Computer aided analysis of inflammatory muscle disease using magnetic resonance imaging

by

James V Jack

A Doctoral Thesis

Submitted in partial fulfilment
of the requirements for the award of

Doctor of Philosophy
of
Loughborough University

25th September 2015

Abstract

Inflammatory muscle disease (myositis) is characterised by inflammation and a gradual increase in muscle weakness. Diagnosis typically requires a range of clinical tests, including magnetic resonance imaging of the thigh muscles to assess the disease severity. In the past, this has been measured by manually counting the number of muscles affected.

In this work, a computer-aided analysis of inflammatory muscle disease is presented to help doctors diagnose and monitor the disease. Methods to quantify the level of oedema and fat infiltration from magnetic resonance scans are proposed and the disease quantities determined are shown to have positive correlation against expert medical opinion. The methods have been designed and tested on a database of clinically acquired T1 and STIR sequences, and are proven to be robust despite suboptimal image quality.

General background information is first introduced, giving an overview of the medical, technical, and theoretical topics necessary to understand the problem domain. Next, a detailed introduction to the physics of magnetic resonance imaging is given. A review of important literature from similar and related domains is presented, with valuable insights that are utilised at a later stage. Scans are carefully pre-processed to bring all slices in to a common frame of reference and the methods to quantify the level of oedema and fat infiltration are defined and shown to have good positive correlation with expert medical opinion. A number of validation tests are performed with re-scanned subjects to indicate the level of repeatability. The disease quantities, together with statistical features from the T1-STIR joint histogram, are used for automatic classification of the disease severity. Automatic classification is shown to be successful on out of sample data for both the oedema and fat infiltration problems.

— J. Jack, September 2015.

Acknowledgements

I would like to thank Dr. Helmut Bez and Prof. Eran Edirisinghe for their many useful thoughts and discussions, but also their support at meetings held at the Computer Science department and away in Coventry.

I must also express my gratitude to Prof. Charles Hutchinson and Dr. Terry Jones from the Warwick Medical School at the University of Warwick, who have provided the medical expertise and background information necessary for cross-disciplinary work.

Thank you also to Dr. Mashhuda Glencross for providing early feedback.

Thank you to the Computer Science department and Loughborough University for providing my scholarship and thereby enabling this research. Loughborough University has been a wonderful educational environment for the past eight years of my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The domain of computer-aided analysis in medicine can be split into different groups depending on the intended role.

Computer Aided Detection (CADe) takes the approach of a "second reader", which is essentially a second opinion, and uses visual markers to direct the attention of the doctor or radiologist. For instance, in Computed Tomography (CT) colonography, CADe for polyp detection may function by ranking candidate regions alongside the 3D rendering. The doctor decides whether the highlighted regions need to be treated. CADe systems are of particular interest for investigations that require a good deal of concentration or many small decisions, for instance finding lung nodules and deciding if they are cancerous or benign.

Computer Aided Diagnosis (CADx) goes one step further and aims to take all of the data available and output a diagnosis verdict. For instance, CADx systems have been applied to differentiate benign and cancerous tumours in mammography using (e.g.) texture features. In this case the system makes use of data that is either difficult to perceive or would otherwise require many years of experience. Systems may use non-imaging data as well, e.g. blood serum levels, age, sex.

In the literature the "CAD" term has a broad meaning and can refer to either type of system. Most systems are designed for one disease only, eliminating the need for differential diagnosis within the system itself. The doctor is ultimately responsible for the overall decision, treatment, and progress review.

CAD has been applied extensively in some areas, for instance in the detection of tumours and brain disorders. However, computer-aided approaches for the detection of inflammatory muscle disease have received little attention. In the case of myositis, there are only a few studies that detect or quantify the level of disease from medical imaging investigation.

Doctors working with the Warwick Medical School in Coventry were interested in using computer-aided analysis of their medical imaging investigations to help make diagnosis and treatment decisions. This was primarily for the Idiopathic

Inflammatory Myopathies (IIM), a group of diseases which are responsible for inflammation build up in the muscles. Within this group are many individual diseases with somewhat different presentations, including polymyositis, dermatomyositis, and inclusion-body myositis. Some of these have been linked with a higher probability of malignancy.

The assessment of IIM could include a wide range of tests: bloods, muscle strength test, electromyogram, patient history, biopsy, physical examination, and Magnetic Resonance Imaging (MRI). All of the tests have some degree of variability: either in the way it is performed, or assessed. Muscle strength tests, for instance, are difficult to perform reliably and requires patient compliance. For blood tests, the Creatine Kinase (CK) level is known to be raised in those with the disease, but it is also known to have normal levels in some histopathologically-confirmed cases. That is to say, even when confirmed by microscope, patients with the disease may have normal levels of CK. Therefore, for some tests, the result is an indication but not a certainty. Overall, the doctor is looking for a pattern of results that may indicate the disease, once other diseases have been ruled out.

When considering an IIM disease, MRI is always performed. Inflammation (oedema) in the skeletal muscle can be seen and significant muscle damage may be assessed. The doctor visually inspects each region of the leg on each slice. Oedema must be identified as distinct from blood vessels and imaging errors. Long-term muscle damage is also assessed by inspecting the amount of fat within the muscles, which must be distinct from normal fat within the muscle. The current standard way of assessing the severity of the disease is to count the number of different muscles affected, as seen on transverse slices of the thigh. The doctors were interested in using a computer-aided approach to get a more thorough quantitative analysis of the muscle disease using MRI. Even for an expert with decades of clinical experience, it is sometimes difficult to assess the severity of disease, particularly when it is not severe. The hypothesis was that MRI contained enough information alone (without other test results) to quantify and possibly diagnose the muscle disease. By using a computer-aided approach, the variability of MRI assessment may be reduced.

Treatment for IIM typically includes corticosteroids to reduce the level of inflammation, which also reduces further damage to the muscles. However, the duration of treatment is difficult to determine and may depend on the treatment response; patients may be re-scanned in order to assess changes in disease state. If treatment is stopped too early, relapse of inflammation may occur. On the other hand, the long-term use of steroids has serious side-effects. Therefore, a CAD system that can monitor changes in disease state would have significant clinical benefit.

MRI machines work by sampling the magnetic resonance "echo" that is generated from atoms in the body when they are placed inside a magnetic field and a Radio Frequency (RF) pulse has been applied (this process is explained in detail in section 2.5). MRI is capable of producing very high Signal to Noise Ratio (SNR) images but typically, in clinical situations, machine parameters must be used that attain somewhat lower SNR images. This is due to the length of time the patient must remain inside the scanner and be still. Elderly patients, or those with other conditions, mean that image quality is sometimes poor. Some MRI sequences are more prone to noise and imaging artefacts than others. Although some minimum level of detail is required, any method of practical, clinical use must be robust against poor quality scans.

With respect to the above, the aim of the research presented within the context of this thesis is to provide a computer-aided analysis of inflammatory muscle disease using MRI to help doctors in diagnosis and monitoring. The objectives are to:

- Quantify the level of oedema and fat infiltration.
- Be robust despite suboptimal image quality.
- Measure changes in disease state over time.
- Be as automatic as possible.

The level of disease must be expressed as a decimal number which reflects the total inflammation which can be seen in the MRI scan (c.f. an integer count of affected muscles). Since there is no absolute ground truth as to the level of disease, any disease metric constructed must have positive correlation with the expert opinion of the doctor.

The system is required to work with MRI scans of a quality that would be obtainable in a clinical setting. Although it is impossible to use images with significant imaging artefacts, or those where the required protocol has not been met, the system must be robust against poor SNR.

The detection and quantification of disease over a range of severities would have clear clinical benefit, as would the ability to monitor the disease state over time. It must do this with as little manual intervention as possible.

The current method to diagnose IIM requires a wide range of clinical tests, which all have some degree of variation. The provision of a quantitative method that can be applied to clinically routine scans is expected to reduce the overall variation in the doctor's diagnosis process, and may also help in disease monitoring. This would improve the level of care provided to patients with IIM. With

a reduction in the variation of assessment, it is hoped that the number of clinical tests required in the majority of cases will be reduced, saving time, money, and lessening the strain put on the patient. The above observations provide the motivation for the research presented within this thesis.

This thesis presents a number of contributions that originated from the research conducted. They are:

- Semi-automatic methods to quantify oedema and fat infiltration for IIM using MRI scans of the thigh which correlate with expert medical opinion.
- Visualisation of the expected presentation of disease state in the joint intensity space for IIM.
- Design of feature vectors for the classification of disease state for IIM using the joint intensity space.

In chapter 1 the problem domain has been briefly outlined. In chapter 2 the necessary background is introduced, including a brief overview of the physics of MRI acquisition. Chapter 3 presents a literature review of related research and prerequisite topics, including the appropriate processing of MRI data. Chapter 4 details the pre-processing steps that are necessary to bring different patient scans into a common frame of reference. The joint intensity space is also introduced. In chapter 5 the database of MRI scans is used to show that the appearance of the joint intensity space correlates with the disease state. This identification, together with thorough data analysis, leads to the development of two disease metrics. They are shown to have positive correlation with expert medical opinion. Chapter 6 builds on the results of the previous chapter with the construction of feature vectors for automatic classification of MRI scans. An out of sample database is used to evaluate the results and the use of some selected feature spaces are examined. Chapter 7 concludes with what has been achieved and discusses the direction for future research in this domain.

Chapter 2

Background

2.1 What is myositis?

Myositis is a type of muscle disease (myopathy) that is characterised by inflammation and progressive muscle weakness. It is part of the Idiopathic Inflammatory Myopathies (IIM) group. It is systemic, meaning that it affects the whole body, not just the muscles.

There are different types of myositis, for instance polymyositis, dermatomyositis, and inclusion-body myositis. The exact presentation tends to be different for each disease but many of the symptoms and biomarkers overlap [14].

Estimates on the prevalence of myositis vary but it is approximately 8 in every 100 000 people in the general population [49]. There is no cure for the disease but it is possible to reduce the level of inflammation by taking steroids in most cases.

How myositis starts is not understood, but it is thought that a combination of environmental and immunological factors, with genetic predisposition, are in some way responsible. Initial onset may develop after an infection, where the patient gradually feels weaker over weeks, months, or years. Age of onset varies depending on the exact disease.

The inflammation is an autoimmune response from the body and the subsequent muscle damage causes cells to break open, releasing their content into the extracellular fluid. Substances normally only found inside cells may now be detectable in blood tests (e.g. aldolase, CK, aspartate aminotransferase), however these substances are non-specific to myositis. As part of a normal healing process, the damaged muscle cells are replaced by fat, which also makes the muscles weaker. Myositis may be described as *acute* and *chronic*. Acute describes the disease as active with inflammation. Chronic describes the disease as fat which has infiltrated the muscle (infiltration). The terms are independent so the disease could be chronic and acute concurrently.

2.2 Types of MRI image

MRI machines normally produce images based on the magnetic resonance of hydrogen protons that are present in the material. Since human biology is largely made up of water (H_2O), this allows MRI to produce high quality images of tissues in the body. It is also possible to produce more than one sort of image, depending on the *pulse sequence* used to acquire the samples. Essentially, the type of pulse sequence enhances or suppresses the signal from different types of material as a result of their physical properties (e.g. [8, pp. 243–245]).

To use an analogy, this is similar to using a traditional film-based camera and exposing with different types of film, for instance greyscale and infra-red film. It is not the same as changing the exposure settings, since this only alters the intensity response.

In reality, the MRI machine transmits a Radio Frequency (RF) pulse through the material and then measures an echo some time later. This is explained in more detail in section 2.5.

2.2.1 T1

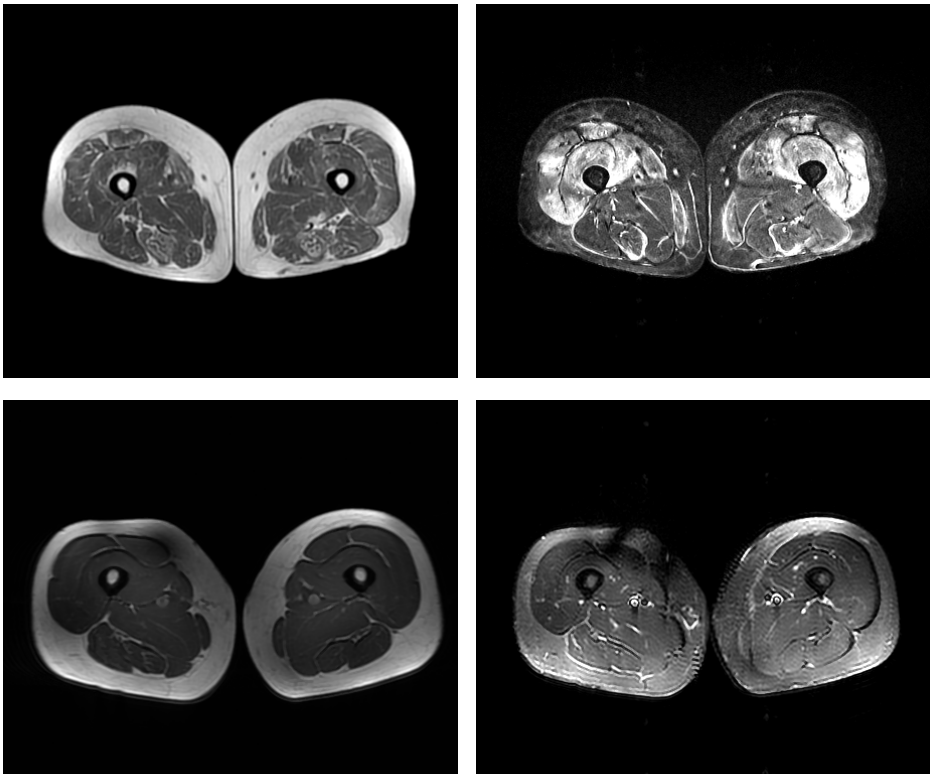
The T1 pulse sequence measure the longitudinal relaxation of the magnetic moments [72, p. 7]. On a typical T1 image the fat and bone-marrow will appear bright (hyperintense), water and muscle have an intermediate appearance, and bone is dark (hypointense). An example of T1 scans are shown in Figure 2.1.

The quality of T1 images can be very high, although some scans investigated have still suffered from noise.

2.2.2 STIR

The STIR sequence suppresses the signal from fat and is useful for imaging inflammation because it generates a high signal [8, p. 244]. On a typical STIR image, blood vessels and muscle surrounded by inflammation appear very bright, healthy muscle has an intermediate appearance, and fat is dark or black. An example of STIR scans are shown in Figure 2.1.

In contrast to T1 images, STIR images are likely to have low SNR and are more susceptible to imaging artefacts such as ringing. It is possible to reduce the noise by increasing the number of excitations used in the STIR pulse sequence, however the length of time spent in the scanner must be reasonable for the patient and so a compromise is made on the SNR.



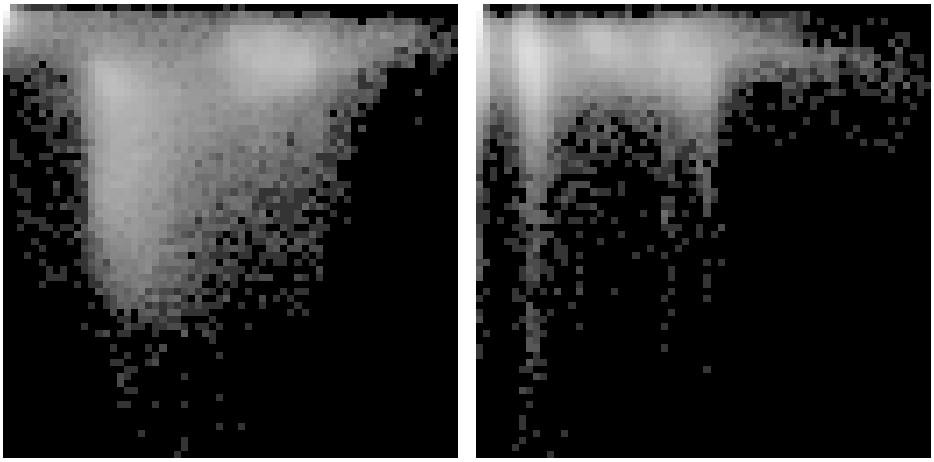
a b
c d

Figure 2.1: (a) and (b) show T1 and STIR scans of a myositis-positive patient. Both are taken at the same location in the thigh. (c) and (d) are T1 and STIR scans from a (healthy) control case. Note how diseased muscle in (b) appears brighter than fat. In comparison, healthy muscle in (d) does not appear brighter than the fat layer. In the T1 images (a) and (c), fat is always brighter than the muscle. Note the presence of imaging artefacts, seen as light and dark patches, in (a), (c) and (d).

2.3 DICOM format

Medical images are recorded in Digital Imaging and Communications in Medicine (DICOM) format and pixel data are normally stored as a 16-bit single channel. Due to differences in hardware, pulse sequence, and biology, there is no expected intensity range (which means that intensity values from one scanner cannot be compared to another without some form of intensity standardisation). STIR sequences typically have a smaller range than T1 sequences, but generally have more than 256 unique intensities.

DICOM images also contain a library of information that can be looked-up to help interpret and display the images. This metadata is recorded at the point of acquisition. For instance, the information might relate to the MRI settings that were used to acquire the images, the spatial resolution of each pixel, the time of acquisition, patient details, slice location, etc. Note that, for this research, personally identifiable data had been removed beforehand.



a b

Figure 2.2: Joint histogram examples for (a) myositis-positive and (b) healthy patients. Both have 64 bins in each dimension. The horizontal axis, from left to right, represents increasing T1 values. The vertical axis, from top to bottom, represents increasing STIR values. No ROI has been applied.

2.4 T1 and STIR joint histogram

T1 and STIR images may be combined to generate a scatter plot and the resulting clusters are thought to have diagnostic use [34]. When the scatter plot data is summed into bins it may be called a joint histogram, or bivariate histogram (see Figure 2.2).

A joint histogram can be generated by visiting pixels from the two images in lock-step. Their joint intensity forms a pair (s, t), where s is the STIR intensity and t is the T1 intensity. The joint histogram is then a tally of occurrences for all such pairs. It is assumed the images are discretised prior to generating the joint histogram. Due to the number of background pixels which have very low intensity, the joint histogram typically has a very large concentration in the top-left corner which can prove problematic when visualising the data. The example in Figure 2.2 has been log-transformed. In the case where a Region Of Interest (ROI) has been applied, the joint histogram may be generated using only the pixels in the ROI, generally revealing better structure.

Interpretation of the joint histogram is best done visually. Area D in Figure 2.3 is the approximate area of the histogram where diseased muscle is expected to appear, characterised by intermediate T1 signal (water and muscle) and high STIR signal (blood vessels and diseased muscle). Anatomical locations for the thigh are shown in Figure 2.4. The high water concentration of diseased muscle leads to this appearance in the joint histogram.

The appearance of joint histograms can vary considerably, even among healthy subjects, owing to differences in the amount of fat, muscle and water present in

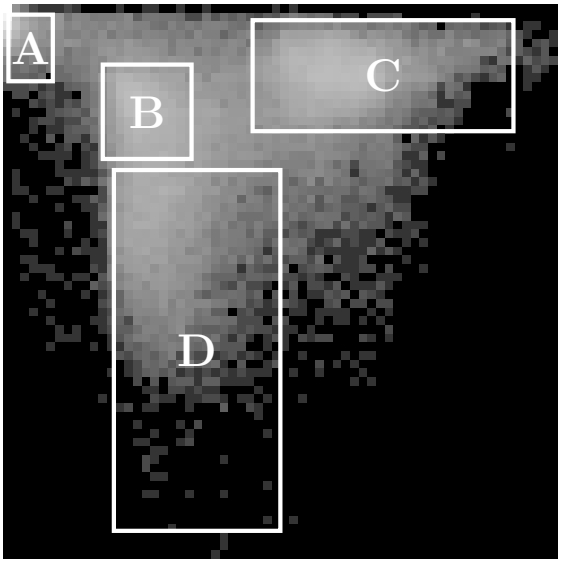


Figure 2.3: Approximate areas of the joint histogram correlating to types of tissue. (A) is the femur bone. (B) is normal muscle. (C) is fat, fascia, and bone marrow. (D) is diseased muscle. Note that the top-left bin has high density due to the large number of low intensity background pixels in both T1 and STIR images; an ROI has not been applied.

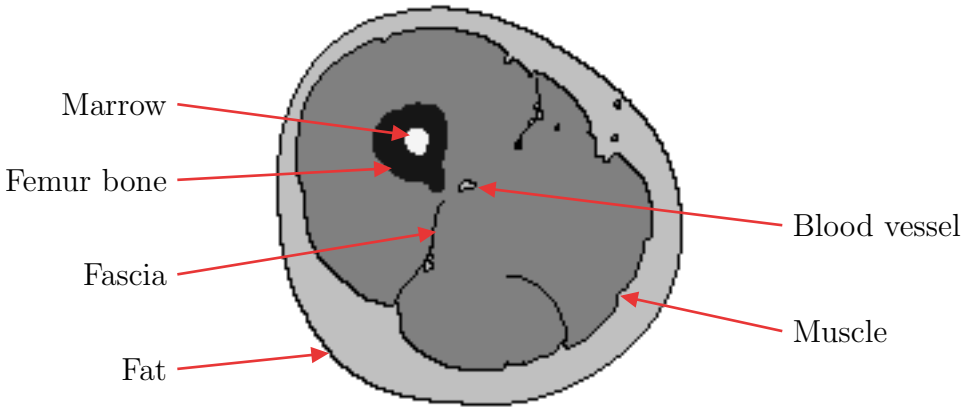


Figure 2.4: Basic anatomy of the (right) thigh in the transverse plane. The view is pointing towards the pelvis. On the MRI scans presented here, it would appear on the left-hand side. For more detailed anatomy see [16, pp. 212–214].

the thigh. Without some method of intensity standardisation, it is not generally possible to compare histograms between patients because the images are not calibrated such that a certain tissue always produces the same signal intensity. Magnetic Resonance (MR) scans performed on the same scanner may be comparable, although the variation in magnetic field strength may be confounding.

