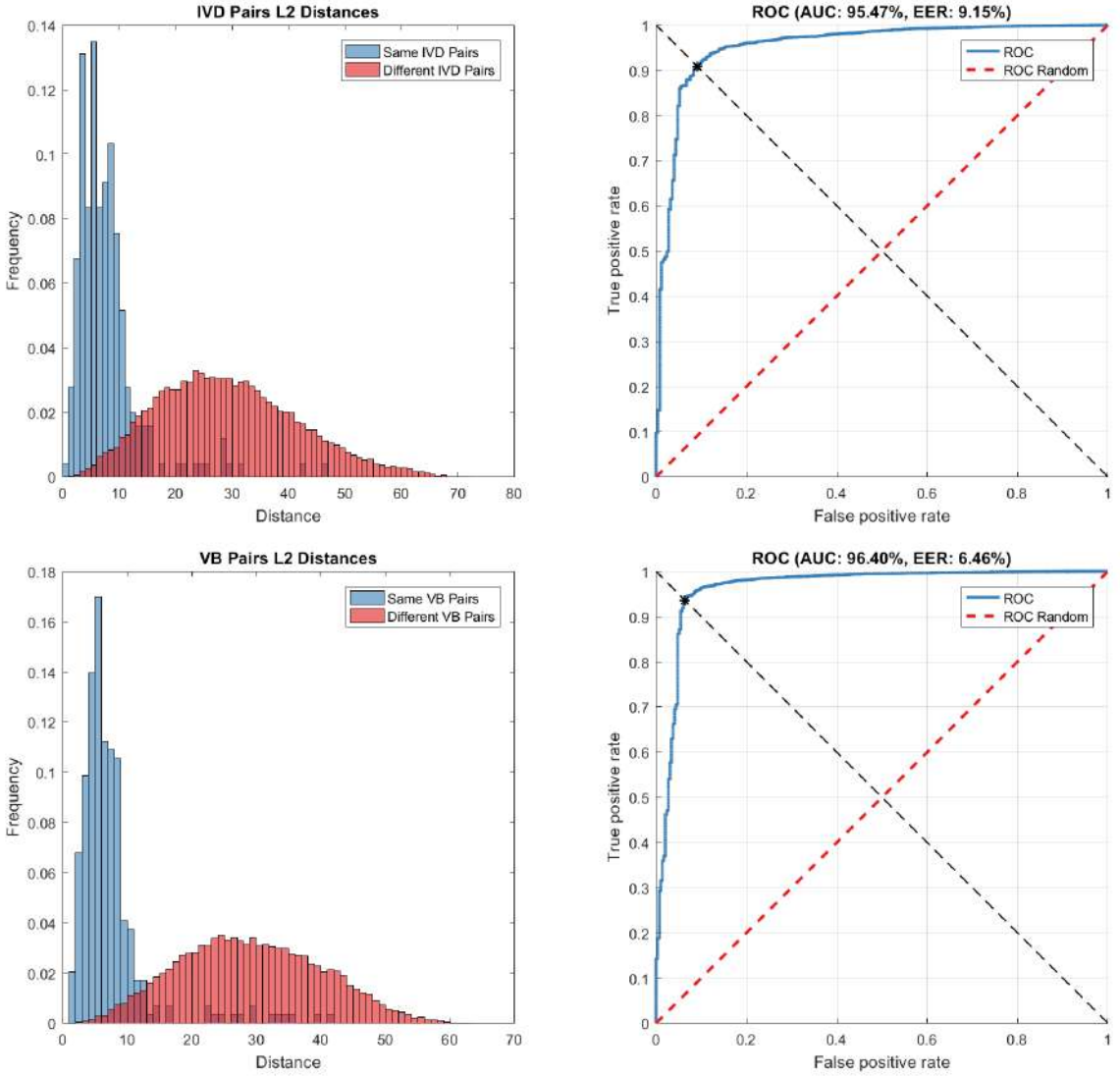


Figure 7.9: ***Results – Siamese-B.*** *Siamese-B in Table 7.1 trained on both VB and IVD pairs from the* ***TwinsUK*** *dataset. The main difference from Figure 7.8 is that we train this CNN with both IVD and VB inputs simultaneously. This is similar to Figure 7.11 but that was trained on the* ***OSCLMRIC*** *dataset.* ***Top Left:*** *Histogram of IVD pairs distances in the* ***TwinsUK*** *test set. Positive pairs in blue, negative in red.* ***Top Right:*** *The ROC of the classification of positive/negative IVD pairs.* ***Bottom Left:*** *Histogram of VB pairs distances in the* ***TwinsUK*** *test set. Positive pairs in blue, negative in red.* ***Bottom Right:*** *The ROC of the classification of positive/negative VB pairs.*

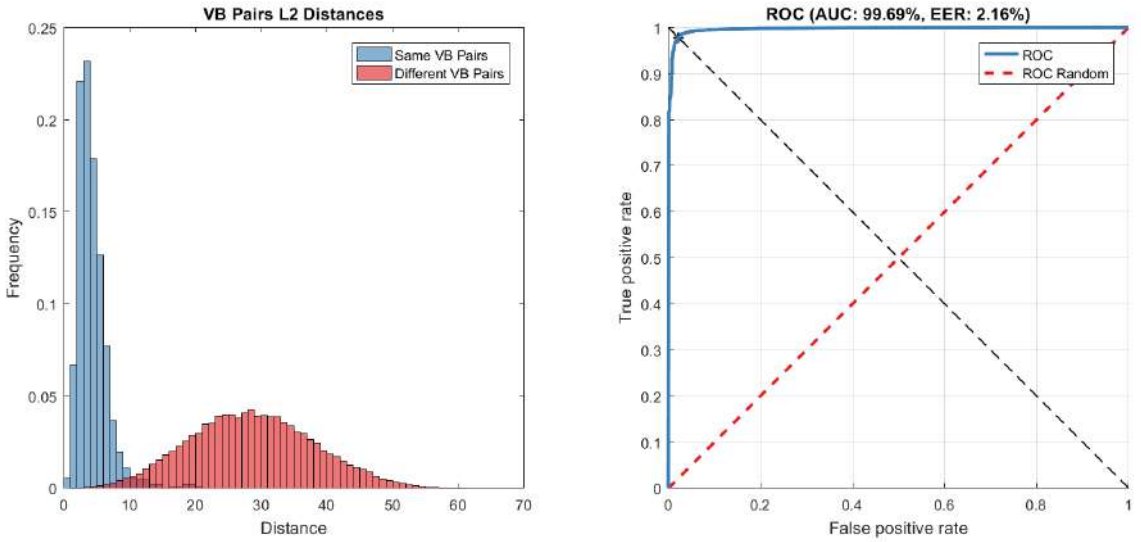


Figure 7.10: ***Results – Siamese-C.*** *Siamese-C in Table 7.1 trained only on VB pairs from the* ***OSCLMRIC*** *dataset. The* ***TwinsUK*** *counterpart, Siamese-A can be seen in Figure 7.8.* ***Left:*** *Histogram of VB pairs distances in the* ***OSCLMRIC*** *test set. Positive pairs in blue, negative in red.* ***Right:*** *The ROC of the classification of positive/negative VB pairs.*

7.4.2 Benefits of Pre-training on Disc Degeneration Classification

To measure the performance gained by pre-training using the longitudinal information we use the convolutional weights learnt in the Siamese network, and train a classification CNN (see Figure 7.7) to predict the **disc degeneration** radiological grading. For this classification task we use the 920 subjects in **TwinsUK** that possess gradings and split them into the following sets: 670 for training, 50 in validation, and 200 for testing. Subjects with follow-up scans (> 1 scans), the same scans used to train the self-supervised Siamese networks, are only used in training and not for testing so, in essence, the Siamese networks will never have seen the subjects in the classification test set.