2.5 An introduction to the physics of MRI

A thorough understanding of how MRI works is outside the scope of this document. However, a brief overview is given here so that a general understanding may be gained. For detailed explanations please see [59; 28; 26; 8; 72; 29; 52].

The human body is made up of approximately 60% water (dihydrogen monoxide molecules) and therefore has an abundance of hydrogen atoms. Hydrogen atoms are useful in MRI because they contain only one positively charged proton which, depending on its direction of spin, generates a magnetic dipole moment, north or south (Figure 2.5). Without the presence of an external magnetic field, the magnetic dipole moments of hydrogen are arranged randomly and their net magnetic field is zero.

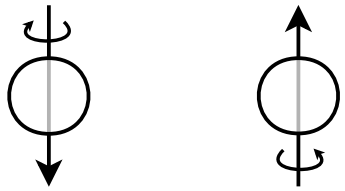


Figure 2.5: A hydrogen atom has a net positive charge and generates a small electric field depending on the direction of spin. The vertical arrows represent the magnetic dipole moment.

When placed inside the MRI machine, the presence of a strong, fixed magnetic field (denoted B_0) causes the dipoles to become aligned with the magnetic field, however their net magnetic field is still zero, since approximately half are pointing north. Gradually over time, a small fraction of the magnetic dipole moments flip to point in the same direction as B_0, causing their net magnetic field (denoted M_0) to increase exponentially, up to some asymptotic maximum (Figure 2.6).

The direction of B_0 is called the longitudinal direction (the Z axis), and two other axes (X and Y) perpendicular to B_0 define the transverse plane, with the Y axis pointing up. With no other influences, M_0 points in the same direction as B_0 (Figure 2.7).

Protons placed in the magnetic field B_0 also precess (wobble) around the axis of B_0 at a rate dictated by the Larmor equation, $\omega = \gamma B$, where γ is the gyromagnetic ratio (a physical property of the nucleus) and B is the strength of a magnetic field (Figure 2.8). For hydrogen, the gyromagnetic ratio γ is 42.6 MHz/T, which in a 1.5 Tesla MRI magnet gives a Larmor frequency of about 64 MHz.

If an oscillating Radio Frequency (RF) pulse is transmitted perpendicular to B_0, then the protons also precess around the direction of the RF wave (it is an electromagnetic wave, therefore it has a magnetic component). This only occurs when the frequency of the RF pulse matches the Larmor frequency ω, causing

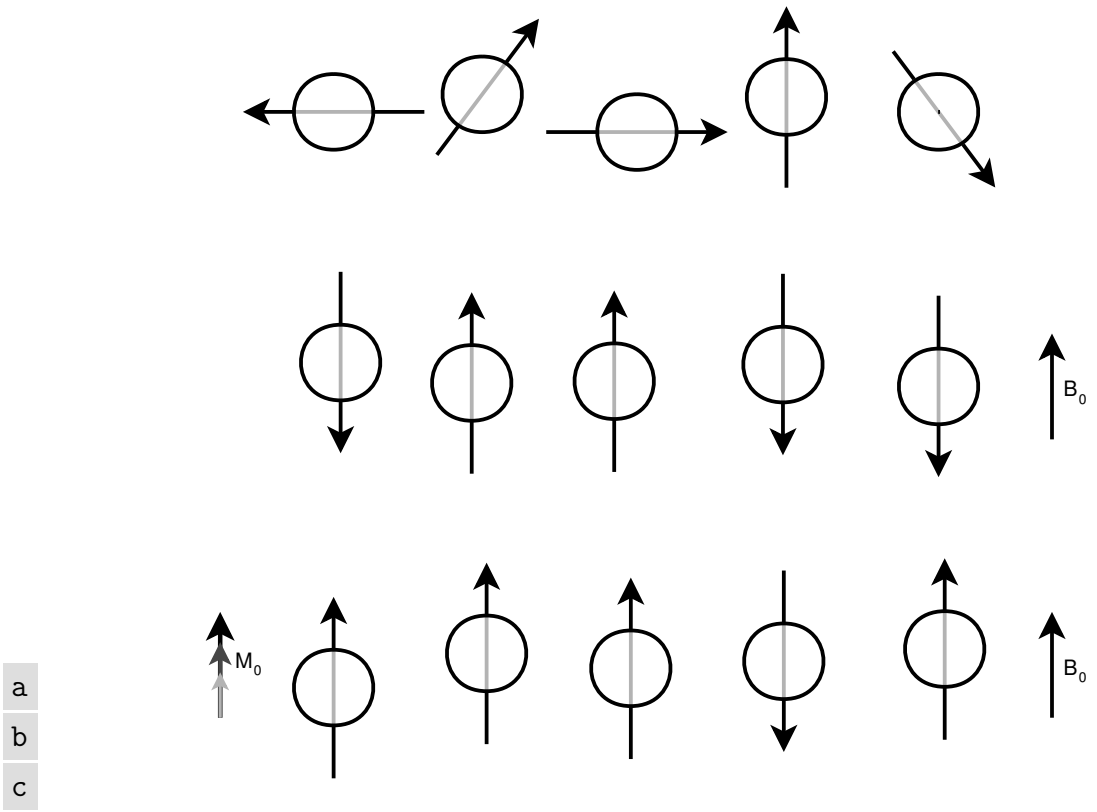


Figure 2.6: The effect of a magnetic field on the arrangement of magnetic dipole moments (hydrogen protons). (a) With no external field the dipoles point randomly. (b) When placed inside a static magnetic field (B_0) the dipoles align with the direction of the field, although their net magnetic field is still zero. (c) Over time, a small fraction of the dipoles flip to point in the same direction as the external field and their net magnetic field (M_0) increases up to some maximum.

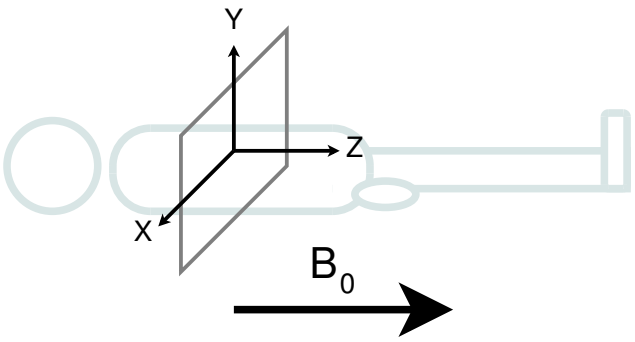


Figure 2.7: Arbitrary axes for a patient in an MRI scanner.

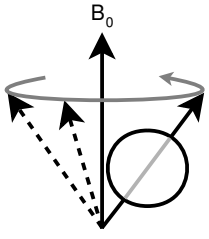


Figure 2.8: Precession of a hydrogen proton around the axis of B_0.

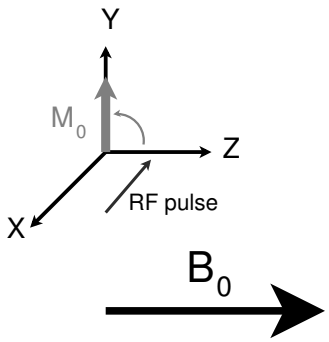


Figure 2.9: With a RF pulse, M_0 is flipped 90 degrees into the X-Y plane.

resonance. The effect is that M_0 is flipped into the transverse plane, said to be a 90 degree flip from the Z axis (Figure 2.9).

When the RF pulse is on and M_0 has flipped into the transverse plane, the protons have a higher energy state and are in phase with each other. Once the RF pulse is turned off, then over time the protons lose their excess energy (transferring it to the surrounding lattice of molecules) and will gradually align with B_0, eventually becoming out of phase again. This exponential process is called relaxation. T1 is thus called the longitudinal relaxation time, or spin-lattice relaxation time. It is worth noting that T2 measures the decay of magnetisation in the transverse plane.

A magnetic field gradient may be generated by using a pair of coils such that the field is (e.g.) north at one end of the machine and south at the other end of the machine, causing a sliding scale of magnetic field strength from one end to the other, with a centre point where the field is zero. One such gradient is denoted G_Z, which is in the same direction as B_0. This essentially causes B_0 to vary (e.g.) from 1.4 T to 1.6 T, which in turn means the rate of precession of hydrogen through the body varies from about 59.6 MHz to 68.2 MHz (Figure 2.10). Thus, the range of frequencies transmitted in the RF pulse can select which protons are flipped (both the position and thickness of the image slice).

The signal of M_0 is measured using RF coils, possibly the same coils which generated the RF pulse. Time to Echo (TE) is the amount of time to wait after the RF pulse before measuring the signal. Time to Repeat (TR) is the amount of time to wait between RF pulses.

Readout is the process of taking samples of M_0 over some small period of time. The readout signal represents the sum of signals from protons resonating throughout the whole slice (not just a single voxel).

To gain spatial information in one dimension, another magnetic field gradient G_X is applied along the X axis (perpendicular to B_0) during readout. For instance, for a 3×3 matrix of voxels in a slice, protons in the left column would have a slightly

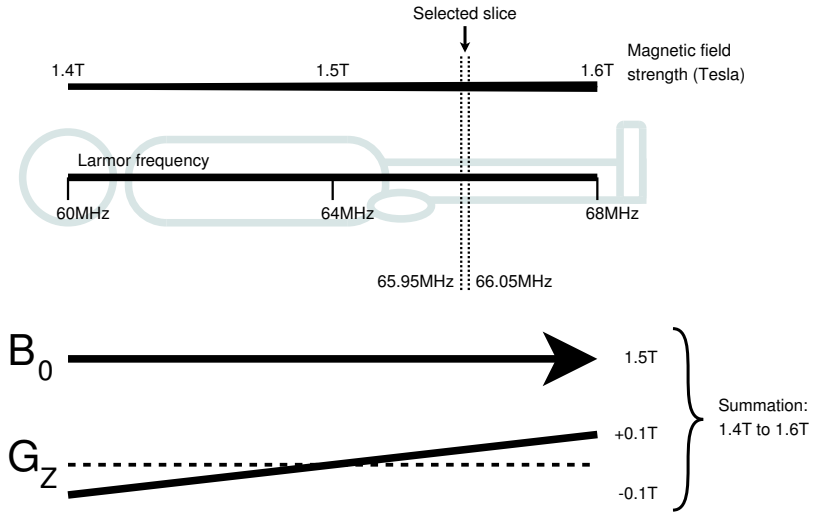


Figure 2.10: Slice selection is achieved by using the addition of a linear field gradient (generated by e.g. Maxwell coil pairs) to alter the precessional frequency of hydrogen through the body in the direction of the Z axis.

lower precessional frequency, the middle column would remain as before, and the right column would have a slightly higher precessional frequency. This is called frequency encoding. After M_0 is sampled during readout, a decomposition of the signal would reveal 3 different frequencies. Immediately after readout is complete G_X is switched off.

So far it has been possible to select hydrogen protons using G_Z and a RF pulse (for slice selection), and later using G_X during readout for frequency encoding (column information). It is not possible to just use frequency encoding in both X and Y directions at the same time, since this would mean many pixels would be represented by the same frequency. To isolate the final dimension requires a process called phase encoding, which must be repeated many times because the MRI machine is not able to measure phase directly (otherwise each pixel would be uniquely encoded by frequency and phase).

Phase encoding starts by using another magnetic field gradient G_Y, applied along the Y axis for a short period of time, just after the RF pulse, but *before* readout. At this point, all protons in the slice still have coherent phase. Consider the 3×3 matrix again. As the gradient G_Y is applied, the precessional frequency is higher for those protons in the top row voxels, remains unchanged in the middle row, and is lower for those in the bottom row. The change in frequency also means a change in phase (positive for higher frequency, negative for lower frequency). After G_Y is switched off, the frequency for all protons in the slice will again be the same because they experience the same magnetic field strength. However, the phase change that was induced in each row is retained, such that each row has a distinct phase. The length of time that G_Y is turned on controls the amount

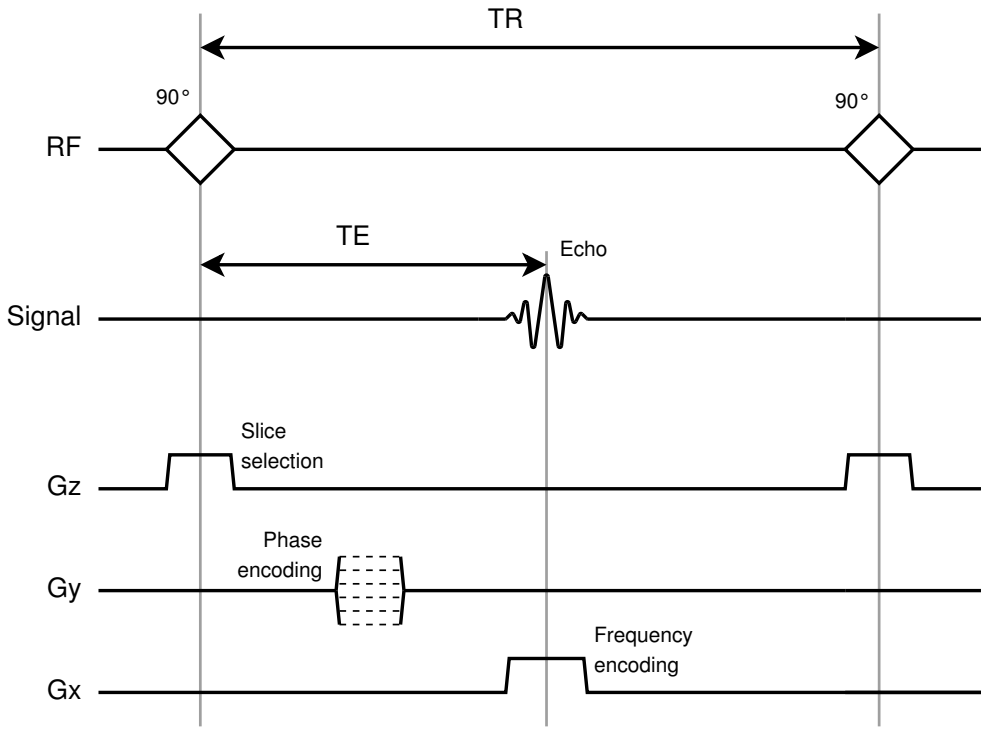


Figure 2.11: Example of an MRI pulse sequence diagram. Time goes from left to right.

of phase change between rows. The process of phase encoding can only be done once per TR and therefore the whole process must be repeated a number of times, equal to the number of rows in the matrix, each time using a *different* field gradient G_Y. Note that when stronger field gradients are used the overall signal amplitude becomes weaker due to increased dephasing effects.

The steps so far are: slice selection, phase encoding (rows), and frequency encoding (columns). A simplified example of an MRI pulse sequence digram is shown in Figure 2.11. The resulting precession of hydrogen protons means there is a changing magnetic field flux which can be measured by the receiver coils (a moving field induces an alternating current in the coil). The samples of this signal are stored in k-space, which is a complex matrix. During readout, the samples are stored in a single row of k-space. The process is repeated with a different phase encoding gradient and the samples are stored in the next row of k-space. Once every row in k-space has been filled, all the necessary data for the final image has been collected.

Going from left to right in k-space represents frequency. Going from top to bottom in k-space also represents frequency because the change in phase encoding gradient (G_Y) at each step is equivalent to a change in frequency. Therefore it is possible to use a two dimensional Fourier Transform (FT) to get real and imaginary images from the spatial frequency domain.

Normally, the image received from the MRI scanner is a magnitude image, $m = \sqrt{r^2 + i^2}$, where r is the real component, i is the imaginary component, and m is the final image.

The MRI machine may perform other pre- or post-processing steps. For instance, signals picked up at the receiver coils are subject to the inverse square law, and therefore signals from objects closer to the receiver coils are stronger. The MRI machine adjusts for this effect automatically.

2.6 Scanning protocol

Slices are taken in the transverse plane, starting from the pelvis and moving to just above the knee (Figure 2.12). Since myositis is systemic, the disease could be visible in any muscle, however a large bulk of muscle is preferred because the associated inflammation can appear in patches. Therefore slices taken close together are less useful than those spread apart. Since acquisition is done in a clinical setting, the exact position will be determined by an experienced radiologist.

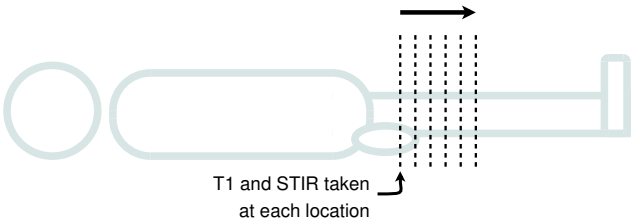


Figure 2.12: Approximate slice locations on a patient.

Slice thickness was 8mm with a 2mm gap between images. For all databases a set of 6 contiguous slice pairs were chosen to start just below symphysis pubis and ended before the top of the knee. The Field of View (FOV) was typically 40cm×40cm. Pixel spacing was between 0.74mm and 1.04mm.

2.7 Cohorts and acquisition

The term *image pair* (or *slice pair*) will refer to one T1 image and one STIR image that have been taken in the same slice location, for the same patient, at approximately the same time. A *case* will refer to a set of image pairs over some contiguous region of the thigh. Scans will normally have between 6 and 12 image pairs. Many slice locations are sampled because myositis can be "patchy" and therefore it may not be visible on all slices. Indeed, if a biopsy is to be taken, the appropriate region may be located on MRI first.

The cases in each database (detailed below) have exactly 6 slices. Acquisitions came from 3 different 1.5 Tesla MRI scanners: Optima MR450w, Optima MR360, and Gyroscan Intera.

2.7.1 Training database

As part of a clinical trial, 12 symptomatic (volunteer) patients and 9 asymptomatic controls were scanned from 2013 to 2015. These scans followed a strict acquisition protocol.

A small historical database of 10 patients was also utilised, taken between 2001 and 2005. The acquisition protocol for these scans were similar to the current protocol, although there was a greater range of TR and TE times. The essential aspects were the same: slice locations for T1 and STIR were aligned and a contiguous set of image pairs were manually selected from the same body region as newer scans.

The variation in protocols were viewed as a way to ensure that the methods employed were robust against small changes in acquisition parameters. It should be noted that different MRI machines are not always capable of exactly the same parameters. The image quality is not just dictated by the sequence parameters but also the individual scanner.

2.7.2 Out of sample database

A further 12 symptomatic patients were scanned in 2015 and were used as an out of sample database. These followed acquisition protocol. These scans were not made available until near completion.

2.7.3 Re-scan database

In some instances, a subject was scanned more than once. Re-scans occurred between 2013 and 2015. In summary: 1 control subject and 1 symptomatic patient were rescanned during the same session. 2 symptomatic patients were scanned 3 times on different occasions. 1 symptomatic patient was scanned on 2 different occasions.

2.7.4 Expert grading

Scans in the training and out of sample databases where graded by a medical expert. One grade measures the severity of acute disease (oedema) while the other grade measures chronic disease severity (infiltration). Table 2.1 shows the ordinal scale used for both severities.

Grade	Ordinal value	Meaning
Normal	0	No current disease
Mild	1	Diseased
Moderate	2	Diseased
Severe	3	Diseased (densely packed/widespread)

Table 2.1: Ordinal grading scale for acute and chronic disease.

2.8 Conclusion

This chapter has presented the medical, technical and theoretical background upon which this research has been built. In the next chapter, a detailed review of existing literature is carried out, to cover important topics, particularly with regards to the appropriate processing of MRI data.

Chapter 3

Literature Review

In this chapter, a review of the literature is presented. Note that, due to the very narrow problem domain, it has not been possible to review a great quantity of very similar work. However, algorithms used in the detection and diagnosis of different diseases and other related published literature have been reviewed.

3.1 General surveys and background

Wang and Summers [71] give a broad overview of machine learning and CAD in the radiology setting, including segmentation, registration, cluster analysis, neural networks, etc. Cluster analysis has many applications in radiology and the k-means clustering, hierarchical clustering, DBSCAN, and mixtures of Gaussians methods are reviewed [71, pp. 937–938]. Registration and optimisation is explained to be part of machine learning [71, pp. 941–942]. Segmentation methods used in medical imaging are also reviewed by [17].

Terms used in the literature are Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx), both of which may be referred to as CAD. CAD aims to improve diagnostic accuracy, reduce reading time, or increase detection sensitivity. Usually, CAD systems are designed to deal with only one specific disease [71, p. 942]. CADx has been used extensively in particular areas of medicine, for instance: breast tumour detection and diagnosis, polyp detection in colon cancer, and pulmonary embolisms [71, pp. 942–943]. The brain and liver have been the subject of increasing research [15, p. 201]. More recently, MRI and contrast-enhanced MRI have been used in the diagnosis, prediction, and monitoring of rheumatoid arthritis [6].

Note that the distinction between detection and diagnosis lies in how the information is likely to be used: *detection* normally refers to CAD systems where the algorithm is viewed as a second reader and its performance is measured in

conjunction with a physician, whereas *diagnosis* systems (CADx) are considered standalone and therefore the required sensitivity is much higher [15, pp. 200–201].

Registration is a crucial step for any process that will combine data from two or more sources [20, p. 1381]. In the context of medical image registration, unregistered images may lead to an incorrect diagnosis being reached [43, p. 236].

Since there is much published literature on the subject of registration, the review has focused on published surveys. Registration as a general process is reviewed by [77]. Medical image registration is reviewed by [1; 43; 32; 51; 31]. In particular, [31, pp. 276–277] looked at multimodal medical image registration and found that Correlation Coefficient (CC), Correlation Ratio (CR), and Mutual Information (MI) were the most popular intensity-based measures. CC assumes some linear relationship between the intensities in the images. CR uses any functional relationship, while MI uses the statistical dependency between intensity values.

Using CR for the registration of multimodality images was first proposed by Roche et al. [64]. Using the brain scans of ten patients, Roche et al. found that CR had a lower mean registration error than MI for T1-to-T2 inputs. They argued that, since tissues are never represented by a single intensity value, it is important to use a measure that considers nearby intensities in the intensity space and this property is not utilised by MI.

Despite this, MI has been used extensively to register both multimodal and monomodal images [23, p. 28].

Kassner and Thornhill [42] reviewed the use of texture analysis for neurological MR applications with particular focus on CADx. They first point out that experienced radiologists spot patterns in brain scans and rely on their judgement to make a diagnosis, but that inevitability there are components of human variability and differences in visual perception, and that these can be mitigated by using automatic tools. For neurological applications, it was hypothesised that a "texture signature" could be found to describe different diseases by examining the changes in gray-level transitions found in MR images. Statistical descriptions of gray-level changes include first-order (e.g. mean, variance, gradients), and second-order measures such as Gray-Level Co-occurrence Matrix (GLCM) and Run-Length Matrix (RLM). The GLCM is a tabulation of a particular relationship occurring between pairs of pixels in the input image for the number of gray-levels present. Often the input image is quantized such that a relatively small number of gray-levels exist (e.g. 8) to prevent sparsity. The image may then be used with some relationship considering pixels that are at distance $d = 1$ pixel away and at angle $\theta = 45°$, resulting in an 8×8 matrix. Multiple GLCMs can be constructed systematically e.g. $d = 2$ and $\theta = 0°, 45°, 90°, 135°$ and the resulting matrices av-

eraged element-wise to give a direction-invariant GLCM for $d = 2$. They note that 14 textural features can be computed on the GLCM (e.g. contrast, correlation [42, Appendix Table 2]). RLM tabulates the number of consecutive pixels (i.e. in a run) for a given direction: $P_\theta(i, j)$ for gray-level i and run-length j, and 5 features can be computed on it [42, Appendix Table 3]. They suggested that both GLCM and RLM be normalised by dividing by the total number of matrix entries.

Kassner and Thornhill [42] also describe the problem of data "over-fitting", where the feature vector is large enough to describe the noise present in the input data and therefore any possible classification. They go on to briefly describe feature space reduction, classification methods, and cross-validated evaluation. Cross-validation (k-folds) is described as partitioning the data into training and testing sets. The training data is used to train the classifier and its performance is scored on the testing data which is an "out of sample" evaluation. The partition is "rotated" such that the testing data is different and the process performed again until all rotations have been tried. The overall success is the average of the test scores. The term "folds" refers to the split between training and testing data and the number of rotations. For instance, $k = 10$ folds would describe a ratio of 9:1 (training to testing, respectively) with 10 rotations, and the overall performance scored out of 10.

3.2 ROC analysis

Excellent introductions to the topic of Receiver Operating Characteristic (ROC) curves, their use in medicine, and with classifiers, has been given in [18; 25; 69]. ROC curves have been used in many different domains to assess the classification ability of a certain method and is therefore a very useful comparison tool.

The ROC curve plots the True Positive Rate (TPR) over the False Positive Rate (FPR). An equivalent interpretation is sensitivity on the ordinate and (100-specificity) on the abscissa. Points in ROC space represent a trade-off between the benefit of correctly identifying positive instances versus the cost of incorrectly labelling negative instances as positive [18].

To take an example, the output from a Naive Bayes classifier is the probability of the instance belonging to a certain class. Given multiple instances, an ROC curve may be generated by considering the TPR and FPR which is obtained for a set of thresholds, e.g. "probability $\geq t_i$". The raw probabilities do not have an effect on the ROC curve, only their relative ordering (i.e. their rank), and therefore, rather than probabilities, any suitable scoring mechanism that expresses confidence could be used. Table 3.1 and Figure 3.1 show the hypothetical output from a Naive Bayes classifier and its associated ROC curve. The diagonal dashed

line represents a coin-toss classifier that predicts no better than random.

The Area Under the Curve (AUC) of the ROC, also called Area Under the Receiver Operating characteristic Curve (AUROC), is important because it represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance [18] and it is mathematically equivalent to the Wilcoxon statistic W [25, pp. 30–31]. Therefore the area under the ROC curve is a useful metric for comparing classifier performance. A random coin-toss classifier has a theoretical AUROC of 0.5 and therefore any classifier of interest would normally be above this value.

No.	Truth	Score	t	TPR	FPR
			$+\infty$	0.0	0.0
1	P	0.92	0.92	0.2	0.0
2	P	0.91	0.91	0.4	0.0
3	N	0.83	0.83	0.4	0.2
4	P	0.72	0.72	0.6	0.2
5	P	0.50	0.50	0.8	0.2
6	N	0.49	0.49	0.8	0.4
7	N	0.34	0.34	0.8	0.6
8	N	0.33	0.33	0.8	0.8
9	P	0.28	0.28	1.0	0.8
10	N	0.11	0.11	1.0	1.0

Table 3.1: Example of scores from a classifier.

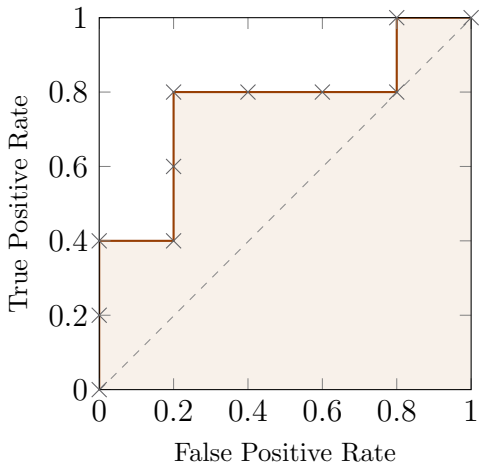


Figure 3.1: Example construction of an ROC curve.

3.3 Imaging artefacts in MRI

There are a wide variety of MRI sequences which aim to express certain properties of the tissues. For instance, the STIR sequence suppresses signal from fat tissue but enhances signal from oedema [53, p. 732]. However, some sequences are more susceptible to imaging artefacts and distortions than others. The STIR sequence suffers from poor SNR and imaging artefacts can be produced as a result of e.g. patient movement, and circulating blood (pulsation) [53, p. 732]. Patient movement may cause imaging artefacts for any sequence but it is not expected to be a major problem when scanning the legs because it is distant from the ongoing movements of the lungs and heart, although some patients may be unable to remain still (e.g. the elderly). The circulation of blood, particularly through the arteries (e.g. any of the femoral arteries, see "Blood vessel" in Figure 2.4), are expected to cause "zipper" artefacts on the STIR images, vertically or horizontally (depending on the phase encoding direction of the scanner). Brown and Semelka [8, pp. 113–117]

explained that circulating blood can also produce ghosting artefacts.

T1 images may also contain the imaging artefacts discussed previously but are generally of better quality. Both T1 and STIR images may show "lighting" artefacts. For instance (a) and (c) in Figure 2.1 (page 7) clearly show anisotropy effects in the fat layer, which theoretically should have homogeneous intensity for the same tissue type. This may be a result of the inhomogeneous static magnetic field used in MRI machines, although, as explored later, there are a plethora of interacting physical phenomena that cause deviations from the "true" image.

Since all imaging sequences used in MRI use Radio Frequency (RF) electromagnetic pulses, imaging artefacts can result from RF interference from other equipment. This could appear on any sequence, and has been seen on T1 and STIR.

As explained in [5], the limitations of the electronics and hardware of the MRI machine prevent perfect sequencing in terms of (e.g.) timing and magnitude. The performance of the RF receiver coils depend upon its exact geometry and physical construction. The gradient fields used for slice selection, frequency encoding, and phase encoding may also suffer from non-linearity, which can cause geometric distortion and eddy currents.

3.4 Intensity correction for MRI

Unlike other imaging techniques such as CT, the intensity from MR is generally uncalibrated [56; 57]. For instance, the intensity value of fat tissue for a patient scanned on one MRI unit may not result in the same intensity for the same patient on a different unit, even when using similar sequence parameters [63]. This is due to differences in the machines [57; 5] and their magnetic field inhomogeneity [70]. Indeed, Nyúl, Udupa and Zhang [57] stated that intensity values cannot be expected to be consistent even if the same patient and the same scanner unit were used.

The effect of the inhomogeneous magnetic field may exacerbate these problems. If a patient is scanned twice on the same unit on subsequent days, the patient's position relative to the magnetic field may not be the same, and therefore the intensity response may be different for the same tissue. However, other factors are also responsible.

Tong et al. [66] presented intensity correction as two separate correction problems to consider: *non-uniformity* refers to the inhomogeneous nature of the magnetic field (and the possible effects of other physically interacting phenomena), while *non-standardness* refers to the problem that tissues do not have a standard intensity value across all slices or from difference scanners.

To help understand the complex nature of MRI, the literature has been explored. It has been shown that the RF signals used in acquisition are absorbed by human tissue and as a result the core body temperature is elevated [48]. It has also been shown that temperature has an effect on T1 relaxation times and thus the signal intensity [58], and conversely it has been shown that it is possible to use MRI to monitor temperature [35]. Consider also that MR samples are taken in the spatial frequency domain [26], and therefore one can see that any change in the physical property of the material being scanned, in any *single* voxel, results in a global change for *all* pixels after application of the Fourier Transform (FT).

Most importantly, Belaroussi et al. [5] explained that the physical properties and shape of the material being imaged significantly affected intensity non-uniformity due to RF standing waves and penetration [5, p. 235].

In summary, the literature has shown that some form of intensity correction (both for non-uniformity and non-standardness) is required when working with MR data, particularly for quantitative measurement [33; 5].

3.4.1 Non-uniformity correction

Belaroussi et al. [5] and Vovk, Pernus and Likar [70] review methods for non-uniformity (inhomogeneity) correction. Vovk, Pernus and Likar [70] divided the methods into two main groups of prospective and retrospective. Prospective methods require additional equipment during acquisition: a *phantom* is a fixed material used for reference. Retrospective methods are image processing techniques applied in a post-processing stage [70].

In the literature, correction of non-uniformity is often modelled by a multiplicative field (sometimes called a *bias field*) over the image, with the form

$$v(\mathbf{x}) = u(\mathbf{x})f(\mathbf{x}) + n(\mathbf{x}), \tag{3.1}$$

where for some location $\mathbf{x}$, v is the acquired signal, u is the ideal signal, f is the bias field component, and n is a noise component (e.g. [12; 5]). If the noise component n is ignored then the log transform separates u and f components to be an additive relationship such that

$$\hat{v} = \hat{u} + \hat{f} \tag{3.2}$$

where $\hat{v} = \log v$, $\hat{u} = \log u$, and $\hat{f} = \log f$ (e.g. [70; 12; 65]).

Now that the mathematical form has been introduced, a brief overview of some of the approaches used in retrospective methods will be discussed, as outlined in [70]. Filtering methods assume that the bias field is a slowly-varying function

over the image. One such technique is called *homomorphic unsharp masking*, where the input image is divided by a low-pass version. Surface fitting methods use a parametric surface (e.g. using polynomials or splines) to model the bias field. Segmentation-based methods view intensity inhomogeneity as a segmentation problem, which essentially minimises the scatter of tissue classes, with or without *a priori* knowledge of the expected tissue intensities. Histogram based methods are fully automatic, requiring no *a prior* knowledge.

An outline of all proposed methods is not possible, however some specific methods from the literature will now be discussed.

Sled, Zijdenbos and Evans [65] proposed the N3 method which is a histogram based method and stands for "non-parametric non-uniformity normalization". N3 works by considering the probability density functions of the variables in Equation 3.2 such that

$$V = U * F \tag{3.3}$$

where $*$ is the convolution operator, $V = \texttt{pdf}(\hat{v})$, $U = \texttt{pdf}(\hat{u})$, $F = \texttt{pdf}(\hat{f})$, and are approximated to be independent random variables. Sled, Zijdenbos and Evans [65] go on to explain that the high frequency component of U represents the detail of the true signal which is blurred due to the presence of the bias field (density function F). In an iterative process, the frequency content of U is maximised. The approach proposes a sharpened distribution of V on which the bias field $\hat{f}$ (constrained to be smoothly varying) is estimated that also has density function U approximately close to the proposed distribution [65, p. 88]. This problem would be intractable (considering all possible fields) but the bias field search space is reduced by approximating F as a Gaussian.

Tustison et al. [67] proposed an update to the N3 method called N4ITK. In this update, a new (parallelisable) B-spline estimator is used, and on each iteration the residual bias field is estimated (as opposed to the whole bias field, as was done in N3). They state this allows for dense control points, for the correction of high-Tesla fields [67, p. 1311]. N4ITK was made publicly available as part of the Insight segmentation and registration ToolKit (ITK).

Ji et al. [41] describes another segmentation based method for non-uniformity (inhomogeneity) correction. Anisotropy effects in the input image were corrected by modelling the bias field as a multiplicative component of the true image,

$$I = b \cdot J + n \tag{3.4}$$

where I is the observed image, J is the true image, n is Gaussian noise, and b is the bias field modelled by the linear combination of a number of orthogonal polynomials (notice the similarity with Equation 3.1). Tissue class membership was

determined via a generalised rough fuzzy c-means algorithm which had positive, boundary, and negative regions per class. Each pixel in the image had an associated intensity distance vector which was calculated by considering the average intensity distance for a neighbourhood around the pixel to all unique intensities in the image. Global thresholds are established by calculating the average minimum distance of all distance vectors and the same for the average maximum of all distance vectors. These thresholds were used to determine positive, boundary and negative regions for each class. Pixels in a positive region definitely belong to one class, and those in a negative region are excluded from a given class. The goal of the generalised rough fuzzy c-means algorithm was to then minimise an objective function representing the spread of intensities for all classes for all pixels in a positive or boundary region [41, p. 647].

3.4.2 Intensity standardisation

Nyúl and Udupa describe their method of intensity standardisation which was developed over several publications [55; 56; 57] where the core idea has remained constant. It is assumed that it is desirable to remove the "tails" of the histogram which are considered outliers or a result of MR artefacts [55, p. 596]. In a training step, a landmark in the histogram space is computed (e.g. the mode, median, or percentile) and transformed to a standard intensity scale space using upper and lower percentiles (defining an Intensity of Interest (IOI) range). The mean of the transformed landmarks μ_{mean} is calculated. They note that this only has to be done once. In the second step, intensities of the input images are transformed by mapping piecewise linearly over the landmarks to the standard intensity space. Note that more than one landmark may be used, for instance when the histogram is bimodal, in which case there are two mean transformed landmarks found in the training step. They note that the histogram shape, and hence the selection of landmarks, is dependent on the imaging sequence and the body region.

Robitaille et al. [63] designed brain-specific histogram landmarks for use with the method of Nyúl and Udupa. They use the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which is a large mutlicentre database of brain scans and associated medical data of healthy controls and Alzheimer's Disease (AD) patients. Robitaille et al. explain that the problem of intensity standardisation is a particular problem in large multicentric databases, such as ADNI. They identify landmarks in the histogram space specific to different brain tissues and show the subsequent standardisation has less error than using the L_4 landmarks set (described in [57]). They noted that in some domains "overstandardizing" [63, p. 7] may be a problem, depending on the pathology in question.

Hellier [27] used a Guassian mixture to model the intensity histogram from (e.g.) two images and the parametric correction was computed as a least squares problem. In this work, there is no requirement that the images be registered.

Jäger et al. [39] approached intensity standardisation as a minimisation problem of the distance between probability density functions of the acquired sequence images (see also Jäger and Hornegger [38]). The multidimensional joint histograms are aligned non-rigidly using a registration method (e.g. MI, sum of squared differences, normalised cross correlation)[39, p. 273]. Joint histograms were equalised prior to registration in order that low-probability tissues were raised in the joint histogram space and high-probability tissues were suppressed. This was done to ensure the registration was not biased towards high-probability tissues [38, p. 139]. Still, Jäger and Hornegger [38] note that tissue classes with few voxels may not be registered reliably (referring to histogram space) and is also dependent on successful inhomogeneity correction [38, p. 139].

3.4.3 Discussion

From reviewing methods for non-uniformity correction and intensity standardisation, it is clear that both are important steps but it is not clear in which order they should be corrected.

Madabhushi and Udupa [50] used nearly 4000 MR datasets to explore the relationship between intensity standardisation and inhomogeneity correction. Their results indicated that inhomogeneity correction methods were indifferent to whether the input was standardised or not, but intensity standardisation methods performed better on inhomogeneity corrected input. Therefore inhomogeneity correction should precede intensity standardisation.

Considering all the work reviewed for intensity correction, it can be seen that working with MRI requires careful adjustment to reduce inhomogeneity and bring scans into a common frame of reference using intensity standardisation (with care regarding the specific pathology being considered).

3.5 Using MRI for disease detection

In this section, very important non-automated MR research for myositis and other muscle diseases is presented.

In the general case, MRI has been identified as a suitable imaging modality in the evaluation of soft tissue abnormalities [4, pp. 253–254], which would include the IIM disease group.

Kim et al. [45] found that total signal intensity (as scored by a visual scale

on affected muscles) was mildly associated with clinical findings such as aldolase ($r = 0.363$, $p < 0.004$), histopathological scoring ($r = 0.339$, $p < 0.05$), and muscle power ($p < 0.05$). These results were for dermatomyositis and polymyositis patients versus a control cohort.

Van De Vlekkert et al. [68] stated that using MRI increased the accuracy of diagnosis for Idiopathic Inflammatory Myopathies (IIM). Using only muscle biopsy tests to diagnose IIM resulted in a False Negative Rate (FNR) of 0.23. Biopsies with positive MRI findings (hyperintense signal) reduced the FNR to 0.19. Using MRI when biopsies were negative resulted in a FNR of 0.06.

Hutchinson [34] has shown that the combination of T1 and STIR signal may identify suspected disease in myositis.

Fischmann et al. [19] demonstrated that the change in T2 relaxation over time was significantly higher in patients with oculopharyngeal muscular dystrophy ($p = 0.01$) compared with controls. Longer time intervals between scans suggested larger changes in T2 relaxation.

Kim et al. [44] showed that non-quantitative fat infiltration scores (visually graded on a set scale) correlated with mean T2 relaxation times ($p < 0.001$) for skeletal muscle in Duchenne muscular dystrophy [44].

Note that T2 relaxation times are computed with a least-squares model from multiple echoes and represent a purely quantitative figure. Although useful for research, calculating this quantity requires long scanning times, and is not performed in the standard protocol for myositis.

In Herlidou et al. [30], nine senior radiologists were presented with MRI scans of the thigh and asked to give five different texture scores (coarseness, contrast, fat infiltration, complexity, and force) for which definitions had been provided. The scans were of 14 normal subjects and 17 patients suffering from muscular dystrophy but the radiologists did not have access to any related diagnostic information (muscular dystrophy must be considered as distinct from myositis). Some scans were presented twice to measure intra-observer variability. The variability of the scores was significant in both the inter-observer ($p < 0.001$) and intra-observer cases [30, pp. 1395–1396]. This shows the inherent subjective nature of visually assessing muscle diseases using MR images.

In summary, this section has shown encouraging evidence from non-quantitative, semi-quantitative, and quantitative methods, that MRI contains substantial data applicable to muscle diseases, which may be used in computer-aided analysis.

3.6 Quantification of myositis

A literature search for published articles relating to the automatic detection or diagnosis of myositis was conducted and only one article was found.

Bartlett et al. [3] used the histogram of pixel intensity values from STIR images to quantify the amount of diseased muscle. The method requires an initial estimate of the pixel intensities representing healthy tissue, which was either taken as the average of pixel intensities taken from a normal subject, or the manual placement of a small ROI on healthy muscle. A Gaussian curve is fitted to the peak intensity on the histogram, which identifies the distribution of healthy muscle, and this is subtracted from the histogram. Due to diseased muscle typically appearing brighter on STIR images than healthy muscle, the diseased muscle is assumed to be pixels with intensities to the right of the Gaussian. However, different histogram shapes are to be expected, since patients have different proportions of healthy and diseased muscle, and as the authors explain [3, pp. 185–187], in some cases a healthy peak could not be identified in the histogram. In this situation the peak with the maximum intensity was taken as the diseased muscle, a Gaussian curve fitted to the peak (using a pre-determined HWHM value), and then subtracted from the histogram. The remaining, most intense peak, is then assumed to be healthy muscle and a Gaussian curve is fitted to this peak. In either case the quantity of diseased muscle can be estimated by the residual diseased histogram.