We transfer and freeze convolutional weights of the Siamese networks and only train the randomly-initialized fully-connected layers (marked as Cyan in Figure 7.7). We also experimented with fine-tuning the convolutional layers but we find the dif-

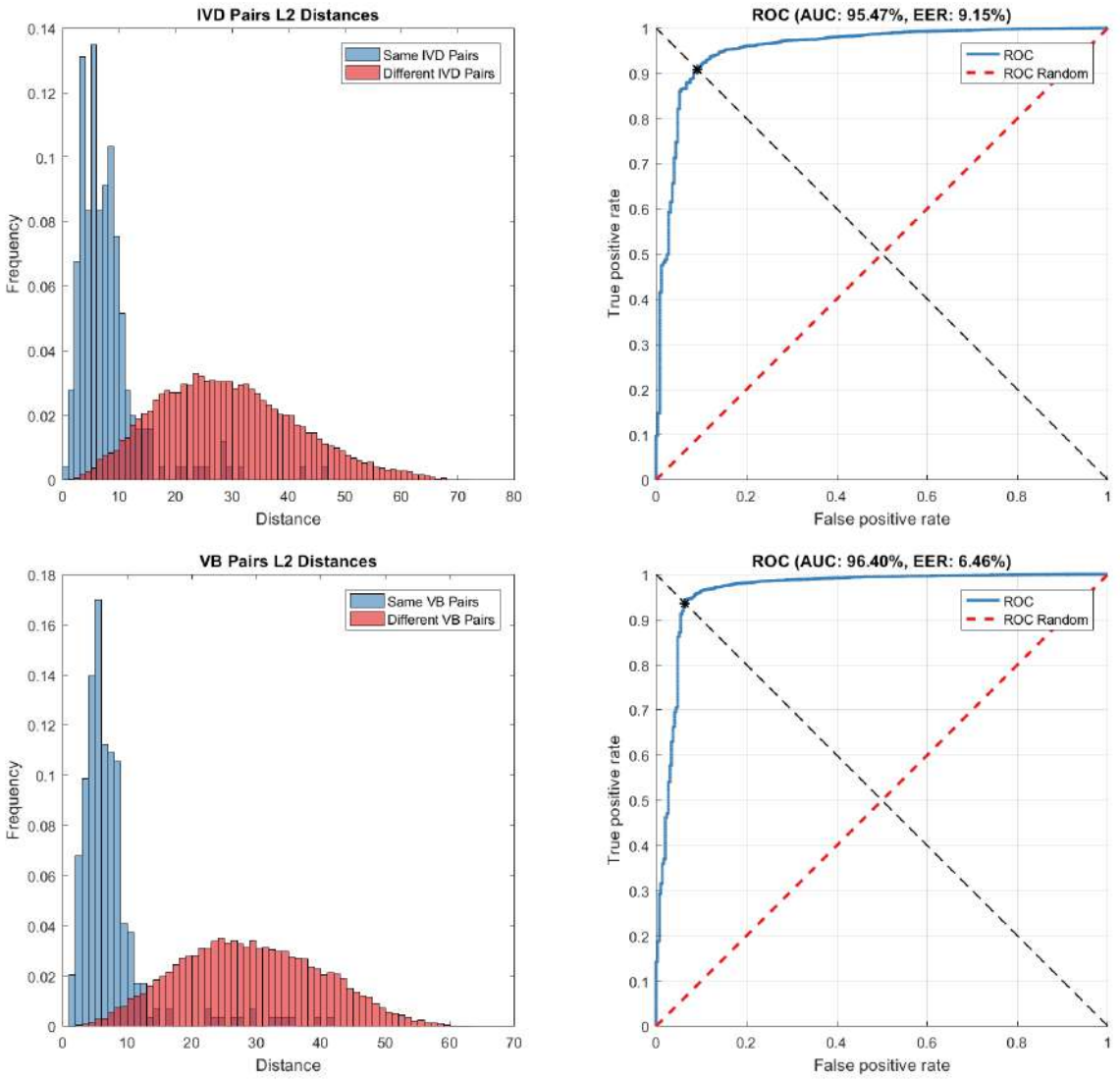


Figure 7.11: ***Results – Siamese-D.*** *Siamese-D in Table 7.1 trained on both VB and IVD pairs from the* ***OSCLMRIC*** *dataset. The CNN is similar to Siamese-C in Figure 7.10 trained only on VB pairs.* ***Top Left:*** *Histogram of IVD pairs distances in the* ***OSCLMRIC*** *test set. Positive pairs in blue, negative in red.* ***Top Right:*** *The ROC of the classification of positive/negative IVD pairs.* ***Bottom Left:*** *Histogram of VB pairs distances in the* ***OSCLMRIC*** *test set. Positive pairs in blue, negative in red.* ***Bottom Right:*** *The ROC of the classification of positive/negative VB pairs.*

ference in performance to be negligible. For comparison, we also train: (i) a CNN from scratch, (ii) a CNN with a frozen randomly initialized convolutional layers (to see the power of just training on the fully-connected layers) as a baseline, and (iii) a CNN using convolutional weights of a CNN trained on a fully-annotated spinal MRI dataset, **Genodisc**, with multiple radiological gradings as in Chapter 4. The performance measure is the average class accuracy, calculated as the average of the diagonal elements of the normalized confusion matrix.

Figure 7.12 shows the performance of all the models as the number of training samples is varied at [120, 240, 361, 670] subjects or [225, 444, 667, 976] scans. It can be seen that with longitudinal self-supervision pre-training, fewer data is required to reach an equivalent point to that of training from scratch, e.g. the performance is **74.4%** when using only 667 scans for pre-training, whereas training from scratch requires 976 scans to get to **74.7%**. This performance gain can also be seen with a lower amount of training data. As would be expected, transfer learning from a CNN trained with strong supervision on the **Genodisc** dataset (see Chapter 4) is better, with an accuracy at least **2.5%** above training from scratch. Unsurprisingly, a classifier trained on top of fully random convolutional weights performs the worst.

For comparison, Figure 7.13 compares performances of several networks using convolutional weights transferred from different Siamese networks (note that we do not use augmentations on training here). Surprisingly, transferring from **OSCLMRIC** self-supervised Siamese network works better than transferring from a Siamese network trained on **TwinsUK** itself, with the performance nearing that of transferring from a fully supervised network trained on **Genodisc** (Chapter 4). Adding augmentations during training might further boost the performance gain transferring from **OSCLMRIC**, which seems to be true for other cases in Figure 7.14. Performance with using both VB and IVD pairs Siamese network seems to be slightly beneficial only when the number of training data is low i.e. only the Siamese network trained

on **TwinsUK** benefits from slightly more supervision as shown in Figure 7.13.

Since longitudinal information is essentially freely available when collecting data, the performance gain from longitudinal pre-training is also free. It is interesting to note that even though the Siamese networks are trained on a totally different task of distinguishing similarity/dissimilarity of VB or IVD pairs, when transferred still manage to achieve better performance at predicting radiological grading of IVDs than starting from scratch.

7.4.3 Zygosity

In **TwinsUK**, unsurprisingly, there exist MRIs of twins which can be monozygotic (identical) or dizygotic (fraternal). Figure 7.15 shows some examples of VB pairs of twins, both monozygotic and dizygotic. Since, we have essentially trained a model to distinguish between pairs of VBs (Siamese-A), we can now compare distances between the following pairs: (i) the same individual at different time points, (ii) two totally different individuals, (iii) monozygotic twins, and (iv) dizygotic twins. The Siamese-A network was not trained on any twin pairs i.e. twin pairs were completely omitted during training. Of the 42 subjects in the test set in Section 7.4.1, 21 are twins; 5 monozygotic and 16 dizygotic. Figure 7.16 shows the distribution of distances for the four types of pairings. The distances of the four pairing are:

$$d_{same} < d_{monozygotic} < d_{dizygotic} < d_{different} \tag{7.7}$$

which suggests that: (i) twins in general have a smaller distance than random strangers but still have a larger distance compared to pairs of MRIs of the same individual, (ii) scans of the same individual at different time points have the smallest distance than any other scans, and (iii) more interestingly, monozygotic twins have a smaller distance than dizygotic twins. Since the network produces such a clear separation of

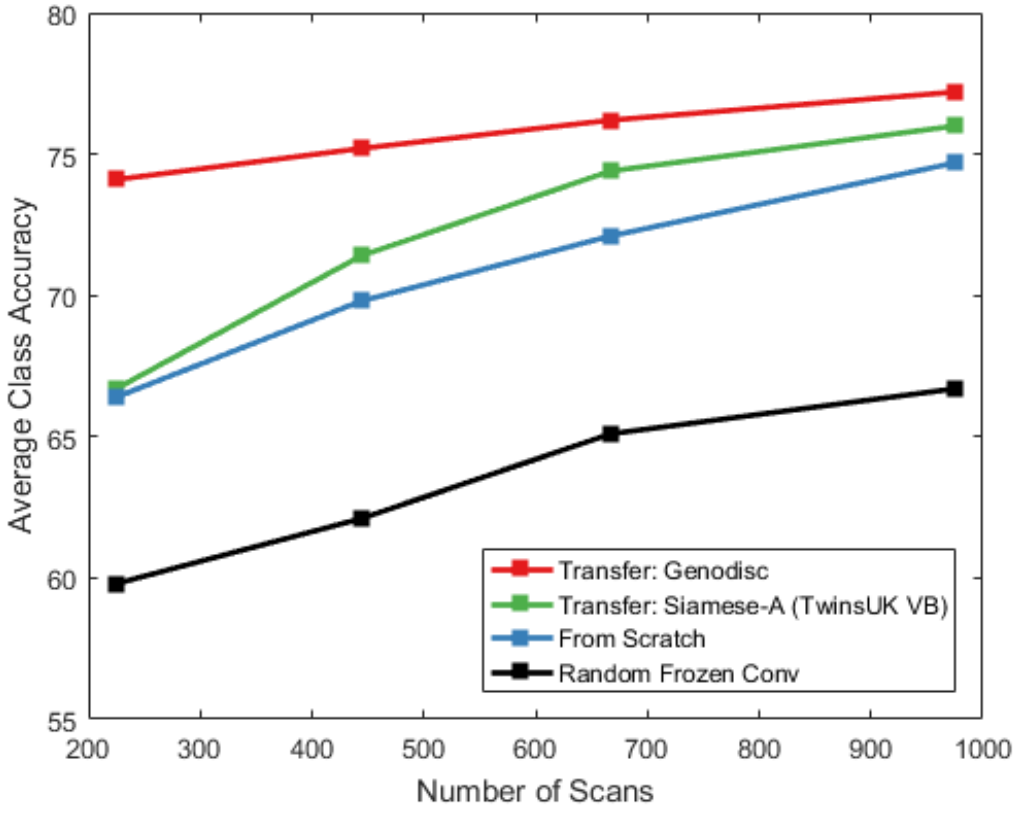


Figure 7.12: ***Disc Degeneration Accuracy with Train-time Augmentation.*** *Accuracy of* ***disc degeneration*** *classification as the number of training samples is increased; increasing from left to right. Transferring weights from a self-supervised network improves performance over training from scratch, showing its benefit, and this performance boost persists even at 976 scans. Transfer learning from a CNN trained on a strongly-supervised dataset (IVD radiological grading classification) is better and provides an 'upper bound' on transfer performance. Note, even with 976 scans in the training set, the performance has not plateaued hinting at further improvements with the availability of more data.*

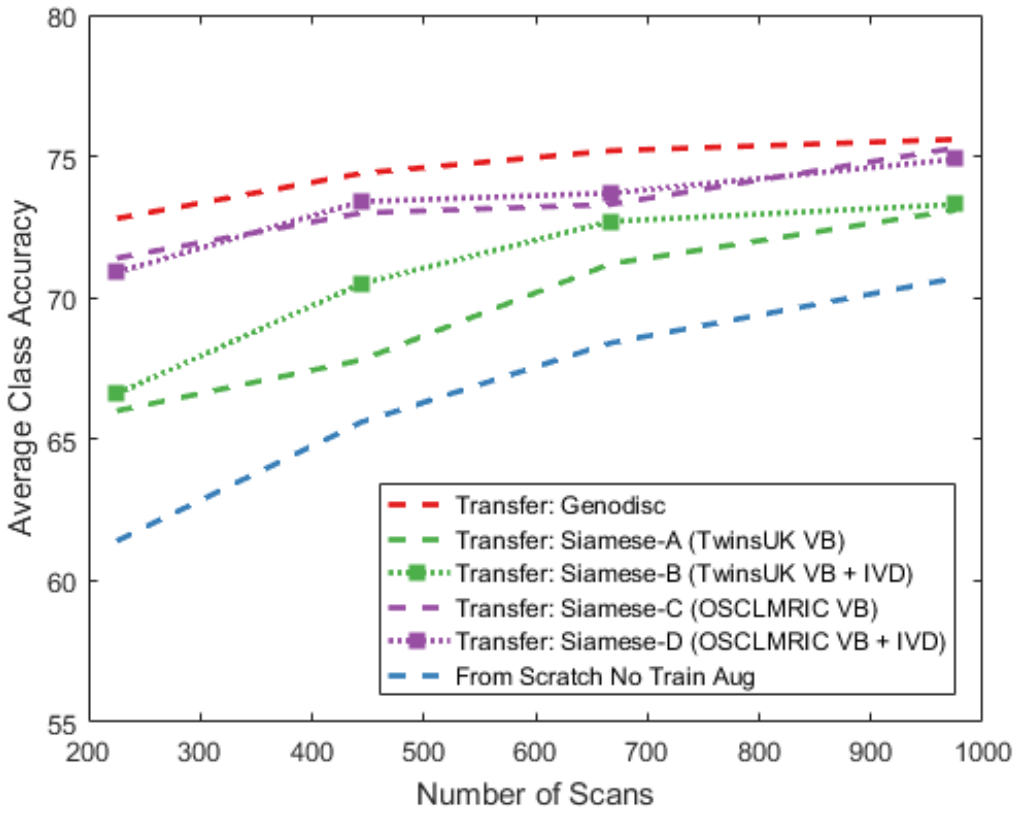


Figure 7.13: ***Disc Degeneration Accuracy with No Train-time Augmentation.*** *Accuracy of* ***disc degeneration*** *classification as the number of training samples is increased; increasing from left to right. It can be seen that the performance gains in self-supervision pre-training on the same dataset,* ***TwinsUK****, going from left to right are* ***+5.2%****,* ***+4.9%****,* ***+4.3%****, and* ***+2.6%*** *(comparing Siamese-B shown as a dotted green line against a CNN trained from scratch shown in cyan). The most interesting comparison in this plot is between the CNNs pre-trained on the* ***OSCLMRIC*** *dataset shown as purple lines and those pre-trained on the* ***TwinsUK*** *dataset (the same dataset we test for* ***disc degeneration*** *grading). It can be seen this pre-training nearly reaches the performance of pre-training on the* ***Genodisc*** *dataset which is stronger in terms of supervision. Generally, the gap between a pre-trained CNN and a CNN traine from scratch gets smaller as the training sample increases. Note, unlike Figure 7.12 there is no train-time augmentation for these comparisons. Benefits of train-time augmentation can be seen in Figure 7.14.*

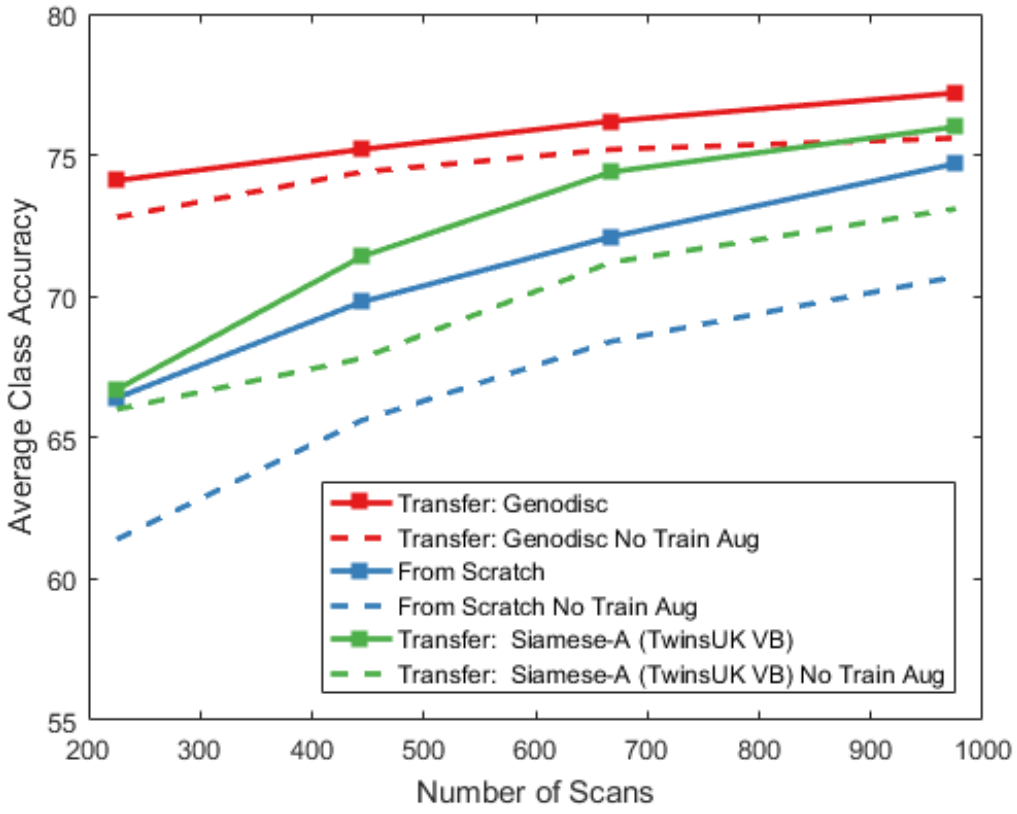


Figure 7.14: ***Disc Degeneration Accuracy: Train-time Augmentation vs No Train-time Augmentation.*** *Accuracy of* ***disc degeneration*** *classification as the number of training samples is increased; increasing from left to right. Transferring convolutional weights from different networks with and without train-time augmentations. Performance gain seems to be consistent i.e. training with augmentations consistently gives better performance.*

distances for different pairings, this means that there is a correlation between genetics and the shape of the spine. Moving on, an interesting observation is that genetically identical individuals, monozygotic twins, possess visually similar spines but due to epigenetic factors, monozygotic spines are still slightly dissimilar when compared to spines of a unique individual at different points in time. This suggests that the shape of the spine is affected more by environmental factors than time or ageing.