Bartlett et al. used a phantom filled with a tissue-simulating solution with five normal subjects to assess the variability of the healthy tissue peak in the histogram. The standard deviation of the intensity value for the identified peak was found to be 6.3% for ten scans (five normal subjects scanned twice). This is used as justification that other scans in the study would have similar intensity values for healthy tissue. 6.3% is assumed to be relative standard deviation, RSD $= 100\frac{\sigma}{\mu}$ (a scaled version of the coefficient of variation).

Two measures of the disease were determined: extent and intensity. The extent of the disease is the number of pixels classified as diseased as a percentage of the pixels classified as diseased or healthy. For disease intensity, the x-axis of the residual diseased histogram is linearly scaled such that the healthy tissue peak corresponds to a value of 100. The intensity measure is then calculated as the mean of the pixel intensities in the residual diseased histogram (values above 100). The authors note that often a few pixels have very high intensities due to the presence of blood vessels and these do not represent diseased muscle. Therefore, histogram bins are excluded from the calculation when they contain less than 2% of the number of pixels found at the healthy tissue peak.

Bartlett et al. gave results as Area Under the Curve (AUC) considered with

different clinical categories as ground truth (Table 3.2). Separating remission vs. non-remission gave 86% sensitivity and 83% specificity.

Test	Visual score	CMI
Remission vs. mild/moderate/severe	0.94	0.93
Remission/mild vs. moderate/severe	0.92	0.92
Remission/mild/moderate vs. severe	0.68	0.65

Table 3.2: Data from Bartlett et al. [3]: AUROC considered for different clinical categories.

The main concern with the approach taken is that it depends on the quality of the Gaussian fit to the underlying distribution of muscle tissue (normal, and diseased) in the MR image. It is possible that a different distribution would fit the distribution better, something that might take into account the "tails" of the distribution, which may be an important presentation of the pathology in myositis (c.f. [55, p. 596]).

From the discussion in section 3.4, it is clear that correction of non-uniformity would be of benefit (note that publication was in Feb. 1999, although correction methods had been published e.g. [12; 65]).

The technique requires that a peak be found in the intensity histogram, and in cases where there is no definitive peak, the selection of a peak becomes somewhat arbitrary (cases with approximately 50% diseased pixels may have broad distributions). This is a greater concern when considering multiple MR scanners. The intensity measure of the disease is sensitive to changes in the location of the healthy muscle peak.

By definition, the histogram as presented in Bartlett et al. [3, p. 185] can only occur in the most severe cases; when the number of hyperintense pixels exceeds the number of pixels considered healthy. In the training and out of sample databases, all subjects had less than $\frac{1}{2}$ oedema (of total muscle). Although research in all stages of disease can be of use, it is of clinical and practical benefit to have methods which focus on the lower severity range.

To quantify the level of fat infiltration, STIR is unsuitable because the sequence is designed to suppress signal from fat. It is already known that the T1 image can adequately identify fat infiltration for myositis [34; 30, p. 1396].

3.7 Detection of neuromuscular diseases

König et al. [46] used a computer-aided approach with Ultrasound (US) acquisitions of the upper arm to detect muscle disease. Between 2 and 4 images were acquired on 7 healthy subjects and 11 subjects with myositis. For each image an

ROI was manually defined and subsequent processing was performed on this region only. Several different features were selected to represent the textural appearance of the disease on the scans. This included the mean, standard deviation, skewness, kurtosis, and entropy of the intensity histogram. Other features were based on the GLCM and two-dimensional discrete wavelets. The initial feature vector had 375 elements. Feature space reduction was performed using supervised principal component analysis, of which the top 5 features were selected for training.

Classification was performed with 3 different methods: k nearest neighbour, Fisher's linear discriminant, and linear Support Vector Machine (SVM). Results indicated that the SVM classifier had fewer errors compared to the others. Using the average of two ROIs, the system attained 90% sensitivity and 84% specificity.

It was assumed that US highlights the disease based on the replacement of muscle by fat [46, p. 1493], as caused by the disease. However, this represents long-term (chronic) disease, not active (acute) disease. US is not normally performed in routine assessment of myositis due to poor response from inflammation (oedema). Although the system successfully differentiates betweens healthy and diseased individuals, it is not clear if the system can successfully differentiate the degree of disease state.

3.8 Quantification of meniscus tears

Ramakrishna et al. [61] developed an algorithm to detect tears in the meniscus. The meniscus is a supporting structure located in the knee joint and is different from cartilage and bone. The meniscus may become broken when subjected to rotational forces.

Although the problem domain is different, Ramakrishna et al. presented a similar situation in that a diagnosis was previously made by a radiologist through visual inspection of MR scans. An objective measure of the severity of the breakage was sought. Due to the variation in signal intensity that is expected from MR scans (e.g. between different hospitals and scanners), the authors designed an algorithm with an adaptive threshold which could be calculated automatically for each case.

In a preprocessing stage, the ROI is first selected by choosing a region with the lowest mean intensity as calculated using all the slices in an image cube [61, p. 1311]. A "mean image" is described by

$$M(i,j) = \text{mean}\left\{I_k(i,j) | I_k \in S_{\text{initial}}\right\} \tag{3.5}$$

where M is the mean image of all sagittal slices I_k in the set S_{initial}. Matrix M is then split horizontally into submatrices denoted W_n. Each W_n is 16 rows high.

The average intensity for each W_n is found and the minimum is selected; this submatrix is denoted n^*. The centre row of n^* is the centre of the ROI denoted I_k^{ROI} which has size 180×100. In a process that is not detailed here, the set S_{final} contains only some slices of S_{initial}.

After the preprocessing stage an adaptive threshold is calculated in order to binarise the images in I_k^{ROI}. The threshold is 0.2μ, where

$$\mu = \text{mean}\left\{I_k^{\text{ROI}} | I_k \in S_{\text{final}}\right\}. \tag{3.6}$$

The binarised images have a number of segments, and these are filtered based on shape and size parameters (e.g. length-to-width ratio, area). This process is repeated on the set of slices with an increased threshold, in steps of 0.05μ up to 0.8μ. The threshold that resulted in the maximum number of filtered segments is selected. Let τ^* be this threshold. The filtered segments are used to extract objects from the original image.

The authors define two measures: breakability and degenerability. Breakbility is measured by the maximum number of segments that result when considering a set of five thresholds $0.7\tau^*, 0.6\tau^*, \ldots, 0.3\tau^*$ applied to the extracted objects. A score is assigned to each of the five thresholds. Degenerability uses the same five thresholds on the extracted objects to measure the rate with which the area of the resulting segments decrease.

The authors found that tears running through consecutive slices have a higher chance of being true tears, and so continuity scores are calculated

$$\hat{\alpha}_k = 0.25\alpha_{k-1} + 0.5\alpha_k + 0.25\alpha_{k+1}, \tag{3.7}$$

$$\hat{\beta}_k = 0.25\beta_{k-1} + 0.5\beta_k + 0.25\beta_{k+1}, \tag{3.8}$$

where α_k is breakability for slice k, and β_k is degenerability for slice k. $\hat{\alpha}_k$ and $\hat{\beta}_k$ are then combined into a final score.

The sensitivity and specificity of the algorithm was found to be 83.87% and 75.19% respectively. An experienced radiologist attained sensitivity of 77.41% and specificity of 81.39%. I.e. the algorithm was slightly better at diagnosing positive cases as positive, but worse at identifying negative cases as negative, compared to a radiologist.

Köse, Gençalioğlu and Şevik [47] also present a different method for the automatic diagnosis of meniscus tears. The authors present several different techniques to locate the meniscus ROI in MR images taken in the sagittal plane. The techniques were used on a collection of approximately 500 archived images from 30 patients.

The authors used vertical and horizontal histograms to utilise the knee morphology, but note that the histograms should be smoothed, normalised, and scaled to account for differences in presentation (e.g. the angle between the tibia and femur). If the vertical and horizontal histograms do not match reference histograms then the original input image may be rotated and the histograms recalculated. Then the extreme values in the histograms are used to find the meniscus ROI.

A second histogram based method uses an edge-filter on the input image first. The authors used five different edge-filters (Sobel, Prewitt, Roberts, Robinson, and Frei-Chen). The vertical and horizontal histograms are then only computed for pixels above a threshold and the resulting histograms are normalised. The extreme values in the normalised histogram are then used to find the meniscus ROI. Out of the five filters, results showed Sobel, Prewitt, and Frei-Chen gave the best horizontal localisation.

A statistical segmentation technique was also evaluated. Characteristic images of the bone region are used to find statistical properties of the bone texture. In a training step, the bone region from an input image is segmented and a small seed image is randomly selected out of the bone region. Many small images of the same size are then randomly selected from the bone region and the average error threshold is computed against the seed image. This process is repeated many times and the characteristic images with the minimum average error are selected and the statistical properties calculated on these images. Segmentation is performed by taking a small sample image around the centre of a pixel in the image and calculating the statistical properties. The difference is calculated against the characteristic images' properties. When the difference is under some threshold, the centre pixel is considered bone, otherwise it is considered non-bone. This is done for all pixels in the image. The resulting segmentation is enlarged using an adaptive region growing technique. Finally, the vertical and horizontal histograms of the segmented bone region are computed and used to find the meniscus ROI.

The authors state that the most invariant morphology for locating the meniscal region is the bones, although the statistical bone segmentation technique resulted in only 82.0% of meniscal regions accurately located when using both top (femur) and bottom (tibia) bones. When using only the femur bone the statistical segmentation technique resulted in 94.0% of meniscal regions accurately located. The histogram based method which had two maximums as reference points achieved 99.0% accuracy in localising the ROI vertically, but performed badly overall (38%) due to horizontal inaccuracy. The normalised histogram method using edge-filtered images attained a horizontal accuracy of up to 98.0%.

The meniscus ROI is binarised using Otsu's method and the meniscus is identified using triangle templates of various sizes. The search region is restricted, based

on statistical information of the most likely locations. Tears are identified as low intensity pixels surrounded by high intensity pixels and the ratio of meniscus to tears is calculated using real areas (as computed using the metadata recorded in the DICOM file). If this is above some threshold value then the diagnosis is positive. Results showed the algorithm successfully diagnosed 88.3% of 30 knees with tears and 95.7% of 70 knees without tears.

In summary, Köse, Gençalioğlu and Şevik [47] presented several different methods in attempt to identify the ROI. Using an edge filter was successful in identifying the meniscus due to the greater number of edges in that area. Vertical and horizontal histograms can also be used to locate objects. The meniscus was binarised using an automatic technique called Otsu's method, which assumes the image has "background" and "foreground" pixels, whereas Ramakrishna et al. [61] used a simple adaptive threshold technique. In either case, thresholding worked well because that part of the process must detect tears, which are visible as lines running through the meniscus and therefore naturally lends itself to segmentation. Ramakrishna et al. [61] utilised the information from consecutive slices whereas Köse, Gençalioğlu and Şevik [47] did not. Examining consecutive slices meant the algorithm gave higher scores when the presentation was more obvious. Köse, Gençalioğlu and Şevik [47] only gave a binary diagnostic decision, whereas Ramakrishna et al. [61] also quantified the extent of the tears.

3.9 Multiple sclerosis lesion segmentation using joint histogram

Zeng and Zwiggelaar [75] propose the use of T1 and T2 sequences to build a joint histogram and present an automatic method to segment MS lesions in the brain.

They explain that the brain is largely made up of grey matter and white matter tissues and that the joint histogram therefore has an expected appearance, with the different tissues separated. They show that this cannot be done with standard histograms for T1 and T2 separately.

The cluster centres for grey matter and white matter tissues are calculated and a rectangular region in the joint histogram is computed using the cluster coordinates. The presentation of MS lesions in MR images is hypointense T1 signal and hyperintense T2 signal (both with respect to normal white matter), and therefore the rectangular region is assumed to contain the MS lesion.

The cluster centres are calculated using a class-adaptive Gaussian Markov random field model with 5 initial classes to segment the joint histogram, and the number of classes is increased until the white matter and grey matter regions are

distinct. Once this has been done for the middle slices in the set, the cluster centres are propagated to neighbouring slices by using a small circular region around the centres which is used as an initial estimate on the neighbouring slice.

Unwanted data in the joint histogram are removed by creating a mask of the most dense areas (using fuzzy C-means with 2 clusters) and then using morphological operators to refine it. The lesion pixels are selected from the original T1 and T2 images using the histogram mask and the rectangle region to pick out the intensity values of interest. After the regions have been selected from all the slices, regions which are not in roughly the same location in neighbouring slices are removed to reduce false positives.

3.10 Multiscale vessel enhancement filtering

Frangi et al. [21] describe the idea of filtering an image or volume for shapes that are tubular at a certain scale s. Tubular objects are assumed to be blood vessels. They show the technique on examples from magnetic resonance angiography (MR for blood vessels) and X-ray.

Frangi et al. start by introducing the Taylor expansion for the input image L at a point $\mathbf{x}_o$,

$$L(\mathbf{x}_o + \delta\mathbf{x}_o, s) \approx L(\mathbf{x}_o, s) + \delta\mathbf{x}_o^T \nabla_{o,s} + \delta\mathbf{x}_o^T \mathcal{H}_{o,s} \delta\mathbf{x}_o \tag{3.9}$$

where s is the scale parameter, $\nabla_{o,s}$ is the gradient (first-order) of image L at $\mathbf{x}_o$, and $\mathcal{H}_{o,s}$ is the Hessian matrix (second-order) of image L at $\mathbf{x}_o$. The Hessian is a square matrix which contains the second-order partial derivatives of the function.

They state that in this framework differentiation is the convolution of the input image with the derivative of Gaussian and give the equation

$$\frac{\partial}{\partial \mathbf{x}} L(\mathbf{x}, s) = s^\gamma L(\mathbf{x}) * \frac{\partial}{\partial \mathbf{x}} G(\mathbf{x}, s), \tag{3.10}$$

where $G(\mathbf{x}, s)$ is a D-dimensional Gaussian such that,

$$G(\mathbf{x}, s) = \frac{1}{(2\pi s^2)^{\frac{D}{2}}} e^{-\frac{\|\mathbf{x}\|^2}{2s^2}}. \tag{3.11}$$

Although not explained in the paper, it is important to realise that the expression $(L * G)'$ can be rewritten $L * (G')$, meaning that the differentiation of image L can be done by differentiating the function G and then convolving with L. In Equation 3.9 the third term describes local curvature by using the Hessian matrix, and this can be calculated by pre-computing the second-order partial derivatives

of G for a given scale s and then convolve with L.

Frangi et al. stated that the second-order Gaussian is particularly suitable for vessel detection since the function is similar to the expected change in intensities presented by a blood vessel.

That is, a blood vessel is essentially a tube of a certain diameter filled with moving fluid (with the central fluid moving moderately quicker than fluid towards the tube walls) and the second-order Gaussian has a shape well suited for this structure.

Information on the local second-order structure of the image is extracted using eigenvalue analysis of the Hessian matrix. The eigenvector and eigenvalue give the principal direction of curvature in the local neighbourhood. For a volume they order the eigenvalues such that $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$ and suggest that

$$|\lambda_1| \approx 0 \tag{3.12}$$

$$|\lambda_1| \ll |\lambda_2| \tag{3.13}$$

$$|\lambda_2| \approx |\lambda_3| \tag{3.14}$$

so that a blood vessel may be found. The polarity of any λ_k indicates whether it is a bright object on a dark background or vice versa.

Frangi et al. suggest a vesselness measure for two-dimensional images,

$$\mathcal{V}_o(s) = \begin{cases} 0 & \text{if } |\lambda_2| > 0, \\ \left[1 - \exp\left(-\dfrac{\mathcal{S}^2}{2c^2}\right)\right] \exp\left(-\dfrac{{\mathcal{R}_\mathcal{B}}^2}{2\beta^2}\right) & \text{otherwise,} \end{cases} \tag{3.15}$$

where

$$\mathcal{R}_\mathcal{B} = \frac{\lambda_1}{\lambda_2}, \qquad \mathcal{S} = \sqrt{\sum_{j \leq D} \lambda_j^2},$$

and β varies the input from "blobness" measure $\mathcal{R}_\mathcal{B}$, and c varies the input from "second order structureness" measure $\mathcal{S}$. The condition of Equation 3.15 may be reversed depending on the modality and the appearance of the vessel (bright on dark or vice versa). In order to assess vessels at different scales Frangi et al. suggest calculating $\mathcal{V}_o(s)$ for a range of s values and selecting the maximum.

3.11 Conclusion

In this chapter, previous work from similar and related domains has been reviewed. Some very important insights have been gained and this knowledge will be leveraged in Chapter 4 where the MRI scans are carefully prepared before disease analysis.

Chapter 4

Pre-Processing

Now that the background has been introduced and literature from similar domains explored, it is possible to appreciate the inherent requirement for pre-processing MRI data. The processes described in this chapter are fundamentally important in preparing the data in order that quantitative measures of inflammatory muscle disease may then be constructed (Chapter 5).

4.1 Overview

Figure 4.1 shows an outline of the steps required to process the scans. A summary of the steps will be detailed here, and the subsequent sections in this chapter go into further detail. In Step 1-A, a contiguous set of 6 slice pairs are selected from the scanner output and are then loaded from DICOM format in Step 1-B. The T1 and STIR images are aligned in Step 1-C and then scaled to a common pixel size in Step 1-D. An initial ROI is computed in Step 1-E. In Step 1-F the slice pair is corrected for non-uniformity effects. Fat and muscle tissue centres are found in Step 1-G. A refined ROI is computed in Step 1-H to reduce partial volume effects seen in the joint intensity space. Step 1-I is a specially developed tool to remove the femur (bone and bone marrow) using a semi-automatic process. In Step 1-J the largest arteries and veins in the leg are masked out. Step 1-K performs intensity standardisation suitable for the pathology under consideration. In Step 1-L the intensity space is segmented to get muscle and fat regions.

These steps are essential in order that the data is normalised to a standard frame of reference for the problem domain, and that all data which is not useful in diagnosis is discarded before further analysis (to help prevent any unexpected bias).

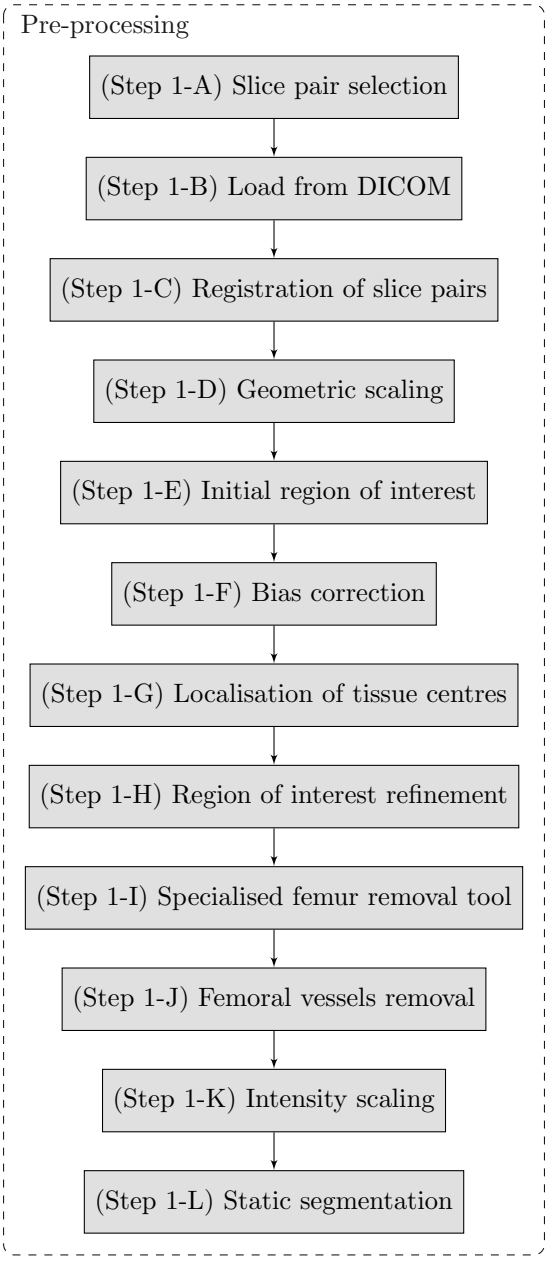


Figure 4.1: Process diagram showing the steps in pre-processing the data.

4.2 Notation

Some notation is introduced here, which is used in subsequent sections. The notation is the same which appears in [37].

An image may be thought of as a function $f : \mathbf{X} \to \mathbb{R}^+$, where

$$\mathbf{X} = \mathbb{Z}_m^+ \times \mathbb{Z}_n^+ = \left\{(x_1, x_2) \in \mathbb{Z}^2 : 1 \leq x_1 \leq m, 1 \leq x_2 \leq n\right\}, \tag{4.1}$$

which forms a rectangular lattice [62], where m is the image height, n is the image width, (x_1, x_2) is the coordinate, and $\mathbb{Z}$ is the set of integers. The same image

may be represented equivalently by the structure

$$\mathbf{f} = \{(x, f(x)) : x \in \mathbf{X}\}, \tag{4.2}$$

where x is the pixel location. For convenience the two source types, T1 and STIR, will be represented by $\mathbf{t}$ and $\mathbf{s}$, respectively, for the remaining chapters.

A bivariate histogram (or joint histogram) may be computed in the following way. The intensities of the input images $\mathbf{t}$ and $\mathbf{s}$ must be discretised for pixels in a Region Of Interest (ROI) e.g. for some mask $g(x) = 1$. The discretised images will be denoted $\hat{\mathbf{t}}$ and $\hat{\mathbf{s}}$, respectively, assumed to have 128 intensities, although other ranges may also be used.