7.5 Summary

In this chapter, we have shown that it is possible to use self-supervision to improve performance on a radiological grading classification task. In short, we show that the performance of the pre-trained CNN on the supervised classification task is (i) superior to that of a network trained from scratch; and (ii) requires far fewer annotated training samples to reach an equivalent performance to that of the network trained from scratch. We hope to explore the benefits of adding more auxiliary tasks in the near future. The performance improvement is nearing that of transfer learning from a CNN trained on a fully annotated dataset given that the target training set itself contains enough data. Furthermore, having a distance measure between vertebral pairs opens up the possibility of identifying people using their MRIs. This chapter concludes our study on radiological gradings and we move on to summarize the whole thesis while briefly touching upon some extensions including looking at the relationship between MRIs and disability in the next chapter.

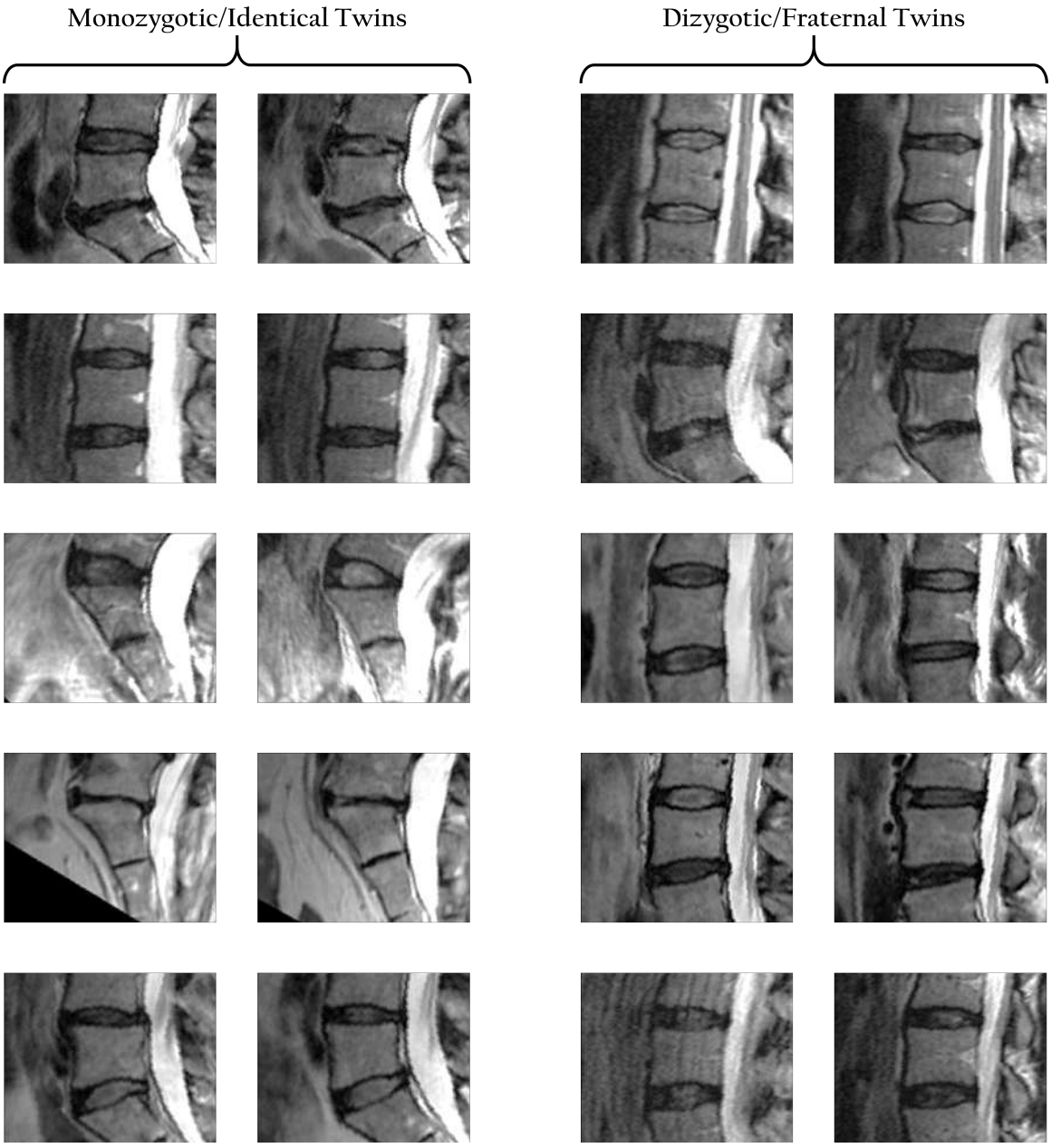


Figure 7.15: ***Vertebral Body Examples Of Monozygotic and Dizygotic Twins.*** ***Left:*** *VB pairs from monozygotic/identical twins.* ***Right:*** *VB pairs from dizygotic/fraternal twins.*

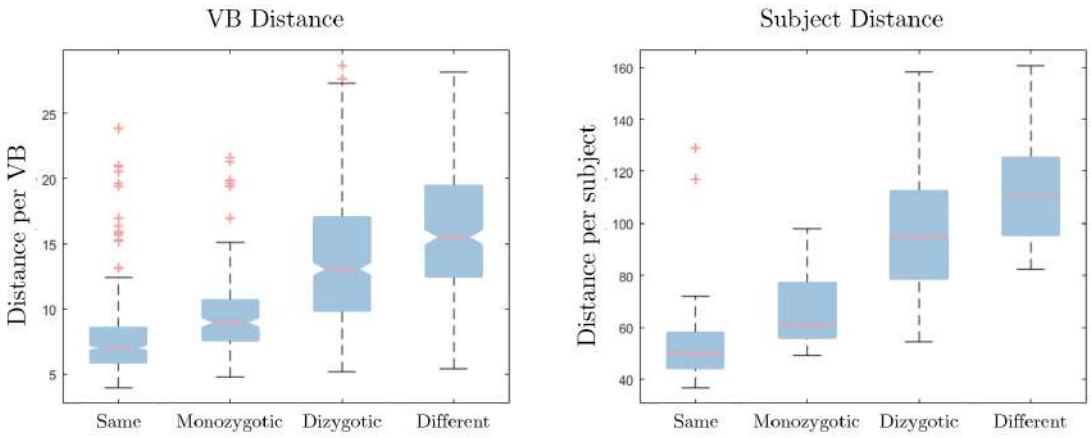


Figure 7.16: ***Box & Whisker Plots – Distances of Various Types of Pairings.*** ***Left:*** *Distances on a per VB basis.* ***Right:*** *Distances on a per subject basis, calculated by tallying up the distances of the 7 VBs for each subject pairs.'Same' refers to pairs of the same individual scanned at different times, and 'Different' refers to pairs of totally different individuals.*

Chapter 8

Summary & Extensions

To conclude, here in this section we summarize the contributions of this thesis, from the research covered in Chapters 3–7, followed by several extensions which include a brief foray into mapping to clinical disability, a web demo built upon the work done in Chapter 4, and discussions of future work.

8.1 Summary

In this thesis we have proposed several methods for automated reading of spinal MRIs including: (i) predicting radiological gradings, and (ii) producing weak localization of evidence predictions which we term evidence hotspots. We examined the possibility of using multiple modalities for predictions, and found ways to learn from essentially free information, in the form of longitudinal scans. We have also built a live demo for automated grading of spinal MRIs which will allow easy access to a portion of this research.

In Chapter 3, we look at different datasets and the necessary pre-processing steps used in subsequent chapters. The three datasets were examined and found to be both in subject demographics and acquisition protocols. We built upon the method proposed by Lootus et al. (2013), improving the labelling performance (95.6%), and

refined the detections such that the outputs are now extremely tight bounding boxes through regression of the vertebral bodies corner points. We explored two different methods for regression: (i) SDM on SIFT features, and (ii) CNN regression. We concluded that the CNN regressor slightly outperforms SDM.

In Chapter 4, we proposed a scheme for classifying multiple radiological gradings of intervertebral discs. We trained a multi-task CNN classifier to predict the gradings simultaneously per disc. Each volume of the intervertebral disc input was extracted by modifying pairs of vertebral body volumes, upper and lower vertebrae with respect to each disc, detected in Chapter 3. We experimented with multi-tasking and found that performance increases with the addition of more task, at least for our tasks using data in the **Genodisc** dataset.

In Chapter 5, we explored the possibility of localizing pathologies of the intervertebral discs only using the classification CNN trained in Chapter 4. The localizations produced from the CNN inherently pinpoint the exact regions in the inputs influencing the predictions i.e. evidence hotspots. We explored different methods to produced these localizations or saliency maps; the best of which is excitation backpropagation. We showed that these hotspots can be produced for any of the gradings the CNN was trained for and we explored the quantitative performance of these qualitative localizations. The code with a demo to run several of the methods to produce the saliency maps, particularly those used in Chapter 5, is available from `https://github.com/amirjamaludin/MatConvNet-Saliency-Visualization`.

In Chapter 6, we returned to predicting radiological gradings and explored whether it is beneficial to add different types of sequences and planes. We found that certain gradings do benefit from certain addition; **disc herniation** seems to benefit with the addition of axial scans while a majority of the gradings learnt in Chapter 4, apart from **Pfirrmann grading** and **disc narrowing** which were graded by the radiologist only using T2-weighted sagittal scans, benefited from adding T1-weighted sagittal scans.