For each entry (u, v) in a bivariate histogram, a subset $\xi(u, v) \subset \mathbf{X}$ is defined by

$$\xi(u, v) = \{x : \hat{s}(x) = u \wedge \hat{t}(x) = v\} \tag{4.3}$$

where $0 \leq u \leq 1$ and $0 \leq v \leq 1$. The bivariate histogram is then

$$\mathbf{h} = \{((u, v), \tfrac{1}{N}|\xi(u, v)|) : (u, v) \in [0, 1] \times [0, 1]\}, \tag{4.4}$$

where $N = |\{x : g(x) = 1, x \in \mathbf{X}\}|$, $\mathbf{h}$ is of size 128×128, and $|\mathbf{s}|$ is used to denote the number of elements in a set $\mathbf{s}$. In words, the bivariate histogram is a normalised tally of the occurrences of joint intensity (STIR pixel, T1 pixel) for all pixels in the ROI. In the instance where a ROI has not been defined then it includes all locations x.

4.3 Slice pair selection (Step 1-A)

For each sequence, the MRI machine writes a set of DICOM files into a single directory. Any number of directories may be available depending on what investigations were performed. The correct body region must be identified. In general, it is necessary to check the sequence parameters from the DICOM meta-data to ensure acquisition parameters are correct. Table 4.1 shows some example data. *Slice Location* is important to check between sequences to ensure that slice pairs refer to the same physical slice plane. Although some data is recorded automatically by the MRI machine, other data is entered manually and occasionally this is erroneous.

A set of 6 contiguous slice pairs are selected from T1 and STIR directories. Slices are selected below symphysis pubis and the gluteus maximus muscle is excluded. Any gross imaging artefacts are avoided, if present.

DICOM Key	Name	Example value
0008-103E	Series Description	Axial T1 Thigh
0018-0050	Slice Thickness	8 mm
0018-0087	Magnetic Field Strength	1.5 T
0020-1041	Slice Location	-28.35 mm
0028-0030	Pixel Spacing	0.88 mm

Table 4.1: Example DICOM keys checked before loading.

The selection of the number of slice pairs to use per case was determined by a number of factors. Utilisation of all available scans (including historical scans) restricted the number of usable contiguous slice pairs, due to variation in the exact body location which was imaged. However, due to the "patchy" nature of the disease, adequate coverage of the leg muscle was assumed to be important. Since the standard protocol dictates that slices are acquired at 10 mm intervals, then 6 contiguous slices surveys a region of leg 6 cm in length. A greater number of slice pairs per case would be preferred, however the overall number of scans meeting this requirement would be greatly reduced. Therefore 6 slices was selected as a sensible minimum based on the consideration of the above factors.

It was assumed statistically superior to use the largest muscles (indeed, this is why the protocol is for the thighs) and therefore slices towards the upper thigh were selected in preference to those closer to the knee. Note that this process need only be performed once per case.

4.4 Loading scans (Step 1-B)

DICOM files were loaded into memory. Images were normalised to [0,1] range (per slice). A table-like data structure was used such that all data pertaining to a "set" could be looked up: Table 4.2 shows an example.

Column name	Type	Example
Set	Integer	2
Slice	Integer	5
Type	String	'T1', 'MASK'
Control	Boolean	True/False
Diagnosis	String	'positive'
File	String	'IM6'
Image	Object	data structure

Table 4.2: Example data in a row of the lookup table. The Image object also contains important data from the DICOM data.

4.5 Slice pair registration (Step 1-C)

Registration is a process of finding a transformation that maps from one image space to another such that the two images may be superimposed and thus the corresponding objects in the scene will be in alignment. In general, registration is required as images may have been taken at different points in time (multitemporal), with different sensors or machines (multimodal), with differing sequence settings (multispectral), or from different viewpoints (multiview). There are many clinical and research applications for registration [51; 43]. Note that there is no clear distinction between multimodal and multispectral terms.

One image is said to be the *reference* image, which is static, while the other image is called the *sensed*, *target*, or *moving* image, which is transformed or *mapped* to align with the reference image [43, p. 236].

Registration can be split into four distinct categories based on the nature of the transformation used. These are rigid, affine, projective, and elastic [51, p. 8]. The coordinate transformation is considered rigid when only translation and rotation are considered and can be defined in 2D by

$$x' = \mathbf{R}x + t \tag{4.5}$$

where $t = (a_5, a_6)^T$ is a shift vector for translation and $\mathbf{R}$ is the rotation matrix

$$\mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{4.6}$$

When parallel lines are mapped onto parallel lines the transformation is named as affine. In 2D it is defined by

$$x' = \mathbf{A}x + t \tag{4.7}$$

where

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \tag{4.8}$$

is the affine matrix [20, p. 1381]. The affine matrix is a composition of rotation, scale, and shear transforms: $\mathbf{A} = \mathbf{RSE}$ [23, p. 346]. Projective transformations map lines on to lines, although parallel lines may not remain parallel after it is applied. It can be defined in 2D by

$$x' = \mathbf{B}x \tag{4.9}$$

where $\mathbf{B}$ is an arbitrary 3×3 homogeneous matrix [51, p. 8]. An elastic transformation (also called a warping, mesh, deformable or curved transformation) may

be described by mapping lines onto curves. It might be represented by a local displacement field such as $x' = x + t(x)$.

Registration is used to ensure that the T1 and STIR images are in alignment. Since the time taken between acquisitions of the sequences is minimal, it was assumed that a constraint of rigid registration would be sufficient. As already discussed in section 3.4, it is clear that some registration may be required even with a strict acquisition protocol.

The affine transformation may be limited to scaling and translation by adjusting Equation 4.7 such that

$$x' = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} x + t. \tag{4.10}$$

The search space that must be considered is $[s_1, s_2, t_1, t_2]$, a brute-force search of which is infeasible, partly due to the number of operations required for a single registration of two input images of reasonable size. The search space can be viewed as an optimisation problem where the registration measure must be either minimised or maximised. An optimiser may use a gradient descent method [23] to estimate s_1 and s_2 parameters, although the application of a scaling transform will have an effect on the translation parameters t_1 and t_2 [74]. The scaling transform should be applied to the center of the objects in the image, as described by

$$x' = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} (x - C) + C + t \tag{4.11}$$

where C is the center [74, p. 257].

With experience, the required adjustment to the *moving* image were small for all parameters $[s_1, s_2, t_1, t_2]$. Registrations for the training database were examined. The range for scale parameters $[s_1, s_2]$ for the training database were found to be small in the general case, but on occasion was much larger than necessary, causing gross misalignment.

The registration was further constrained to translation-only registration for the following reasons:

- A widened search space with parameters $[s_1, s_2]$ (as compared to $[t_1, t_2]$ only) combined with poor SNR of STIR input occasionally resulted in gross misalignment.

- High quality fat suppression achieved in STIR sequences could result in very dark or no signal from fat tissue, which may have contributed to inadequate alignment.

- From experience, the range of (visual) misalignment was small (less than 2 pixels in either dimension).
- In examining a great number of registrations (by fusion of the two images) translation-only alignment was seen to be effective.

Registration based on translation only has an advantage of being computationally simple and is often quick. Note that this could only be achieved due to the specific protocol used in this problem domain, with identical geometry for a slice pair.

4.6 Geometric scaling (Step 1-D)

Utilising the DICOM meta-data, all images were scaled to a common pixel dimension. The scale factor was chosen to target 0.74mm/pixel, chosen such that all images in the database were magnified.

4.7 Region of interest construction (Step 1-E)

Without any Region Of Interest (ROI) the joint histogram generated for a slice pair will result in one very dense peak representing very low T1 and STIR intensity, resulting from the large number of dark pixels from the background region. For an example, see (a) in Figure 2.2 (page 8). In this step the ROI is defined to remove the large background region. Later in pre-processing (Step 1-H and Step 1-I) the ROI is refined.

A simple masking technique has been designed to select a region of a few pixels around the legs, applied to each slice pair. The T1 image is used to create the mask, but may be applied to both images in the pair, as they have been registered. As previously discussed, the STIR image is often noisy, thus using the T1 image was considered advantageous.

Thresholding the T1 image, followed by erosion, creates the first mask

$$\mathbf{m}_1 = \big\{(x, c(x)) : c(x) = 1 \text{ if } t(x) \leq 0.05, \text{otherwise } c(x) = 0\big\} \ominus D(3) \qquad (4.12)$$

where $D(r)$ is a disk-shaped structuring element with radius r, and $\ominus$ is morphological erosion. $\mathbf{m}_1$ is now assumed to cover the majority of the background region, which contains small variations in intensity due to noise. A local standard deviation filter is used to measure variations throughout,

$$j(x) = \text{stdev}\{t(x) : x \in M_x\} \qquad (4.13)$$

where M_x is the 5×5 neighbourhood of $\mathbf{X}$ centred at $x \in \mathbf{X}$. The maximum standard deviation of interest is then

$$\rho = \max\{j(x)m_1(x) : x \in \mathbf{X}\}. \tag{4.14}$$

ρ is the selected threshold for mask

$$\mathbf{m}_2 = \left\{(x, c(x)) : c(x) = 1 \text{ if } j(x) > \rho, \text{otherwise } c(x) = 0\right\}. \tag{4.15}$$

The outermost region where $\mathbf{m}_2$ is 0 defines the background region. The standard binary reconstruction technique is used with $\mathbf{f}_0$ as the initial blank image with seed $f_0((1,1)) = 1$, assumed to be background. Defining $\oplus$ as morphological dilation then

$$\mathbf{f}_k = \left(\mathbf{f}_{k-1} \oplus D(1)\right) \cap \mathbf{m}_2^{\mathrm{c}} \text{ for } k = 1, 2, \ldots \tag{4.16}$$

until $\mathbf{f}_k = \mathbf{f}_{k-1}$, and note that $\mathbf{m}_2^{\mathrm{c}}$ denotes the complement. Using $\mathbf{f}_f$ as the final case then the ROI is

$$\mathbf{g} = \mathbf{f}_f^{\mathrm{c}}. \tag{4.17}$$

The above process is shown in Figure 4.2.

Although $\mathbf{m}_1$ is based on a simple threshold of the raw T1 image, it is reasonable to assume that the background is less than $\frac{1}{20}$ of the brightest pixel before intensity standardisation. This process was found to be sufficient for all cases in the training database.

4.8 Bias correction (Step 1-F)

The N4ITK method [67] was selected as the inhomogeneity correction method (see related discussed in section 3.4) for the following reasons:

- The previous version (N3 [65]) has been used on large MRI databases, e.g. ADNI.
- The updated method (N4ITK) is an improvement over N3.

Other methods were briefly investigated, but it was common to find a reduction in pathologically useful information (e.g. severely reduced intensity spread over tissue centres) where the correction method had worked "too well". (Note that subsection 3.4.2 covered the possibility of "overstandardizing" [63, p. 7], although that discussion was not related to inhomogeneity correction.)

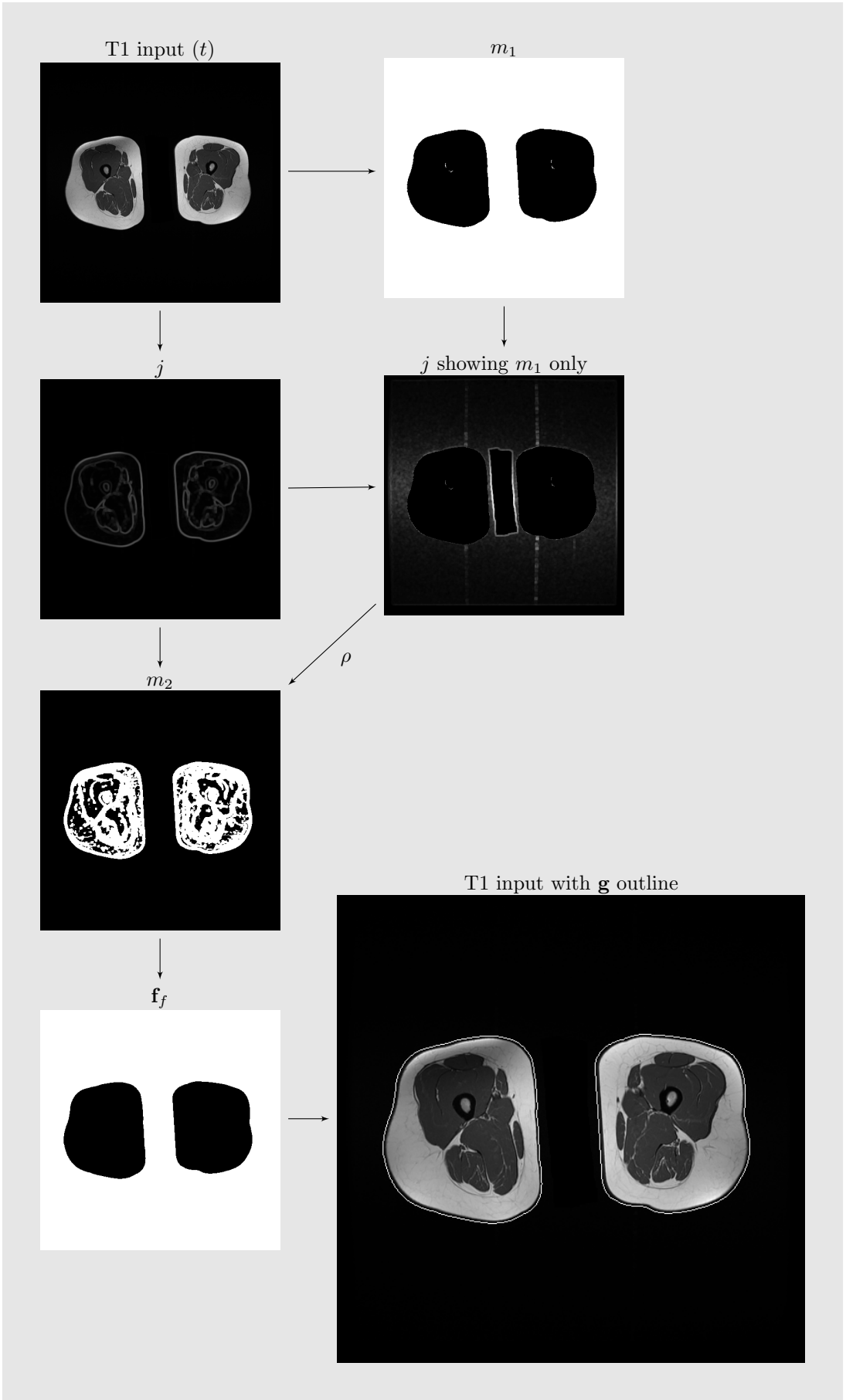


Figure 4.2: Overview of the process for computing the ROI.

4.9 Localisation of tissue centres (Step 1-G)

The bivariate histogram must be generated from the registered and bias-corrected inputs, using the ROI from Step 1-E. The computation for generating the bivariate histogram was described above.

Popular clustering algorithms were investigated, including k-means, DBSCAN, and affinity propagation, however these algorithms produced results that were often unsatisfactory when considering the whole training database. For instance, k-means did not always locate cluster centres satisfactorily, and largely relied on the starting seed locations.

Therefore a technique was developed that would be robust against changes in the bivariate histogram space: some changes are expected due to variation in the quantity of different tissues and their intensity, however some basic rules may be defined on the basis that any bivariate histogram that violated the rules would be most likely the result of improper sequence settings, severe MR artefact, or misregistration.

It is expected that:

- Two main tissue clusters exist in the intensity space. The femur bone and bone marrow is removed in a later step, but this region should not be a significant fraction of the overall leg, and therefore its contribution is not expected to distort the overall shape of the bivariate histogram. In practice, this is always true, even for slice pairs acquired towards the knee, where muscle may be relatively sparse.

- STIR signal from muscle inflammation (oedema) has the same approximate T1 intensity as healthy muscle.

- T1 signal from muscle and fat is sufficiently distinct (e.g. more than 10% of the total T1 intensity range).

Given the above constraints are met, then the following method is seen to be robust. Algorithm 4.1 is a generalised method for finding an approximate peak location in a bivariate histogram over a defined ROI (in intensity space) given template parameters.

The muscle and fat centres were found using different template parameters but using the same method (Algorithm 4.1). With the bivariate histogram H computed (dimensions 128×128), then the templates defined in Table 4.3 were used. (f_s, f_t) is the fat centre, and (m_s, m_t) is the muscle centre. N_1 and N_2 are the number of ranked indexes to average in each dimension.

The results were manually checked for every image-pair in the training database, judged by visual inspection. Note that the parameter N_2 is greater for the

Algorithm 4.1: Find approximate cluster centre location

Data: Joint histogram H of size $y \times x$, search space (y_1, y_2),(x_1, x_2), and N_1, N_2.

Result: Approximate cluster centre (L_y, L_x)

1 $R_i = \sum_{x \in [x_1, x_2]} H_{i,x}$
2 $C_i = \sum_{y \in [y_1, y_2]} H_{y,i}$
3 $J \leftarrow$ list of pairs (i, R_i) ranked in descending order of R_i
4 $K \leftarrow$ list of pairs (i, C_i) ranked in descending order of C_i
5 $L_y \leftarrow$ mean of indexes i in the top N_1 pairs of J
6 $L_x \leftarrow$ mean of indexes i in the top N_2 pairs of K
7 $L_y \leftarrow L_y / y$
8 $L_x \leftarrow L_x / x$
9 **return** (L_y, L_x)

Output		Template parameters: $(y_1, y_2, x_1, x_2, N_1, N_2)$
Muscle centre	$(m_s, m_t) \leftarrow$	Algorithm 4.1$(8, 80, 7, 60, 6, 6)$
Fat centre	$(f_s, f_t) \leftarrow$	Algorithm 4.1$(1, 64, 64, 128, 6, 10)$

Table 4.3: Template parameters for fat and muscle clusters.

fat cluster than the muscle cluster, owing to its elongated spread in the T1 dimension. Using the average of 6 intensity indexes was found to be sufficient for the other template parameters, for all cases investigated. The averaging process in Algorithm 4.1 means that the algorithm was not overtly misled by a local maximum, distant from the main cluster. The use of column and row sums was inspired by the work of Köse, Gençalioğlu and Şevik [47] (see section 3.8).

4.10 Region of interest refinement (Step 1-H)

Algorithm 4.2 takes the first ROI, as described in section 4.7 (Step 1-E), and iteratively "shrinks" it until all pixels on the edge of the mask are above a heuristic value based on the T1 fat centre f_t (see Table 4.3). Essentially the mask is drawn slightly within the fat layer of the legs, thereby removing partial volume effects, resulting from the background region. An example ROI and the effect of refinement is shown in Figure 4.3.

4.11 Femur removal tool (Step 1-I)

The femur is removed from the slice pairs because the data is assumed to carry no diagnostic value (for IIM). In removing the data contributed by the femur, any possible bias presented is eliminated from further analysis. The reasoning here is

Algorithm 4.2: Refine region of interest

Data: T1 image T, region M, cluster value f_t
Result: Region L

$L \leftarrow M$
$R \leftarrow T < 0.3 f_t$
repeat
    $C \leftarrow L \cap (L \ominus D(1))$
    $X \leftarrow \{x : C(x) = 1 \wedge R(x) = 1\}$
    $L(X) \leftarrow 0$
until $X = \emptyset$
return L

Algorithm 4.3: Weak segmentation process.

Data: Image T, seed point s, percentile rank p.
Result: Femur mask M

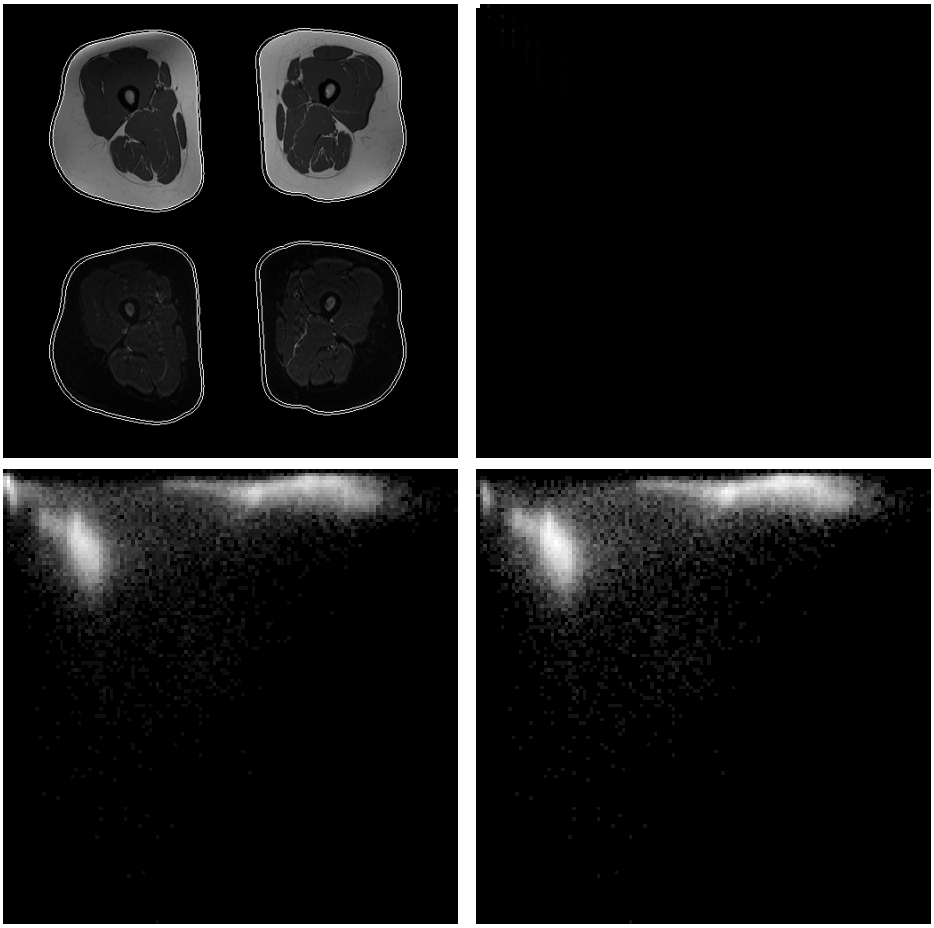
$\eta_{\text{sample}} = \{$ intensities of pixels within a radius of 22.2mm from seed s $\}$.
Find c such that $\dfrac{|\{i : i \in \eta_{\text{sample}}, i \leq c\}|}{|\eta_{\text{sample}}|} \approx \dfrac{p}{100}$.
$N \leftarrow T > c$.
$N \leftarrow N$ with binary reconstruction from seed s.
$N \leftarrow N \oplus D(3)$, where $D(r)$ is a disk shaped structuring element with radius r.
$N_2 \leftarrow N$ with binary reconstruction from point $(1, 1)$, assumed to be background.
$M \leftarrow N \vee \neg N_2$.
return M

simply that by removing data not immediately useful for diagnosis for IIM, the possibility of confounding issues is avoided.