Now since we have only focused on classification on extremely constrained detected intervertebral discs, we also explore whether disc specific predictions can be made directly from raw MR volumes. However, it seems that direct classification from the raw volume is not comparable to per disc volume classification.

In Chapter 7, we looked at training a CNN model only using freely available longitudinal data, more specifically training using baseline and follow-up scans. We proposed a method of training a Siamese CNN taking as inputs, vertebral bodies or intervertebral discs with the end goal of learning similar or dissimilar pairs. This CNN in itself is not that beneficial, except for the task of identifying spinal MRIs of the same subject, but the learnt weights themselves are beneficial for transfer learning. We transfer these weights for **disc degeneration** classification task in the **TwinsUK** dataset which we found to be extremely better than training from scratch. In fact, the best results we get for the classification of **disc degeneration** in the **TwinsUK** dataset, aside from transfer learning from a CNN trained from the **Genodisc** dataset with multiple gradings, was from transfer learning from the **OSCLMRIC** dataset which do no possess any form of radiological labelling (aside from the MRI reports).

8.2 Extensions

8.2.1 Predicting ODI from MRIs

The **OSCLMRIC** dataset is unique in that it possesses clinical information of the subjects. The information can be in the form of scores and they were obtained via surveys and questionnaires. These scores are important in spinal research especially the study to find the correlation between back pain and radiological gradings, labelled by radiologists. However, mapping gradings to pain is ongoing research and it is extremely difficult to establish the relationship between the two especially since pain measures are highly subjective between different subjects. As such, a better approach

would be to map to a more objective measure of disability. One example of such measure is the Oswestry Disability Index (**ODI**) which is made up of 10 separate sections, which can be scored from 0 to 5, where each section describes a unique point of disability. The 10 sections are:

1. Pain intensity
2. Personal care
3. Lifting
4. Walking
5. Sitting
6. Standing
7. Sleeping
8. Sex life
9. Social life
10. Travelling

Details of each section can be seen in Figure 3.13 (Chapter 3). The more disabled a subject is, the higher the **ODI** score. Only 795 subjects in **OSCLMRIC** possess **ODI** scores which we split into 645:75:75, train:val:test sets.

Learning from Disc Features. Since previous works on correlating clinical measurements to MRIs have always started with radiological features of the intervertebral disc, we propose that we use a CNN pre-trained on **Genodisc** to produce intermediate features. As such we use a 3D CNN with a branch point at **FC6**, which is 1024-dimensional, which is a slight modification to the one in Chapter 4 to produce a feature vector for each of the six lumbar disc; T12-L1 to L5-S1. These disc features are concatenated resulting in a feature vector of the dimension 1024×6. We then train a CNN to predict each section of the **ODI** questionnaire for each subject. The convolutional kernels of the CNN is set to be 1×6 so that the CNN is able to learn "inter-disc" features. Figure 8.1 shows the overview of the training procedure. Training details are similar to the procedure previously described in Chapter 4 and Chapter 6. We keep each task for training to be multi-class as each **ODI** section is unique and finding a suitable point to threshold the questionnaire is not trivial.

Preliminary Results. We find that the performances of the multi-class predic-

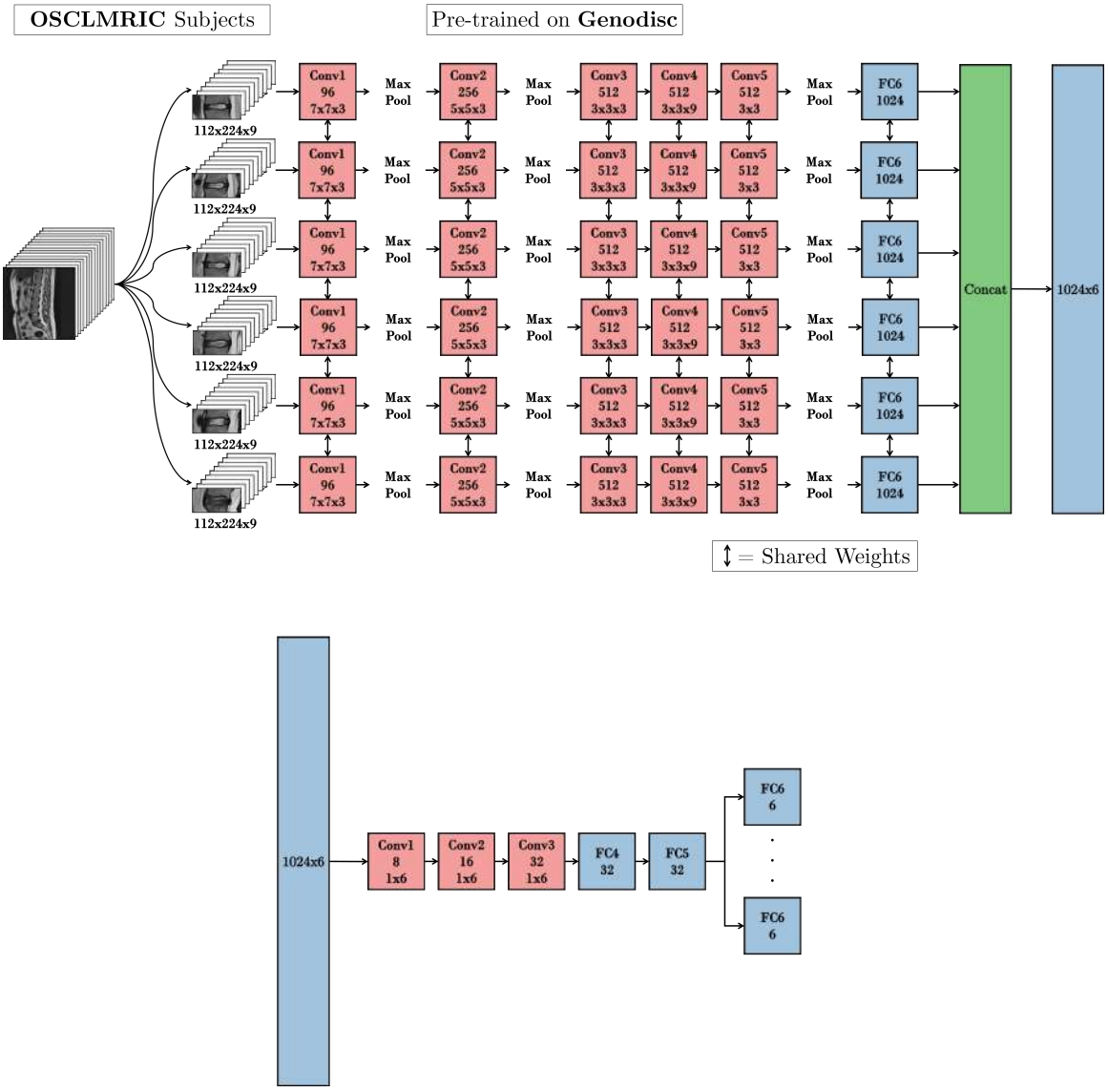


Figure 8.1: ***CNN Architecture – Predicting ODI. Top:*** *A CNN trained on* ***Genodisc*** *is used to produce disc features. These features are then concentrated for each disc resulting in* 1024×6 *feature vector.* ***Bottom:*** *CNN trained to predict the 10* ***ODI*** *scores using the disc features from the CNN trained on* ***Genodisc****.*

tions for all ten sections are still quite low. However, by binarizing the scores (best threshold found via the validation set), essentially grouping up subjects into "normal" or "abnormal" groups, we achieve respectable performances for a couple of the **ODI** sections. Results can be seen in Table 8.1.

	Best Grouping	Class Average Accuracy
Pain intensity	{0,1,2,3,4} ∈ Normal {5} ∈ Abnormal	70.8
Personal care	{0} ∈ Normal {1,2,3,4,5} ∈ Abnormal	53.2
Lifting	{0,1,2,3} ∈ Normal {4,5} ∈ Abnormal	44.5
Walking	{0,1} ∈ Normal {2,3,4,5} ∈ Abnormal	62.1
Sitting	{0,1,2} ∈ Normal {3,4,5} ∈ Abnormal	63.5
Standing	{0,1,2,3} ∈ Normal {4,5} ∈ Abnormal	53.5
Sleeping	{0,1,2,3} ∈ Normal {4,5} ∈ Abnormal	50.8
Sex life	{0,1,2} ∈ Normal {3,4,5} ∈ Abnormal	50.0
Social life	{0,1,2} ∈ Normal {3,4,5} ∈ Abnormal	49.8
Travelling	{0,1} ∈ Normal {2,3,4,5} ∈ Abnormal	52.5

Table 8.1: ***Predicting ODI Results.*** *Binary classification performance for all the 10* ***ODI*** *sections, scored from 0 to 5. Binary threshold for each section found in the validation set. Chance accuracy is 50%.*

8.2.2 SpineNet Online Demo

An on-line live demo of the full automatic grading system is available at `http://zeus.robots.ox.ac.uk/spinenet/`. The back-end of the demo was developed as part of this thesis while the front-end was developed in collaboration with Abhishek Dutta. The current version of the landing page can be seen in Figure 8.2 followed by some examples in Figure 8.3 and Figure 8.4.

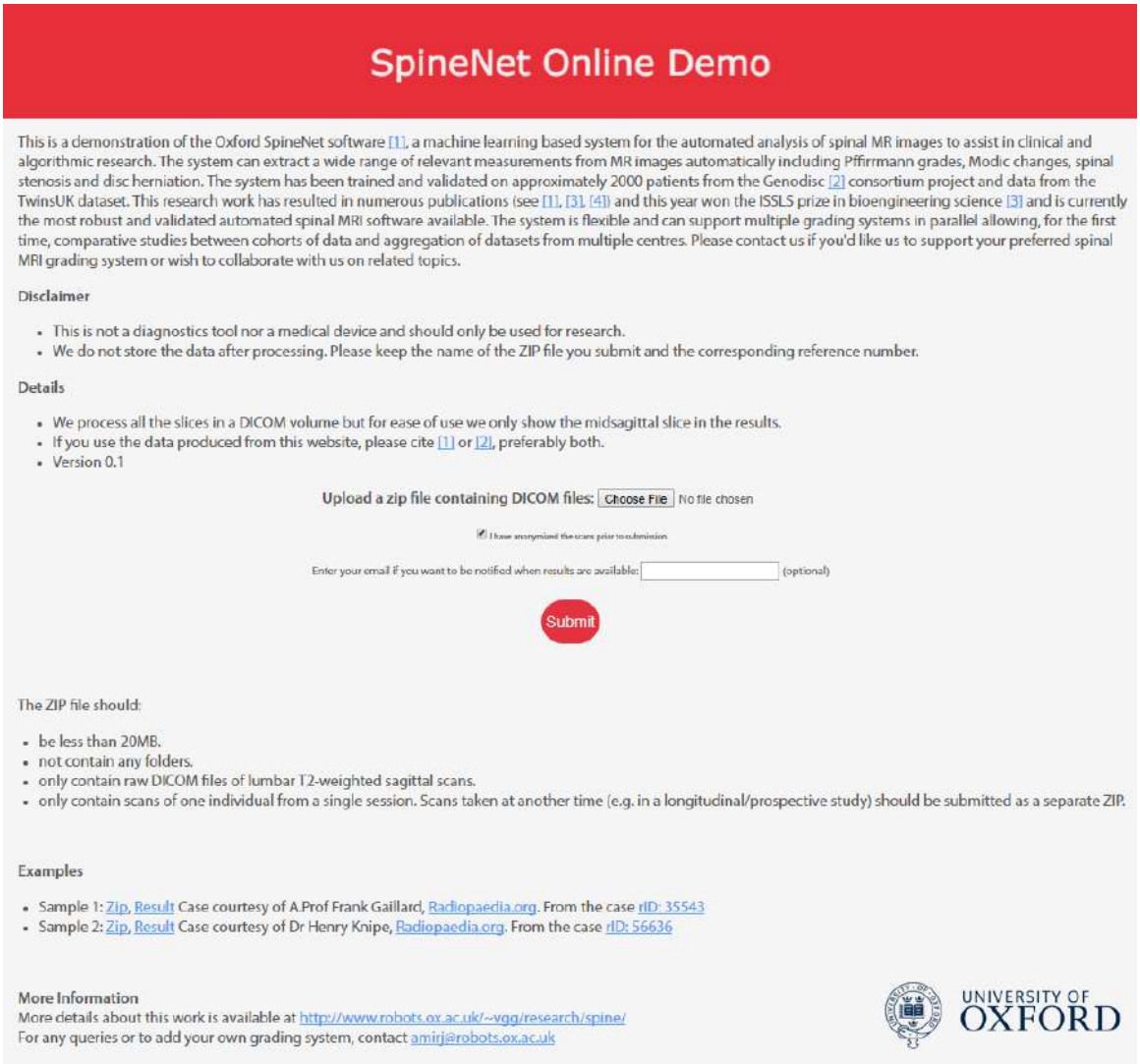


Figure 8.2: ***Landing Page of the SpineNet Online Demo.*** *Users can upload DICOM files (zipped) containing T2-weighted sagittal scans and the system will produce the eight gradings covered in this chapter alongside tight bounding quadrilaterals of the vertebral bodies alongside their level labels as discussed in Chapter 3. Processing a scan takes around a minute from start to finish.*

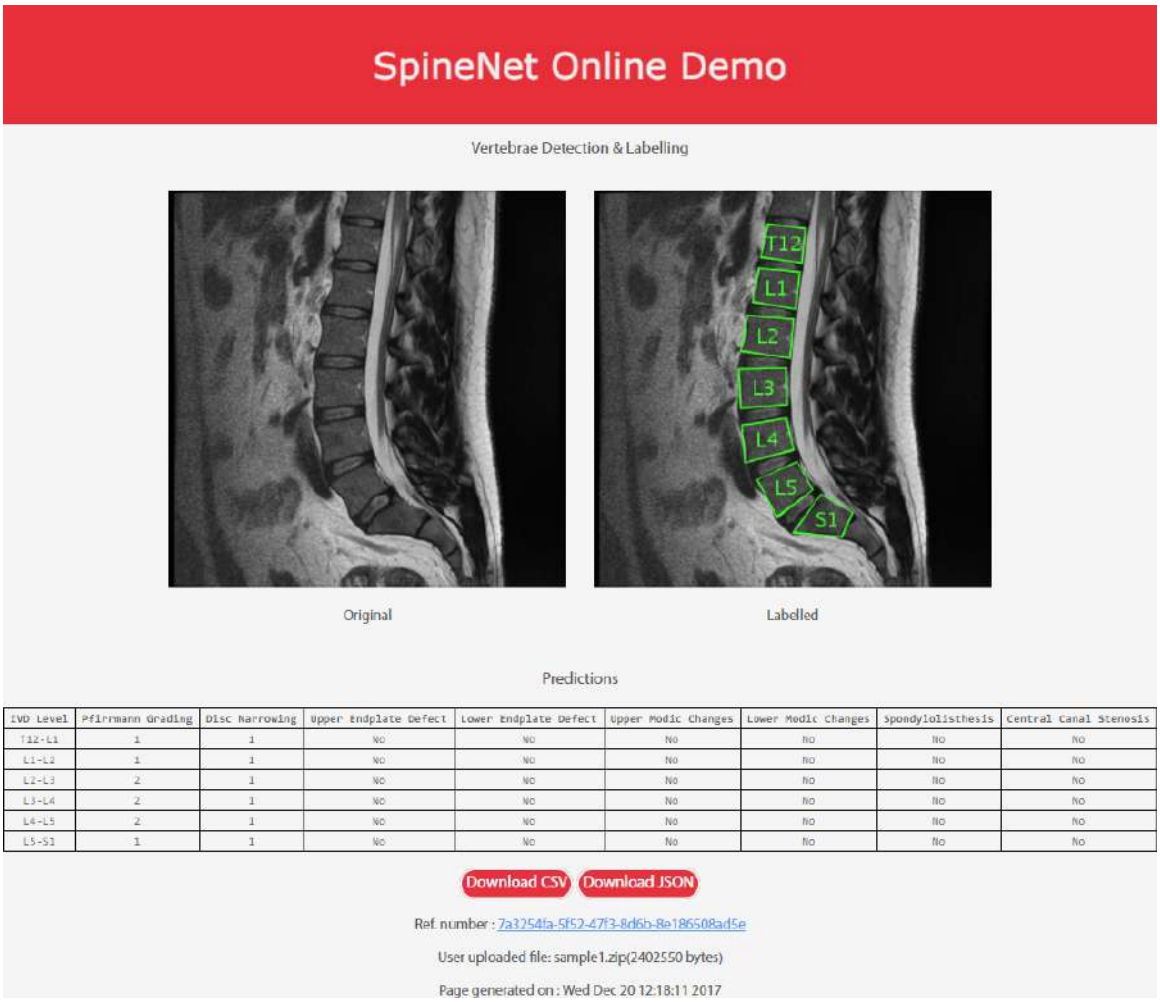
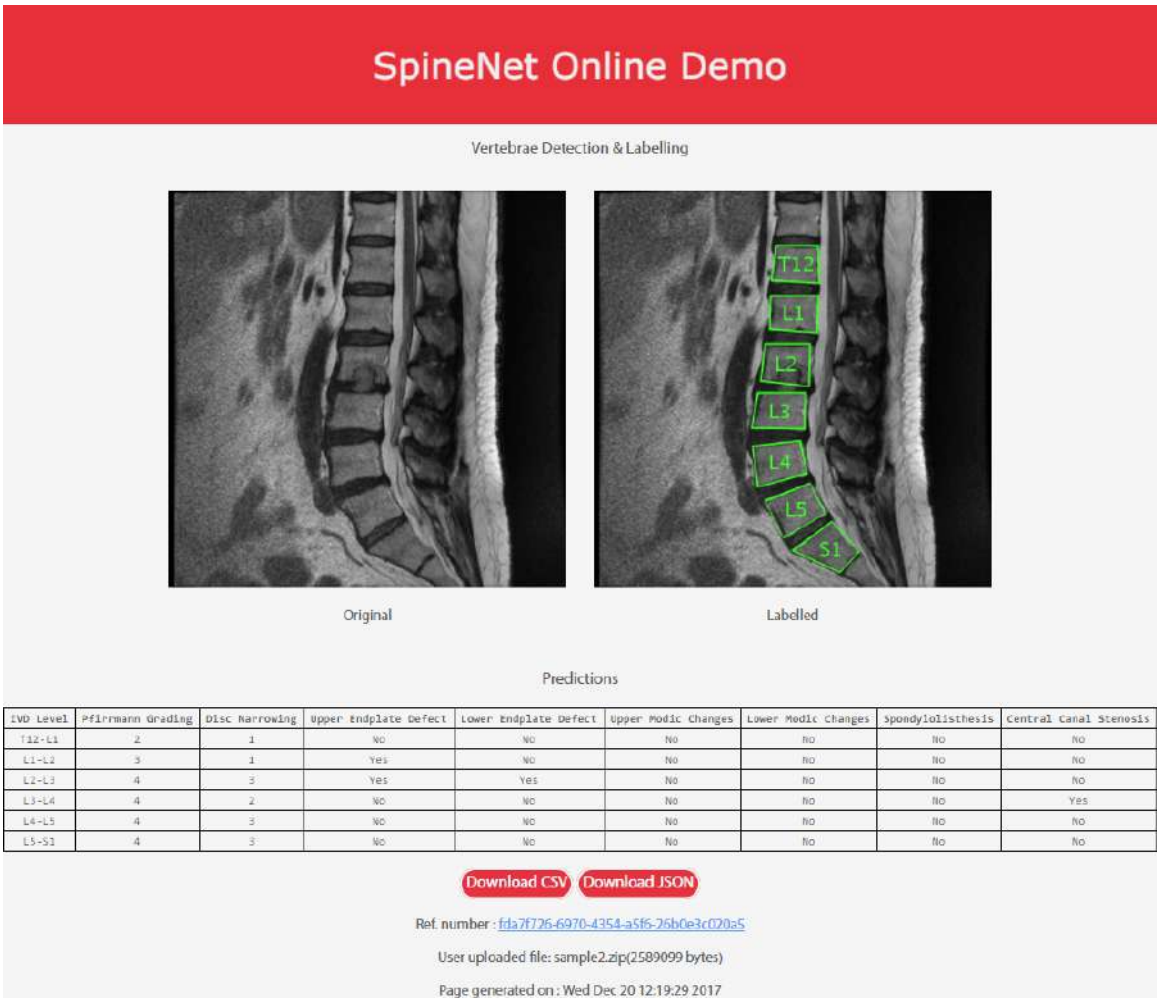