Previously, the femur had been semi-manually segmented for each new scan entered in the training database. These masks were carefully checked to cover the correct region. However, this semi-manual approach was extremely laborious. The original regions will be denoted ω_i, where $i = 1, \ldots, N$.

In general, it was found that thresholding the T1 slice adequately differentiated the femur bone from the surrounding tissues. Given a seed coordinate from the operator, the femur and surrounding region could be extracted. It was also observed that segmentations from adjacent slices were often similar. However, the threshold value was found to be sensitive, as was the exact seed coordinate. Therefore, it was thought that this method was a *weak segmentation process* which could be utilised successfully *en masse*.

Algorithm 4.3 returns a candidate region which is assumed to cover the femur and ∼2.22mm beyond it. The concept is to then trial many points around a seed



a b
c d

Figure 4.3: An example showing the effect on the joint histogram of applying a ROI. (a) is a composite image showing the T1 and STIR input images. The outer contour shows the first ROI as computed in Step 1-E and the inner contour shows the subsequently refined ROI from Step 1-H. (b) shows the resulting joint histogram when no mask is applied, where the signal from background pixels (low signal on both T1 and STIR, as can be seen in the top-left corner) overwhelms the signal from the fat and muscle. (c) shows the resulting histogram with the first ROI applied (Step 1-E). (d) shows the resulting joint histogram with the refined ROI (Step 1-H). Careful comparison of (c) and (d), reveals a reduction in the number of pixels from the fat tissue with hypointense T1, essentially making the definition between fat and muscle clusters more apparent.

location using this weak segmentation process and then rank the candidate regions in a scoring process.

The original masks ω_i were first analysed to better understand the problem domain. The femur centroids and pixel area were computed and statistics were calculated (Table 4.4). Clearly, the femur centroid does not often move more than 3 pixels in any direction between slices. Therefore, it is assumed that a suitable seed point (located on the femur bone) will also tend to be within a certain boundary around the previous seed point (that is, from the prior slice). Anatomically, this is intuitive since the femur bone does not have kinks and slices are 10mm apart.

	L_{Cdiff}	R_{Cdiff}	B_{Cdiff}	Cardinality
Mean	2.483	2.274	2.378	1483
Median	2.221	2.058	2.121	1479
Min	0.209	0.120	0.120	936
Max	6.144	5.507	6.144	2149

Table 4.4: Analysis of femur centroids for the training database. L_{Cdiff} denotes the Cartesian distance between adjacent slices of the left femur centroid. Similarly, R_{Cdiff} refers to the right femur centroid. B_{Cdiff} refers to left and right sets combined. Cardinality is the number of pixels covered (left and right sets).

Algorithm 4.4 makes use of this prior knowledge. An initial seed point s is used for segmentation of the first slice T_1. In instances where the bone is thin, the initial seed point may not be viable (line 2) and another (dark) pixel with proximity is selected. From line 8, the method predicts a region of trial seed points for the current slice based on the prior (successful) seed point. The region defined by Z_1 (line 9) is trialled first (a heuristic based on the results of Table 4.4), which allows the algorithm to finish early for "easy" problems. Region Z_2 is a large fall-back region, used when the number of suitable candidates from Z_1 is low. Given a candidate seed point q, metrics are calculated (line 15 to line 18) on the segmentation C. Clearly bad regions are discarded (line 20), again based on the results of Table 4.4. On line 25, the remaining candidate regions are ranked and the top-scoring region is selected. Figure 4.4 shows two steps in a femur segmentation.

Parameter p was chosen to be 23 and worked in almost all cases. However, in a couple of cases this value was insufficient and $p = 17$ was elected instead. It was noticed that MR scans acquired prior to 2005 were particularly likely to require this adjustment.

With this approach, the weak segmentation process has been made robust to within the expected limits of the training database. Given a larger database, this method may be extended to satisfy further variations in presentation. For instance, line 12 may be expanded to accommodate a larger possible search space, as defined by Z_1 and Z_2 (line 9, line 10). Indeed, the method may be further tailored to specific circumstances. As an example, it could be assumed that any trial seed point with an intensity greater than 0.5 will result in a correct segmentation, and therefore Z_1 and Z_2 may be initially vetted. Further technical optimisations appropriate for the hardware are surely possible. For instance, harnessing graphics processing hardware would allow images to be thresholded in a batch. The overall method is easy to understand and may be augmented for individual requirements.

To evaluate the performance of this tool, the computed regions were compared

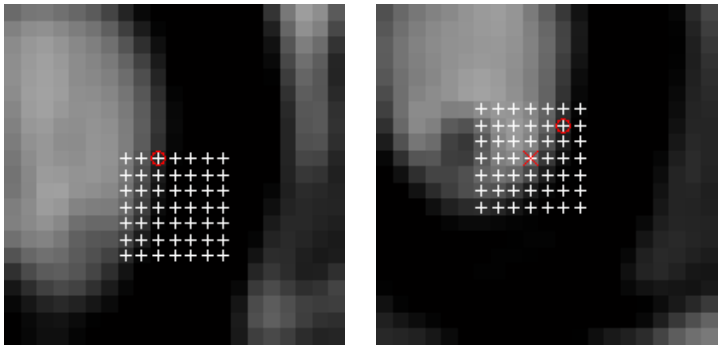
Algorithm 4.4: Segment one femur for a whole set.

Data: T1 slices T_i, $i = 1, \ldots, N$, seed point s, percentile rank p.
Result: Femur masks J_i

$J_1 \leftarrow$ Algorithm 4.3(T_1, s, p)
if $\sum J_1 = 0$ **then**
  $w \leftarrow \{s + x : ||s + x||_2 \leq 2 \wedge x > 0\}$ `// chessboard distance`
  $s \leftarrow$ pixel w with lowest intensity.
  $J_1 \leftarrow$ Algorithm 4.3(T_1, s, p)
  if $\sum J_1 = 0$ **then**
    return $\emptyset$ `// assume bad seed, or try new p`
for $k \leftarrow 2$ **to** N **do**
  $Z_1 \leftarrow \{s + x : ||s + x||_2 \leq 3\}$ `// common trial region`
  $Z_2 \leftarrow \{s + x : ||s + x||_2 \leq 7 \wedge s + x \notin Z_1\}$ `// uncommon trial region`
  Initialise Q as an empty list of candidates.
  for $m \leftarrow 1$ **to** 2 **do**
    foreach point q in Z_m **do**
      $C \leftarrow$ Algorithm 4.3(T_k, q, p)
      $n \leftarrow \sum C$
      $d \leftarrow \frac{2|C \cap J_{k-1}|}{|C|+|J_{k-1}|}$ `// Dice coefficient`
      $g_1 \leftarrow \frac{1}{10}||\text{centroid}(C) - \text{centroid}(J_{k-1})||$
      $g_2 \leftarrow ||\text{centroid}(C) - q||$
      Append to Q the candidate parameters $\{n, d, g_1, g_2\}$.
    $Q \leftarrow Q$ satisfying: $800 < n < 2300 \wedge g_1 \leq 1$.
    if $|Q| > 2$ **then**
      break
  if $|Q| = \emptyset$ **then**
    return $\emptyset$ `// try new p`
  Rank Q by score $d + g_1$, then g_2 (descending,ascending).
  Pick the top candidate, denoted q'.
  $J_k \leftarrow$ Algorithm 4.3(T_k, q', p).
  $s \leftarrow q'$.
return J_i

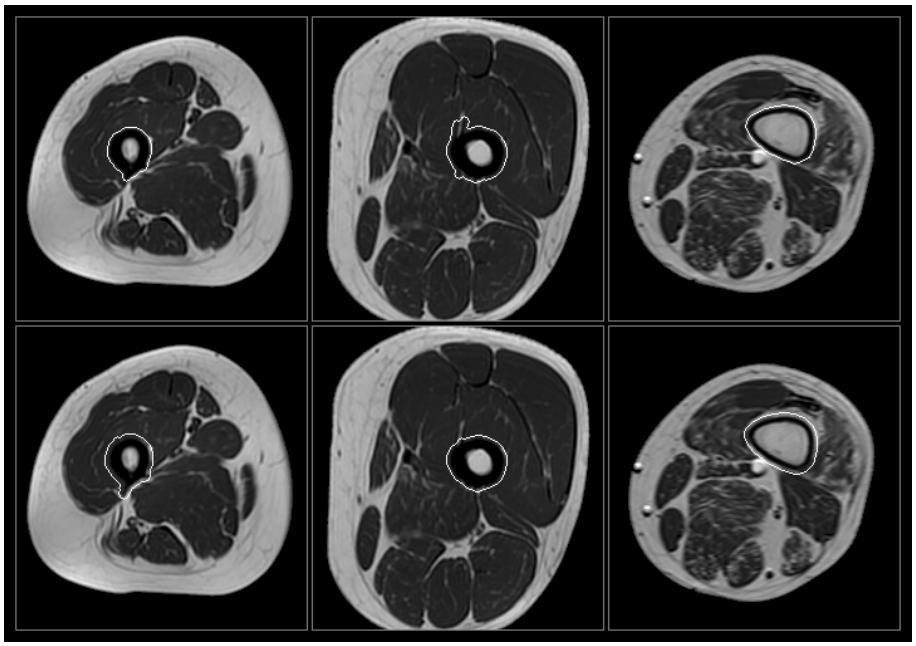
with the original regions ω_i, which are considered a suitable proxy for ground truth. The new computed masks will be denoted ρ_i, where $i = 1, \ldots, N$, and $N = 186$. The test calculated the Dice [13] and Jaccard [36] scores for all T1 slices in the training database. The results are shown in Table 4.5. Note that from visual inspection, it was clear that the new method was just as appropriate as the original masks, given that there is some leeway on exactly what region should be considered for removal (see Figure 4.5). Consider that the main goal of the segmentation is to remove possibly confounding data; not to delineate the exact femur shape.

It is possible to use other methods for the task of bone segmentation. Level set methods, for instance [76], have been applied previously to medical image



a b

Figure 4.4: The second (a) and third (b) steps from a femur segmentation using Algorithm 4.4. Note that (a) is a magnification of Figure 4.5 (a). In both cases the smaller set of Z_1 trial points can be seen as +. The circled point indicates the new chosen seed point. In (b) the $\times$ indicates the previous seed location: the circled point from (a).



a b c
d e f

Figure 4.5: Comparison of old and new femur segmentation methods. (a)-(c) are the original segmentations from ω_i, and (d)-(f) are the new segmentation ρ. Note that (c) was chosen for its difficulty in that the femur bone is both large and thin.

	Dice	Jaccard
Mean	0.95	0.90
Median	0.95	0.90
Min	0.88	0.78
Max	0.99	0.98

Table 4.5: Scores to compare the difference between ground truth (proxy) ω and the proposed method yielding ρ.

segmentation problems. However, the above proposed method is straightforward, easy to implement, suitable for parallelism, produces sufficiently accurate results, and may be easily augmented.

4.12 Vessel removal (Step 1-J)

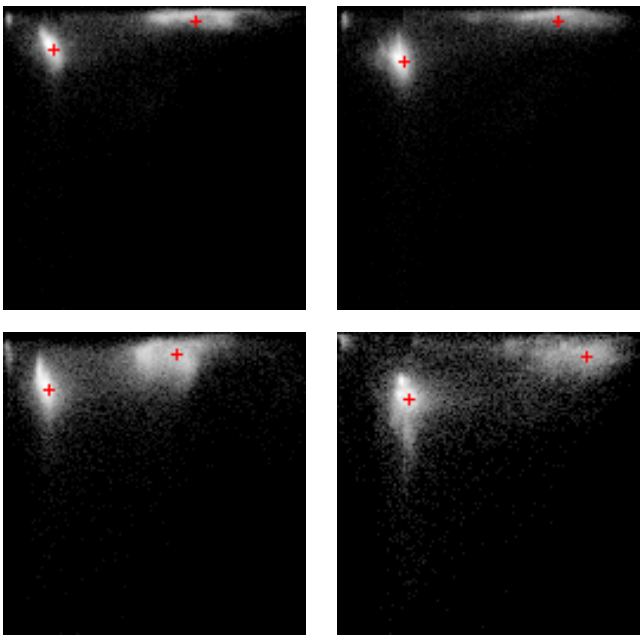
Automatic methods for vessel segmentation were briefly investigated (e.g. [21; 40]), however it was found that the presentation of blood vessels in the STIR sequences were varied and subject to gross imaging artefacts caused by the movement of blood. In some cases, the similarity in appearance between blood vessels and oedema was a major confounding issue.

In a manual process, the two great femoral vessels are masked with the application of a 14.8mm diameter disc placed over the vessel. Anatomically, the vessels are not expected to be bigger than this size.

It is clear this step could benefit from automation, however this aspect is left as a possibility for future improvement.

4.13 Linear intensity scaling (Step 1-K)

By examining many different joint histograms from the training database, it was possible to see that they all had a common theme but with some smaller changes, for instance the shape of the muscle and fat clusters (see Figure 4.6).



a b
c d

Figure 4.6: The appearance of joint histograms can vary. (a), (b), and (d) are control cases. (c) has confirmed disease. Cluster centres have been marked.

As an initial step using image manipulation software, it was possible to manually compare joint histograms between two patients by scaling and translating one joint histogram until it overlapped with the other. This would allow to compare and contrast between scans and patients. The hypothesis was that, by examining enough cases, a general trend could be found to show the disease state in the joint histogram space.

To this end, an automatic method was sought to bring all scans in the database into a common space. This was achieved by using the muscle and fat centres (detailed in section 4.9) and linearly scaling them to pre-defined intensities. Defining T1 and STIR images such that $T = \{t_i : i = 1, \ldots, n\}$ and $S = \{s_i : i = 1, \ldots, n\}$, respectively, where n is the number of pixels, then their intensities were transformed using

$$\lambda_t = \frac{\alpha_t}{f_t - m_t} \qquad \lambda_s = \frac{\alpha_s}{m_s - f_s} \tag{4.18}$$

$$\kappa_t = \beta_t - \lambda_t f_t \qquad \kappa_s = \beta_s - \lambda_s m_s \tag{4.19}$$

$$t'_i = (t_i \lambda_t + \kappa_t) \qquad s'_i = (s_i \lambda_s + \kappa_s) \tag{4.20}$$

$$t''_i = t'_i \text{ clamped to } [0, 1] \qquad s''_i = s'_i \text{ clamped to } [0, 1] \tag{4.21}$$

where subscript s refers to STIR, subscript t refers to T1, α is the desired space between centres, β is the desired transformed centre, and f and m are the fat and muscle centres, respectively (section 4.9).

Note that t'_i and s'_i are referred to as the unclamped intensities, which may fall outside the range $[0, 1]$. Generation of the joint histogram using t''_i and s''_i yields a common intensity space that may be compared to others. The parameters α_t, α_s, β_t, and β_s are chosen to place the tissue centres at appropriate locations in the joint intensity space e.g. $\alpha_s = 0.2$, $\alpha_t = 0.5$, $\beta_s = 0.4$, $\beta_t = 0.65$. An example of this scaling method is shown in Figure 4.7. Note that these parameters are chosen with care to minimise the quantity of data points outside the $[0, 1]$ range.

The ability to compare and contrast between scans and patients is clearly useful. However, there needs to be some discussion on whether it is sensible to linearly scale intensity values in this way.

As was outlined in subsection 3.4.2 it is clear that MR data requires some form of intensity standardisation. However, the method employed must be suitable to the pathology in question (e.g. [63, p. 7]). The method employed by Nyúl and Udupa [55] may not be suitable for this problem domain due to the assumption of undesirable "tails" in the intensity distribution [55, p. 596]. In contrast, for STIR acquisition and the myositis pathology, it is generally the high-intensity end of the muscle distribution (with relatively low frequency) which helps indicate the

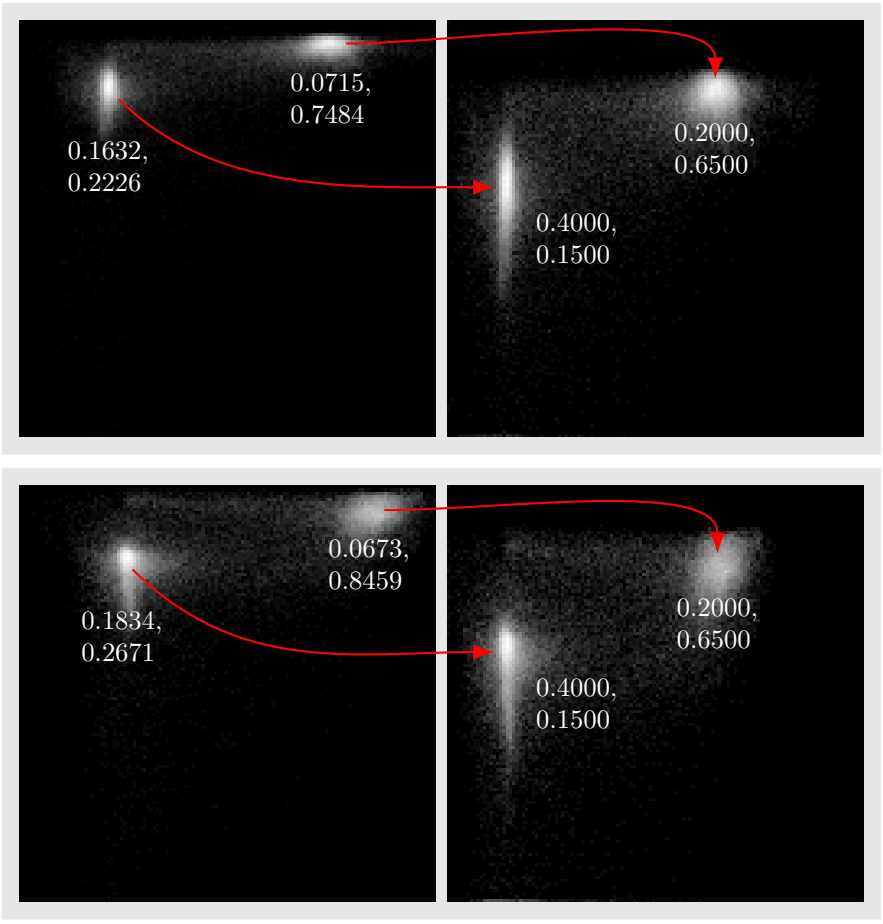


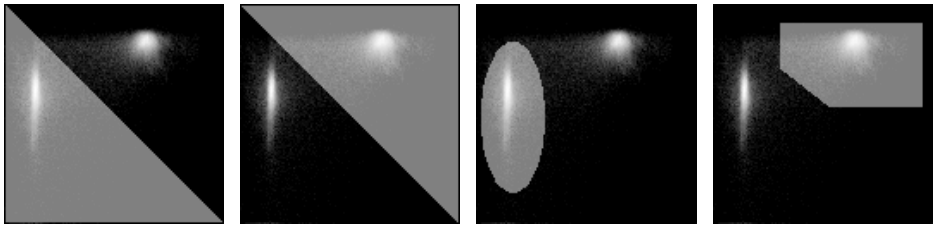
Figure 4.7: Example of scaling joint histograms. Values shown are (STIR,T1) intensities as transformed by Equation 4.21. (a) shows the transformation before and after for a control subject, and (b) shows the same for a symptomatic case.

disease activity (oedema). Therefore it can be seen that Equation 4.20 provides a suitable standardisation method for this problem domain.