Figure 8.3: ***SpineNet Online Demo Example 1.*** *Example output on a normal lumbar spine MRI. Case courtesy of A.Prof Frank Gaillard, Radiopaedia.org. From the case rID: 35543*



IVD Level	Pfirrmann Grading	Disc Narrowing	Upper Endplate Defect	Lower Endplate Defect	Upper Modic Changes	Lower Modic Changes	Spondylolisthesis	Central Canal Stenosis
T12-L1	2	1	No	No	No	No	No	No
L1-L2	3	1	Yes	No	No	No	No	No
L2-L3	4	3	Yes	Yes	No	No	No	No
L3-L4	4	2	No	No	No	No	No	Yes
L4-L5	4	3	No	No	No	No	No	No
L5-S1	4	3	No	No	No	No	No	No

Figure 8.4: ***SpineNet Online Demo Example 2.*** *Example output on a lumbar spine MRI with an acute Schmorl node, part of the* ***Genodisc*** *definition of* ***endplate defects****, of the L2 vertebral body. Case courtesy of Dr Henry Knipe, Radiopaedia.org. From the case rID: 56636*

8.2.3 Future Works

Here, we discuss possible research directions continuing from this thesis.

Age. Age might play a big part in determining back pain as current radiological gradings are typically age independent but will probably benefit from age stratification. For example, a highly degenerated disc, Pfirrmann grade 5, in a young subject might be a bigger indicator of pain compared the same degenerated disc in an old subject.

Global Features. Depending only on disc features might be a limiting factor in our end goal of mapping MRIs to back pain. For example, Segar et al. (2016) suggested that BMI plays a big factor in back or leg pain which suggests that fat, which can be seen in the MRIs, might play some part to mapping back pain. Similar observations can be made from the geometry of the spine, where in abnormal cases present themselves as scoliosis, kyphosis, etc.

Serious Pathology. There are other measurements that can be made from a spinal MRI that were not looked at in this thesis which include serious pathology e.g. cancer, ankylosing spondylitis and several others. Extending the range of predictions to include more serious diseases might prove to be beneficial.

Bibliography

G. Abbati, S. Bauer, S. Winklhofer, P. J. Schüffler, U. Held, J. M. Burgstaller, J. Steurer, and J. M. Buhmann. *MRI-Based Surgical Planning for Lumbar Spinal Stenosis*, pages 116–124. Springer International Publishing, Cham, 2017. ISBN 978-3-319-66179-7. doi: 10.1007/978-3-319-66179-7_14. URL https://doi.org/10.1007/978-3-319-66179-7_14.

R. S. Alomari, J. J. Corso, V. Chaudhary, and G. Dhillon. Lumbar spine disc herniation diagnosis with a joint shape model. In J. Yao, T. Klinder, and S. Li, editors, *Computational Methods and Clinical Applications for Spine Imaging*, volume 17 of *Lecture Notes in Computational Vision and Biomechanics*, pages 87–98. Springer International Publishing, 2014. ISBN 978-3-319-07268-5. doi: 10.1007/978-3-319-07269-2_8. URL http://dx.doi.org/10.1007/978-3-319-07269-2_8.

M. Aslan, A. Ali, H. Rara, and A. Farag. An automated vertebra identification and segmentation in ct images. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 233–236, Sept 2010. doi: 10.1109/ICIP.2010.5651959.

V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

M. C. Battie, A. Lazary, J. Fairbank, S. Eisenstein, C. Heywood, M. Brayda-Bruno, P. P. Varga, and I. McCall. Disc degeneration-related clinical phenotypes. *Eur Spine J*, 23 Suppl 3:S305–314, Jun 2014.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380. URL `http://doi.acm.org/10.1145/1553374.1553380`.

W. Brinjikji, P. H. Luetmer, B. Comstock, B. W. Bresnahan, L. E. Chen, R. A. Deyo, S. Halabi, J. A. Turner, A. L. Avins, K. James, J. T. Wald, D. F. Kallmes, and J. G. Jarvik. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *AJNR Am J Neuroradiol*, 36(4):811–816, Apr 2015.

A. K. Burton, F. Balague, G. Cardon, H. R. Eriksen, Y. Henrotin, A. Lahad, A. Leclerc, G. Muller, and A. J. van der Beek. How to prevent low back pain. *Best Pract Res Clin Rheumatol*, 19(4):541–555, Aug 2005.

K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. 2011.

K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. 2014.

J. P. Cheung, H. Shigematsu, and K. M. Cheung. Verification of measurements of lumbar spinal dimensions in T1- and T2-weighted magnetic resonance imaging sequences. *Spine J*, 14(8):1476–1483, Aug 2014.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. 2005.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference*, 2017.

D. C. Cireşan, L. M. Gambardella, A. Giusti, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *In NIPS*, pages 2852–2860, 2012.

D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 411–418. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40762-8. doi: 10.1007/978-3-642-40763-5_51. URL `http://dx.doi.org/10.1007/978-3-642-40763-5_51`.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. 2015.

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, Feb. 2017. ISSN 0028-0836. URL `http://dx.doi.org/10.1038/nature21056`.

J. C. Fairbank and P. B. Pynsent. The Oswestry Disability Index. *Spine*, 25(22): 2940–2952, Nov 2000.

P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. 61 (1), 2005.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, Sept. 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.167. URL `http://dx.doi.org/10.1109/TPAMI.2009.167`.

S. Ghosh, R. S. Alomari, V. Chaudhary, and G. Dhillon. Computer-aided diagnosis for lumbar mri using heterogeneous classifiers. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1179–1182, March 2011. doi: 10.1109/ISBI.2011.5872612.

B. Glocker, J. Feulner, A. C., D. R. Haynor, and E. Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In *MICCAI 2012 - 15th International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, October 2012.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. 2010.

B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

R. Herzog, D. R. Elgort, A. E. Flanders, and P. J. Moley. Variability in diagnostic error rates of 10 mri centers performing lumbar spine mri examinations on the same patient within a 3-week period. *The Spine Journal*, 17(4):554–561. ISSN 1529-9430.

doi: 10.1016/j.spinee.2016.11.009. URL http://dx.doi.org/10.1016/j.spinee.2016.11.009.

J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. URL http://arxiv.org/abs/1709.01507.

S. H. Huang, Y. H. Chu, S. H. Lai, and C. L. Novak. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. *IEEE Trans Med Imaging*, 28(10):1595–1605, Oct 2009.

A. Jamaludin, T. Kadir, and A. Zisserman. Automatic modic changes classification in spinal mri. In *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*, 2015a.

A. Jamaludin, M. Lootus, T. Kadir, and A. Zisserman. Automatic modic changes classification in spinal mri. In *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*, 2015b.

A. Jamaludin, T. Kadir, and A. Zisserman. *SpineNet: Automatically Pinpointing Classification Evidence in Spinal MRIs*, pages 166–175. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46723-8. doi: 10.1007/978-3-319-46723-8_20. URL https://doi.org/10.1007/978-3-319-46723-8_20.

A. Jamaludin, T. Kadir, and A. Zisserman. *Self-supervised Learning for Spinal MRIs*, pages 294–302. Springer International Publishing, Cham, 2017a. ISBN 978-3-319-67558-9. doi: 10.1007/978-3-319-67558-9_34. URL https://doi.org/10.1007/978-3-319-67558-9_34.

A. Jamaludin, T. Kadir, and A. Zisserman. SpineNet: Automated classification and evidence visualization in spinal MRIs. *Medical Image Analysis*, 41 (Supplement C):63 – 73, 2017b. ISSN 1361-8415. doi: https://doi.org/10.1016/

j.media.2017.07.002. URL http://www.sciencedirect.com/science/article/pii/S136184151730110X. Special Issue on the 2016 Conference on Medical Image Computing and Computer Assisted Intervention (Analog to MICCAI 2015).

A. Jamaludin, M. Lootus, T. Kadir, A. Zisserman, J. Urban, M. C. Battié, J. Fairbank, and I. McCall. Issls prize in bioengineering science 2017: Automation of reading of radiological features from magnetic resonance images (mris) of the lumbar spine without human intervention is comparable with an expert radiologist. *European Spine Journal*, 26(5):1374–1383, May 2017c. ISSN 1432-0932. doi: 10.1007/s00586-017-4956-3. URL https://doi.org/10.1007/s00586-017-4956-3.

J. Koh, V. Chaudhary, and G. Dhillon. Disc herniation diagnosis in mri using a cad framework and a two-level classifier. *International Journal of Computer Assisted Radiology and Surgery*, 7(6):861–869, 2012. ISSN 1861-6410. doi: 10.1007/s11548-012-0674-9. URL http://dx.doi.org/10.1007/s11548-012-0674-9.

I. Kokkinos. Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *CoRR*, abs/1609.02132, 2016. URL http://arxiv.org/abs/1609.02132.

S. Koompairojn, K. Hua, K. A. Hua, and J. Srisomboon. Computer-aided diagnosis of lumbar stenosis conditions. In *SPIE*, 2010.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. 2012.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, Nov. 2015.

M. Lootus. *Automated Radiological Analysis of Spinal MRI*. DPhil thesis, University of Oxford, 2015.

M. Lootus, T. Kadir, and A. Zisserman. Vertebrae detection and labelling in lumbar mr images. In *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*, 2013.

M. Lootus, T. Kadir, and A. Zisserman. Radiological grading of spinal MRI. In *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*, 2014.

A. J. MacGregor, T. Andrew, P. N. Sambrook, and T. D. Spector. Structural, psychological, and genetic influences on low back and neck pain: a study of adult female twins. *Arthritis Rheum.*, 51(2):160–167, Apr 2004.

A. Mahendran and A. Vedaldi. Salient deconvolutional networks. In *European Conference on Computer Vision*, 2016.

H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *ICML*, 2009.

M. T. Modic, P. M. Steinberg, J. S. Ross, T. J. Masaryk, and J. R. Carter. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. *Radiology*, 166:193–199, 1988.

P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. IÅ¡gum. Deep learning for multi-task medical image segmentation in multiple modalities. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, pages 478–486. Springer International Publishing, Cham, 2016. URL `http://dx.doi.org/10.1007/978-3-319-46723-8_55`.

B. Moffit, M. Reicher, R. Lufkin, and J. Bentson. Comparison of T1 and T2 weighted images of the lumbar spine. *Comput Med Imaging Graph*, 12(5):271–276, 1988.

A. B. Oktay and Y. S. Akgul. Simultaneous localization of lumbar vertebrae and intervertebral discs with svm-based MRF. *IEEE Trans. Biomed. Engineering*, 60 (9):2375–2383, 2013. doi: 10.1109/TBME.2013.2256460. URL `http://dx.doi.org/10.1109/TBME.2013.2256460`.

K. T. Palmer, K. Walsh, H. Bendall, C. Cooper, and D. Coggon. Back pain in britain: comparison of two prevalence surveys at an interval of 10 years. *BMJ*, 320(7249): 1577–1578, 2000. ISSN 0959-8138. doi: 10.1136/bmj.320.7249.1577.

S. Peleg, G. Dar, B. Medlej, N. Steinberg, Y. Masharawi, B. Latimer, L. Jellema, N. Peled, B. Arensburg, and I. Hershkovitz. Orientation of the human sacrum: Anthropological perspectives and methodological approaches. *American Journal of Physical Anthropology*, 133(3):967–977, 2007. ISSN 1096-8644. doi: 10.1002/ajpa.20599. URL `http://dx.doi.org/10.1002/ajpa.20599`.

C. W. Pfirrmann, A. Metzdorf, M. Zanetti, J. Hodler, and N. Boos. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine*, 26(17): 1873–1878, Sep 2001.

O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. Technical Report 1505.04597, arXiv, 2015. URL `http://lmb.informatik.uni-freiburg.de//Publications/2015/RFB15`. arXiv:1505.04597 [cs.CV].

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages

1–42, 2015. ISSN 0920-5691. doi: 10.1007/s11263-015-0816-y. URL http://dx.doi.org/10.1007/s11263-015-0816-y.

P. Savigny, S. Kuntze, P. Watson, M. Underwood, G. Ritchie, M. Cotterell, D. Hill, N. Browne, E. Buchanan, and P. Coffey. Low back pain: early management of persistent non-specific low back pain. *London: National Collaborating Centre for Primary Care and Royal College of General Practitioners*, 14, 2009.

S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, and C. Schnorr. Spine detection and labeling using a parts-based graphical model. *Inf Process Med Imaging*, 20:122–133, 2007.

A. H. Segar, J. P. Urban, J. C. Fairbank, and A. Judge. The Association Between Body Mass Index (BMI) and Back or Leg Pain in Patients With Spinal Conditions: Results from the Genodisc Study. *Spine*, 41(20):E1237–E1243, Oct 2016.

J. Shen, S. Parent, and S. Kadoury. Classification of spinal deformities using a parametric torsion estimator. In J. Yao, T. Klinder, and S. Li, editors, *Computational Methods and Clinical Applications for Spine Imaging*, volume 17 of *Lecture Notes in Computational Vision and Biomechanics*, pages 75–86. Springer International Publishing, 2014. ISBN 978-3-319-07268-5. doi: 10.1007/978-3-319-07269-2_7. URL http://dx.doi.org/10.1007/978-3-319-07269-2_7.

H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2528162.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In *Workshop at International Conference on Learning Representations*, 2015.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR 2015*, 2015. URL `http://arxiv.org/abs/1409.4842`.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.510. URL `http://dx.doi.org/10.1109/ICCV.2015.510`.

A. Vedaldi and K. Lenc. MatConvNet – Convolutional Neural Networks for MATLAB. *CoRR*, abs/1412.4564, 2014a.

A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, abs/1412.4564, 2014b.

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.

L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, April 2002. ISSN 0899-7667. doi: 10.1162/089976602317318938.

X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. 2013.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

M. D. Zeiler and R. Fergus. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, chapter Visualizing and Understanding Convolutional Networks, pages 818–833. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53. URL `http://dx.doi.org/10.1007/978-3-319-10590-1_53`.

Y. Zhan, D. Maneesh, M. Harder, and X. S. Zhou. Robust MR spine detection using hierarchical learning and local articulated model. *Med Image Comput Comput Assist Interv*, 15(Pt 1):141–148, 2012.

J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. *Top-Down Neural Attention by Excitation Backprop*, pages 543–559. Springer International Publishing, 2016. ISBN 978-3-319-46493-0. doi: 10.1007/978-3-319-46493-0_33. URL `http://dx.doi.org/10.1007/978-3-319-46493-0_33`.

Q. Zhang, A. Bhalerao, and C. Hutchinson. *Weakly-Supervised Evidence Pinpointing and Description*, pages 210–222. Springer International Publishing, Cham, 2017. ISBN 978-3-319-59050-9. doi: 10.1007/978-3-319-59050-9_17. URL `http://dx.doi.org/10.1007/978-3-319-59050-9_17`.

B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.