4.14 Static segmentation (Step 1-L)

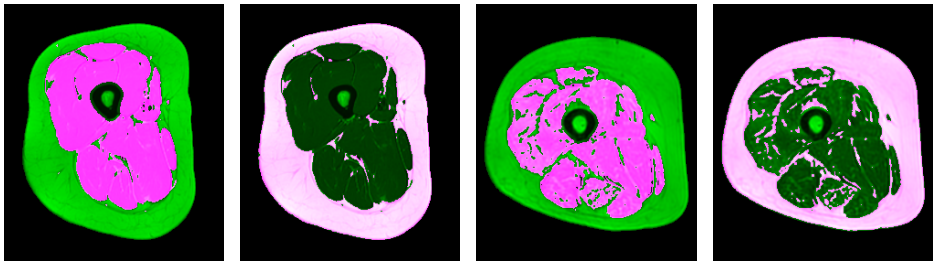
Since the joint histograms are now in a common space it is possible to select regions using static masks (see Figure 4.8). Effectively this is segmenting the image based on the joint T1 and STIR standardised intensities. In the majority of cases the tissues are segmented well (e.g. Figure 4.9). The joint histogram masks allow selection of the tissues within certain bounds specified by the masks.

Since the muscle tissue is of primary interest in the following chapters, the set of locations x considered as muscle will be denoted M henceforth.



a b c d

Figure 4.8: (a) and (b) show segmentations of the intensity space into the two tissues of muscle and fat, respectively. These are used in the proposed system. (c) and (d) show the possibility of arbitrarily defined masks, where the tissue selection criteria can be more refined if required.



a b c d

Figure 4.9: (a) and (b) show muscle and fat segmentations for an asymptomatic control subject. (c) and (d) show the same segmentation process for a patient with acute and chronic disease state (note the quantity of fat within the muscle).

4.15 Discussion

The pre-processing steps have been designed to be as rigorous and robust as possible, with respect to the pathology in question. In general the steps work without manual intervention for the training and out of sample databases. However, in some instances unexpected behaviour could occur.

Registration Step 1-C may fail if either of the two input slices have severe MR artefacts, including intensity inhomogeneity as a result of the bias field. If the subject moved during acquisition, then translation-only registration may be inadequate and pre-processing Step 1-G to Step 1-L would be affected. Geometric scaling in Step 1-D depends upon the correct parameter for DICOM key 'Pixel Spacing'. ROI selection in Step 1-E could fail if the T1 input has unexpected geometry or unusual background intensity. Localisation of tissue centres in Step 1-G may not work if the intensity distribution of tissues are extremely dispersed or chaotic. Intensity scaling (Step 1-K) and adequate segmentation (Step 1-L) depend upon the expected approximate location of tissues within the intensity space.

In general, the pre-processing steps assume that the correct MRI protocol has been followed, and that the subject has the normally expected anatomy.

4.16 Conclusion

In this chapter, the necessary steps to pre-process the MRI scans have been described. A semi-automatic segmentation process was designed to remove the femur, based on the empirical data analysis of femur centroids, which was shown to be comparable to a more labour-intensive process. A process for intensity standardisation, suitable for the disease in question, has been introduced and it has been shown to work via visual analysis of the joint intensity space. Importantly, this means all scans are in a common frame of reference, which will allow generalised observations and comparisons to be made in the next chapter, leading to disease quantification methods.

Chapter 5

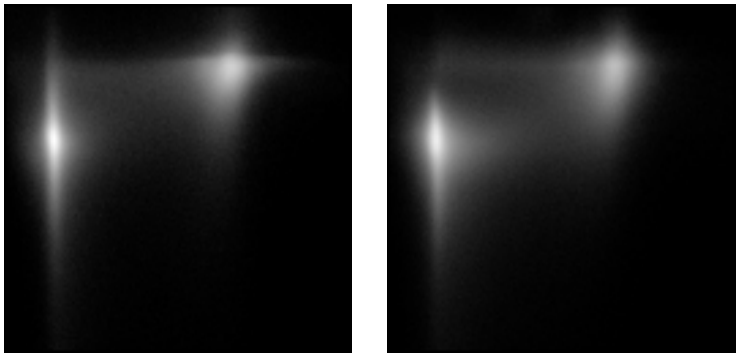
Quantification of disease

In this chapter, the severity of oedema and fat infiltration is quantified such that the disease state of a patient may be systematically measured. Since the level of active disease is of immediate clinical interest, a method to quantify the severity of oedema is presented first. A similar method is applied for the quantification of fat infiltration.

5.1 Analysis of STIR intensity distribution for oedema by cohort

By considering the standardised joint histogram from all subjects meeting a particular criteria, it is possible to learn what is expected in the *average* case and its visualisation in the joint intensity space. With the benefit of this knowledge, an approach is constructed based on the expected distribution shape.

Figure 5.1 shows the average joint histogram for all subjects in the training database split into two cohorts. The first cohort is comprised of control subjects

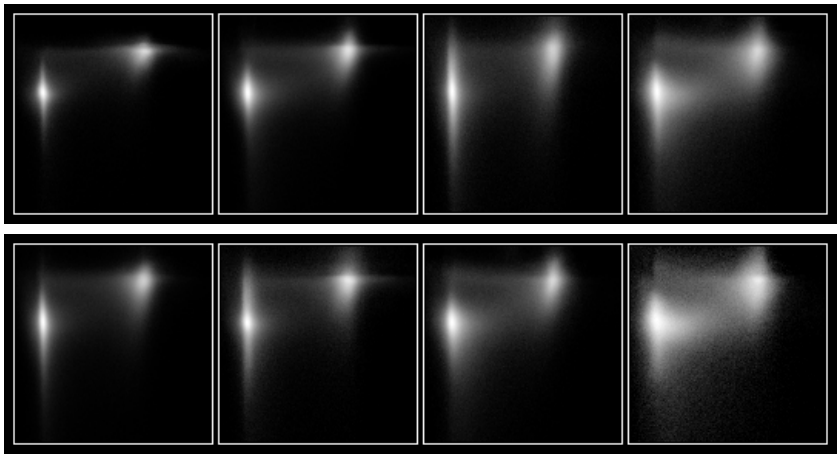


a b

Figure 5.1: Comparison of the average joint histogram for different cohorts in the training database. (a) is the average joint histogram of control subjects (n=9), and (b) is the average joint histogram of all symptomatic patients (n=22).

(volunteers assumed to be healthy) which represents what to expect in the absence of the disease. The second cohort, "symptomatic", is comprised of patients from the training database with confirmed oedema (i.e. acute disease state). The average joint histogram is constructed by concatenating $(\mathbf{s}''(x), \mathbf{t}''(x))$ intensity pairs from all pixels in the final ROI, for all slices, for all cases in the group under consideration. 128×128 bins are used in the range $[0, 1]$ and the joint histogram is normalised such that the area under the surface is 1. In Figure 5.1 it is clear that there is a difference in the joint histograms between cohorts. Generally, it may be noted that the joint histogram for the control cohort is well-defined, with one single value generally dominating the tissue clusters (in each dimension). In comparison, the symptomatic cohort has poorly defined tissue clusters.

Figure 5.2 shows average joint histograms grouped by disease state. It is assumed that the disease causes physical changes in the tissues that are represented in the joint intensity space. This result has been used to guide the development of disease specific metrics.



a b c d
e f g h

Figure 5.2: Comparison of the average joint histogram, grouped by oedema and infiltration grade. (a)-(d) are average joint histograms for subjects with increasing oedema. (e)-(h) are average joint histograms for subjects with increasing infiltration. Note that each subject may have some degree of oedema and fat infiltration, so increasing severity in one state is not completely separable from the other.

Figure 5.3 is an analysis of the distribution for muscle tissue only. (a) suggests that the presence of oedema does lead to a higher frequency of hyperintense pixels. The effect can be seen clearly in Figure 5.4. This observation forms the basis for selecting oedema in STIR images. It would be desirable to model the distribution parametrically, and from the literature, Bartlett et al. [3] fitted a Guassian curve to model the healthy and diseased tissue in the STIR histogram (see section 3.5).

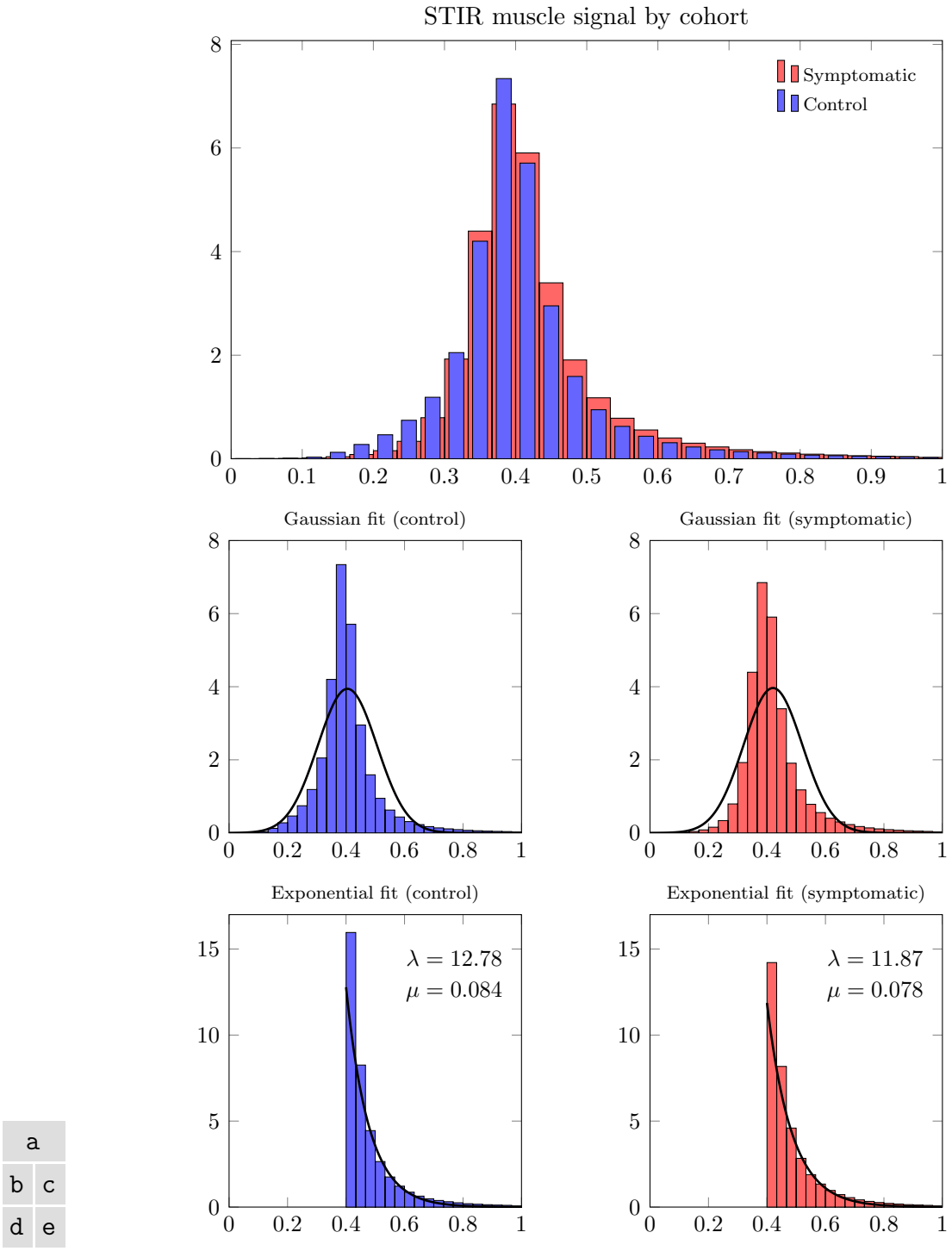
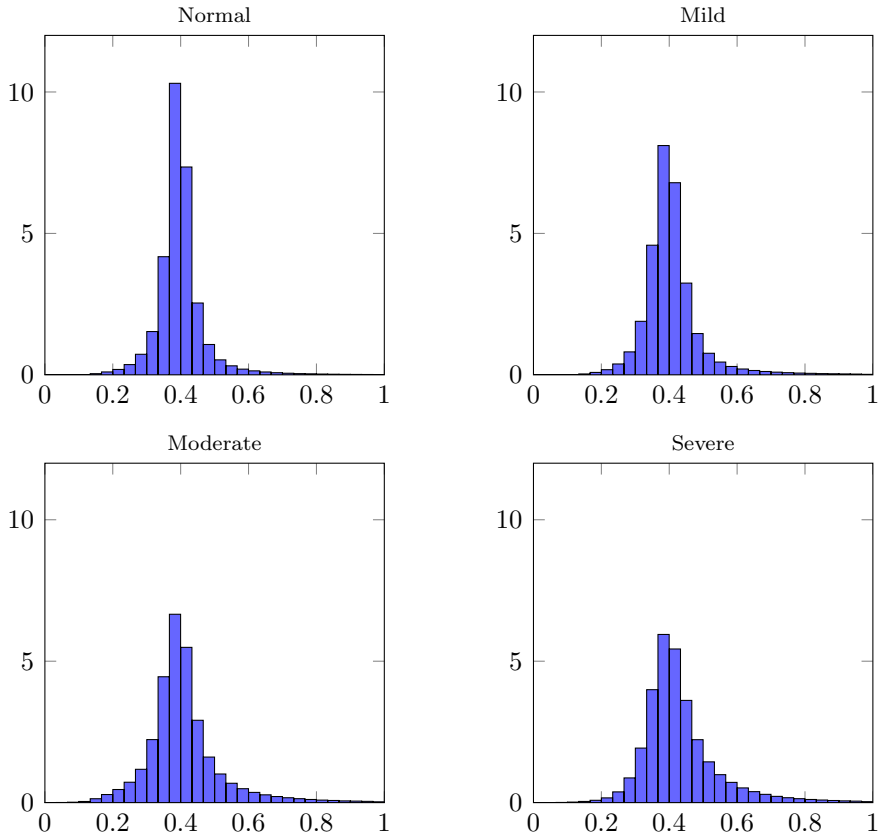


Figure 5.3: STIR muscle tissue histograms in the average case grouped by cohort. All histogram data has been normalised. (a) is a comparison for all STIR muscle signal. (b) and (c) show a Gaussian fit of the data for control and symptomatic cohorts, respectively. (d) and (e) show the right-hand tail of the distribution ($s'' > 0.4$ in this example) extracted and modelled with an exponential distribution.



a b
c d

Figure 5.4: Histograms of STIR signal for muscle tissue by disease grade. (a)-(d) show increasing thickness on the upper tail.

If X denotes a random variable with mean μ and standard deviation σ, then it is assumed the distribution of tissue intensities I_T is approximated by

$$I_T \approx X \sim \mathcal{N}(\mu, \sigma^2) \tag{5.1}$$

where $\mathcal{N}$ is the normal distribution and has Probability Density Function (PDF)

$$P_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.2}$$

Figure 5.3 (b) and (c) show a Gaussian distribution fitted to the muscle intensities for control and symptomatic cohorts, respectively. It is clear that, for the data in the training database, the tissue is not adequately modelled by the Gaussian, particularly in the right-hand region.

Another approach may be taken. Since it is known that acute disease (oedema) gives hyperintense signal (i.e. the upper tail of the distribution) then it is logical to conclude that only the upper tail of the distribution needs to be modelled. Taking a threshold of the intensities (beyond the main peak) would result in a histogram with gradually declining frequencies. It was assumed that this part of

the distribution could be modelled effectively with an exponential distribution. Again, if X denotes a random variable, then it is assumed the distribution of tissue intensities I_T is approximated by

$$I_T \approx X \sim \mathbf{Exp}(\alpha, \lambda) \tag{5.3}$$

where **Exp** is an exponential distribution, α is the location, and λ is the rate parameter (the reciprocal of standard deviation). The exponential distribution has PDF

$$P_{\alpha,\lambda}(x) = \lambda e^{-\lambda(x-\alpha)}, \text{ for } x \geq \alpha. \tag{5.4}$$

Figure 5.3 (d) and (e) show the histogram for STIR intensities above 0.4 with an exponential distribution fitted. It was found that any threshold value marginally beyond the main tissue peak can be sufficiently modelled with the exponential distribution. However, the value 0.4 was chosen since this is the assumed muscle peak after intensity standardisation (see section 4.13). Table 5.1 shows a comparison of the R^2 statistic between the two different models.

Therefore this analysis has indicated a suitable approach to take in selecting acute disease (oedema), using an exponential distribution as the basis.

R^2	Control	Symptomatic
Gaussian (whole)	0.769	0.768
Exponential (tail)	0.889	0.912

Table 5.1: Comparison of goodness of fit for Guassian and exponential distributions for STIR muscle intensity distributions.

5.2 Method for oedema selection

The standardised z-score is used to transform the STIR muscle intensities. The z-transform takes the form

$$z = \frac{x - \mu}{\sigma} \tag{5.5}$$

where x is the original value, μ is the mean, σ is the standard deviation, and z is the transformed value. In the case of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ then the z-score method transforms the data such that it follows the standard normal distribution $\mathcal{N}(0, 1)$. The z-score may be thought of as the number of standard deviations away from the mean of the distribution. This process is used to select a candidate region of oedema that may then be quantified.

The z-transform will now be considered for an exponential distribution. If X

is a random variable following an exponential distribution, denoted

$$X \sim \mathbf{Exp}(\alpha, \lambda), \text{ then} \tag{5.6}$$

$$(X - \alpha) \sim \mathbf{Exp}(0, \lambda). \tag{5.7}$$

The PDF for this distribution is

$$P_{0,\lambda}(x) = \lambda e^{-\lambda x}, \text{ for } x \geq 0. \tag{5.8}$$

The mean and standard deviation are defined as

$$\mu = \sigma = \lambda^{-1} = \frac{1}{N} \sum x_i \tag{5.9}$$

where N is the number of elements. Note that Equation 5.9 may be rearranged to give

$$\lambda = \frac{N}{\sum x_i}. \tag{5.10}$$

The z-transform for an exponential distribution may then be calculated as

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} = \frac{(x - \frac{1}{\lambda})}{\frac{1}{\lambda}} = \lambda(x - \tfrac{1}{\lambda}) \\ &= \lambda x - 1. \end{aligned} \tag{5.11}$$

Now the z-score method will be applied to the problem of selecting a threshold value which is consistent with a range of values for rate parameter λ. If M is the set of pixel locations that represent the muscle which was found in the static segmentation Step 1-L (section 4.14) then a mask of pixels whose intensities will be modelled is given by

$$\mathbf{g}_1 = \big\{(x, c(x)) : c(x) = 1 \text{ if } x \in M \wedge s''(x) > \gamma_T, \text{ otherwise } c(x) = 0\big\} \tag{5.12}$$

where γ_T is the intensity threshold, set to 0.4, and $\mathbf{s}''$ is the standardised STIR image, as computed in Step 1-K (section 4.13). Using Equation 5.10, the rate parameter is calculated as

$$\lambda = \frac{\sum_x \mathbf{g}_1(x)}{\sum_x \left[\left(\mathbf{s}''(x) - \gamma_T\right) \mathbf{g}_1(x)\right]} \tag{5.13}$$

and the z-score criteria is applied with

$$\mathbf{z}_1 = \left\{(x, c(x))\right\},$$
$$c(x) = \begin{cases} 1 & \text{if } g_1(x)[\lambda\left(\mathbf{s}''(x) - \gamma_T\right) - 1] > g_1(x)\gamma_Z \\ 0 & \text{otherwise,} \end{cases} \tag{5.14}$$

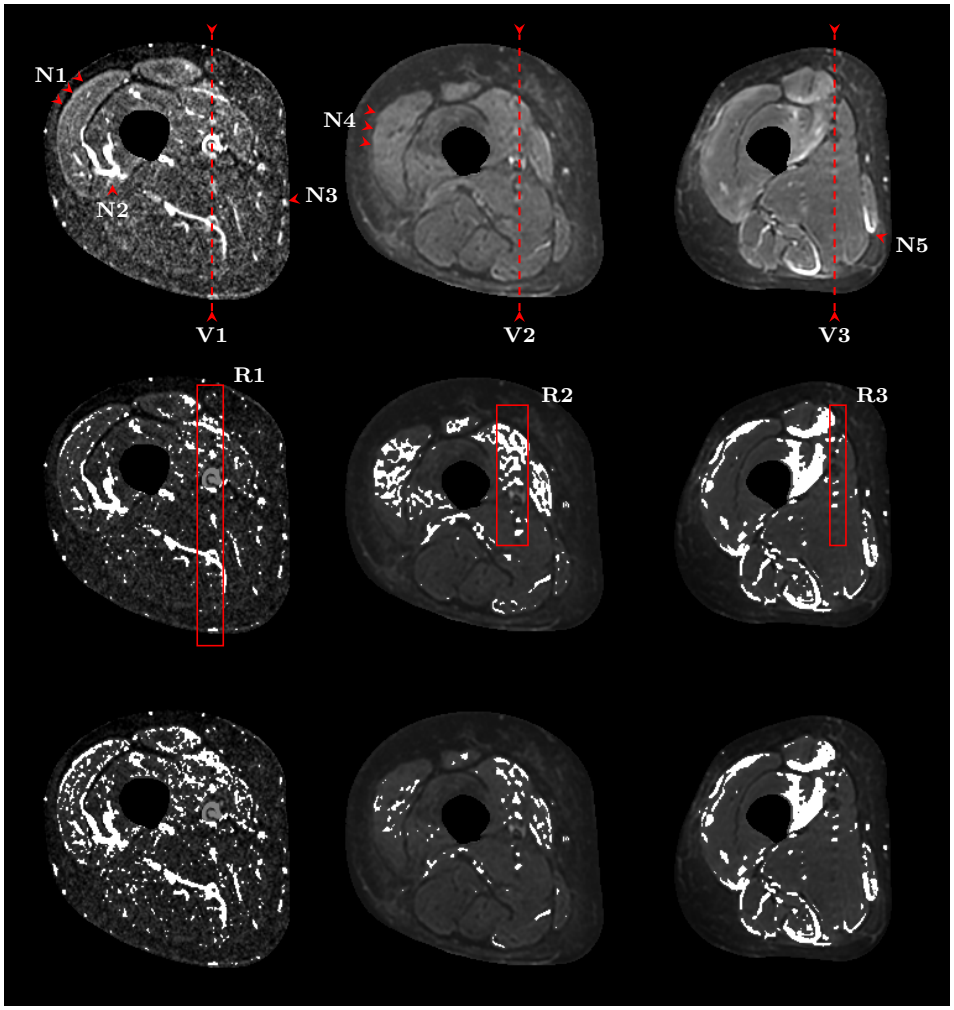
where γ_Z is a threshold value for the z-score, set to 0.1. Note that the actual threshold value for a particular case may be found by adjusting Equation 5.11 to account for threshold γ_T so that

$$x = \frac{\gamma_Z + 1}{\lambda} + \gamma_T. \tag{5.15}$$

When applied to the whole training database, the above method calculates a thresholded value per STIR slice. The minimum threshold value was 0.431 and the maximum was 0.624. This indicates the range of values taken by λ. Figure 5.5 shows the method applied to 3 cases from the training database. Comparison of the method against a simple constant threshold method is provided in the last row.

5.3 Region improvement

It is clear that whatever method is used to determine a threshold value, it is conceivable that any arrangement of MR artefacts may be mistaken for oedema or otherwise disguise it. For instance, Figure 5.5 (a) illustrates a control subject, but the image is somewhat noisy and has appreciable bias field effects in the upper half (residual from Step 1-F) which could be mistaken for acute disease. Further, a large transverse blood vessel may yield hyperintense signal which is not representative of the disease. In comparison, the symptomatic subject in (b) has confirmed acute disease but the oedema has a softer appearance with less hyperintense signal than (a). Figure 5.5 (c) shows how oedema (see arrow N5) could look like a blood vessel (or artefact). Similarly, arrow N2 shows a blood vessel that could be mistaken for oedema (or artefact). Therefore some verification must be obtained from an experienced operator that the candidate region $\mathbf{z}_1$ does not cover large regions of artefact. Some highlighting due to artefact is expected, and cannot be realistically avoided, but so long as this remains at a baseline level then the overall disease metric (discussed in the next section) is assumed to be valid. Due to the z-transform applied in Equation 5.14, it is generally possible to over-



a b c
d e f
g h i

Figure 5.5: A STIR slice from 3 separate cases in the training database are shown in the first row. (a) is a control subject. (b) and (c) have confirmed acute disease (active inflammation) status. The middle row shows the application of the method from section 5.2 used on the slice above. The white region indicates $\mathbf{z}_1$, with the faded slice behind. The last row illustrates a fixed threshold method, which may be compared to the prior result. In (g), too much signal has been selected in the control subject. In (h), the wide-spread oedema (indicated by N4) has not been selected. (i) uses the same threshold value as (f). **Notations**: V1, V2, and V3 indicate the phase encoding direction and passes through the main blood vessel present in the slice (it is dark for V3). Ghosting artefacts can be seen along V1. Minor signal artefacts can be noticed in V2 and V3. The bright edge of N1 is an MR artefact and does not represent oedema. N2 is assumed to be a transverse blood vessel. N3 is an example of a blood vessel or MR artefact in the fat layer. N4 is an example of broad oedema. N5 is an example of concentrated oedema. R1, R2, and R3 indicate regions of $\mathbf{z}_1$ that are partially due to the vessel pulsation artefacts (V1,V2,V3).

estimate the candidate region but more difficult to under-estimate it. Thus, the main scenario which must be considered is the removal of non-indicative regions from $\mathbf{z}_1$.

From visual inspection of Figure 5.5, it can be seen that the presentation of myositis is generally of interest when it occurs in groups, e.g. (e) and (f). Small, individual objects in $\mathbf{z}_1$ tend not to indicate the presence of the disease but rather a poor SNR, e.g. (d). Large objects in $\mathbf{z}_1$ tend to indicate some level of oedema within a muscle.

Consider a new region $\mathbf{z}_5$, identical to $\mathbf{z}_1$ but where objects with fewer than 5 pixels have been removed. For noisy STIR images this reduces the quantity of erroneous oedema pixels, however, the trade-off is that oedema may then be under-estimated in subjects with acute disease. Since there is no data to provide ground truth (with pixel-level granularity), it is only possible to compare overall results from a series of experiments.

A number of different oedema regions will be formed from the initial region $\mathbf{z}_1$,

$$\mathbf{z}_n = \mathbf{z}_1 \text{ where objects with } < n \text{ pixels are excluded, for } n > 1. \tag{5.16}$$

A suitable sample of the set of possible regions will be used in a test in section 5.5.

5.4 Definition of the oedema metric

A naive quantification may be formed with

$$\Psi_{\text{Oedema}} = \sum \mathbf{z}_i,$$

which yields the number pixels thought to represent disease for some region $\mathbf{z}_i$. In pre-processing, the input slices were scaled such that each pixel represents 0.74mm in each dimension. However, it is natural that subjects have a range of physical leg size. Given that all subjects (control and symptomatic) will have at least some baseline level of oedema, then subjects with a greater cross-sectional area of muscle will tend to have greater values of Ψ_{Oedema}. Clearly, the metric must be normalised to account for the quantity of muscle present (not the overall number of leg pixels, which would also contain fat tissue). The equation is re-written

$$\Psi_{\text{Oedema}} = \frac{\sum \mathbf{z}_i}{\sum M} \tag{5.17}$$

where M is the set of muscle pixels, as before. Ψ_{Oedema} now represents the percentage of muscle pixels thought to be oedema.

5.5 Correlation of the oedema metric with acute disease

Since the cases in the training database have been graded by a medical expert for disease severity, it is possible to measure the correlation between the metric and expert opinion. (The ordinal scale is shown in Table 2.1, page 17). The "expert grade" is used as a proxy for ground truth, albeit with relatively coarse granularity when compared to a percentage metric. For the metric to be useful, it must have positive correlation with the expert grading. However, deviations are expected. It is generally understood that there is natural variation in expert grading. For example, if a test were performed with 10 experts grading a set of scans, the variation in the expert grade could be considerable. If the test were repeated, a given expert may not reproduce the same grades (see Herlidou et al. [30]). Fischmann et al. notes that clinically validated scales are subject to poor inter- and intra-rater repeatability, and could be influenced by such factors as fatigue and cooperation [19, p. 1648].

Figure 5.6 illustrates the correlation between the expert grade and oedema metric Ψ_{Oedema} using a set of sample regions $\{\mathbf{z}_1, \mathbf{z}_5, \mathbf{z}_9, \mathbf{z}_{13}\}$, generated using Equation 5.16. Statistics relating to this test are given in Table 5.2. The results show an increase in the correlation coefficient and R^2 statistics up to z_9, meaning that removing some small objects is indeed useful. z_9 may be selected on the basis that both the correlation coefficient and R^2 gain their highest value for this region. Results for z_{13} indicate that objects with fewer than 13 pixels (but more than 8) meaningfully contribute to the Ψ_{Oedema} metric and therefore should remain.

Region	m	c	Pearson corr.	R^2
z_1	0.0226	0.0965	0.6506	0.4233
z_5	0.0231	0.0885	0.6651	0.4424
z_9	0.0236	0.0804	0.6727	0.4526
z_{13}	0.0238	0.0733	0.6707	0.4499

Table 5.2: Statistics for the the oedema metric using different regions of $\mathbf{z}_i$. m and c are the slope and intercept, respectively, of the linear regression line. Pearson corr. is the linear correlation coefficient.

In Figure 5.6 it is clear that values for the Ψ_{Oedema} metric overlap between grades. One possible reason for this is that only one set of data from a medical expert was available. The variance of a single rater would have the effect of "blurring" the class boundaries due to variation from the (unknown) ground truth, although the overall trend would still be evident. A second possible reason for this effect is due to the interaction between the "patchy" nature of the disease and the

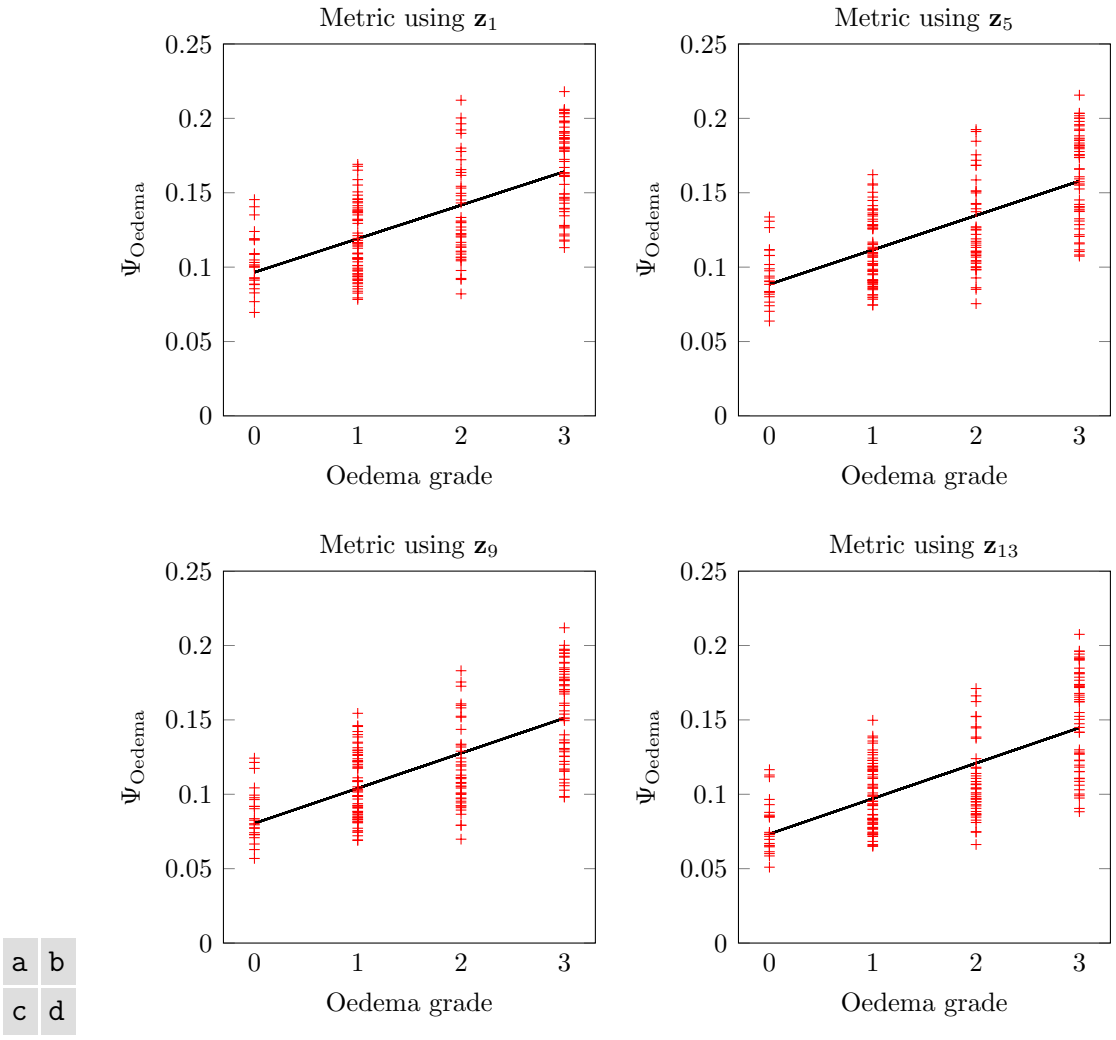


Figure 5.6: Linear regression between the Ψ_{Oedema} metric and expert grade using regions $\{\mathbf{z}_1, \mathbf{z}_5, \mathbf{z}_9, \mathbf{z}_{13}\}$.

fact that only one expert grade is assigned per subject. Since there is a total of 6 slices for each subject and the quantity of disease is not constant throughout the leg, then there is a natural spread of Ψ_{Oedema} values for every subject.

5.6 Towards validation of the oedema metric

Good correlation of Ψ_{Oedema} with the expert grade is an encouraging sign. However, further testing will help indicate the level of robustness. The following three scenarios must be considered to help validate the metric.

1. For a control subject who is rescanned in the same session, the metric must have a similar value.

2. For a subject with confirmed oedema who is rescanned in the same session, the metric must have a similar value. A longer period of time may mean a

real change in the concentration of oedema.

3. For a patient with repeat scans over a long period of time, the change in Ψ_{Oedema} between scans must be seen to correlate with changes in disease state.

Given the number of subjects available in the re-scan database, the following tests cannot be considered a full validation. They will however strongly indicate the potential of the method. The above scenarios will now be dealt with in turn.

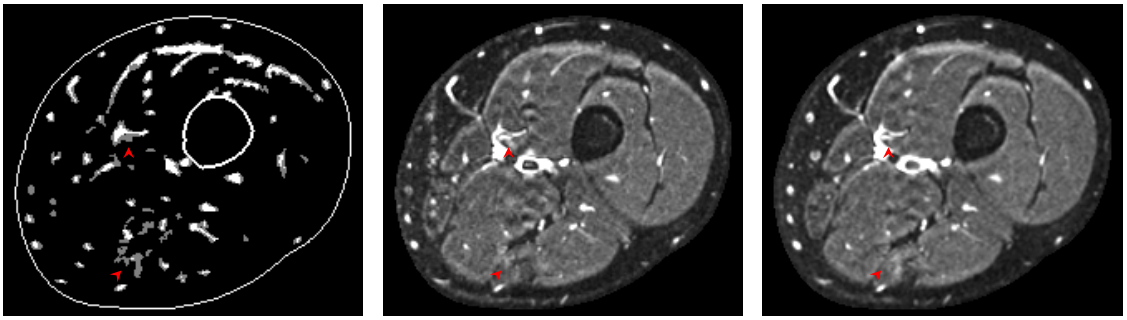
5.6.1 Rescanned control

A control subject was rescanned in the same session, using the same slice locations. Table 5.3 shows the test results. Comparing values of T_1 and T_2 it can be seen that the total number of pixels in the muscle region remained stable. The difference between Ψ_{Oedema} values has a minimum of 0.02% and a maximum of 1.04%. However, taking an average of all Ψ_{Oedema} per scan, the mean values are 9.3% and 9.1% for first and second scans, respectively. The worst match in the set is highlighted and shown in Figure 5.7. In general, the appearance between scans is identical, however imaging artefacts resulting from vessel pulsation (vertically above and below the blood vessel) do have some effect on the metric.

O_1[a]	O_2[a]	T_1[b]	T_2[b]	Ψ_1[c]	Ψ_2[c]	$\lvert\Psi_1 - \Psi_2\rvert$	Jaccard	Dice
3570	3639	41197	41470	0.0867	0.0878	0.0011	0.5325	0.6950
3791	3381	40082	39886	0.0946	0.0848	0.0098	0.3972	0.5686
3723	3399	38481	38395	0.0967	0.0885	0.0082	0.5416	0.7026
3283	3647	36258	36136	0.0905	0.1009	0.0104[d]	0.3485	0.5169
3699	3665	35247	34985	0.1050	0.1048	0.0002	0.5539	0.7129
2926	2543	33661	33421	0.0869	0.0761	0.0108	0.5384	0.7000
				0.0934[e]	0.0905[e]			

[a] Quantity of oedema, $O_i = \sum \mathbf{z}_9$, subscripts denote scan 1 and 2.
[b] Quantity of muscle tissue, $T_i = \sum M$, for scan 1 and 2.
[c] $\Psi_i = \Psi_{\text{Oedema}}$, for scan 1 and 2.
[d] Denoted as the worst match in the set. Shown in Figure 5.7.
[e] Mean of the values above.

Table 5.3: Data table for rescanned control. The six rows in the table correspond to the six slice locations (identical for both scans).



a b c

Figure 5.7: Diagram of the worst match for the rescanned control case. (a) is an overlap digram showing the two $\mathbf{z}_9$ regions. The solid white contour shows the final ROI. Intersection of the two regions is shown in white. (b) is the first scan and (c) is the second. The view window has been cropped to show detail. Arrows indicate regions with different intensity, possibly due to vessel pulsation.

5.6.2 Rescanned acute disease

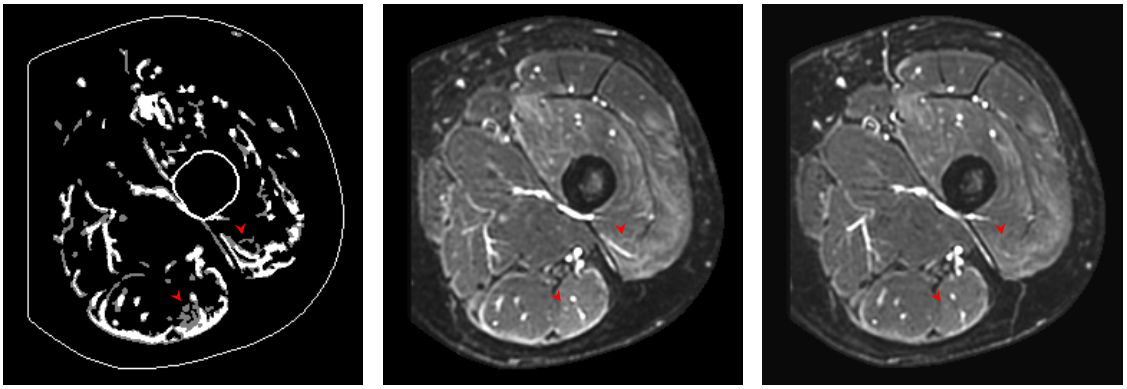
A subject with confirmed oedema was rescanned during the same session. Note that slice locations remained the same. Table 5.4 shows the data from the test on the two regions of $\mathbf{z}_9$.

O_1[a]	O_2[a]	T_1[b]	T_2[b]	Ψ_1[c]	Ψ_2[c]	$\lvert\Psi_1 - \Psi_2\rvert$	Jaccard	Dice
8497	9344	54093	53994	0.1571	0.1731	0.0160	0.4934	0.6607
8870	6651	54421	53932	0.1630	0.1233	0.0397[d]	0.4830	0.6514
6725	7046	54088	53642	0.1243	0.1314	0.0070	0.3750	0.5455
7743	6153	53571	53212	0.1445	0.1156	0.0289	0.3342	0.5010
7473	6472	51526	50951	0.1450	0.1270	0.0180	0.4266	0.5981
6262	5275	48886	48563	0.1281	0.1086	0.0195	0.4250	0.5965
				0.1437[e]	0.1298[e]			

[a] Quantity of oedema, $O_i = \sum \mathbf{z}_9$, subscripts denote scan 1 and 2.
[b] Quantity of muscle tissue, $T_i = \sum M$, for scan 1 and 2.
[c] $\Psi_i = \Psi_{\text{Oedema}}$, for scan 1 and 2.
[d] Denoted as the worst match in the set. Shown in Figure 5.7.
[e] Mean of the values above.

Table 5.4: Data table for rescanned acute disease case. The six rows in the table correspond to the six slice locations (identical for both scans).

In this test, the difference in values of Ψ_{Oedema} between scans has a minimum of 0.7% and a maximum of 3.97%. The worst match in the set was judged by the absolute difference of Ψ_{Oedema} between scans. This slice location from both scans is shown in Figure 5.8. Notice that (b) and (c) are nearly identical except for slight, yet distinct, differences in intensity. This has caused a change in the $\mathbf{z}_9$ region and therefore decreased the Jaccard score (Table 5.4). However, the location of the $\mathbf{z}_9$ region is similar in both scans, supported by a relatively high



a b c

Figure 5.8: Diagram of the worst match in the rescanned acute disease case. (a) is an overlap digram showing the two $\mathbf{z}_9$ regions. The solid white contour shows the final ROI. Intersection of the two regions is shown in white. (b) is the first scan. (c) is the second scan. The view window has been cropped to show detail. Notice that (b) has somewhat pronounced oedema compared to (c). Two specific changes are shown at the arrows.

Dice score.

Taking an average of all Ψ_{Oedema} per scan, the mean values are 14.4% and 13.0% for first and second scans, respectively. It is difficult to interpret the meaning of the decrease, since the test would ideally represent no change in the acute disease state. The change in intensity could be some artefact of the MRI machine. However there are changes in intensity, which can be seen visually. The decrease in Ψ_{Oedema} is in line with that visual perception.

5.6.3 Longitudinal rescans

In the re-scan database there are 2 subjects who were scanned on 3 separate occasions, and 1 subject was scanned on 2 different occasions. These scans have been used to identify the change in disease over time.

Summarised data for the test is shown in Table 5.5 for each of the re-scanned subjects. The data is also plotted in Figure 5.9 so that the change over time is more apparent.

Although there is too little data to perform statistical analysis, it is worth noting that in the cases available, not only does the average metric Ψ_{Oedema} (for 6 slices) appear to be useful, but so does the range of Ψ_{Oedema}.

Since the scans have been taken at different times, it is not possible to compare oedema regions ($\mathbf{z}_9$) directly. However, the level of disease can be visually assessed. Figure 5.10 shows sample slices from the three cases.

Case	Mark	Day	Average	Min (error)	Max (error)
1	$\oplus$	0	0.2274	0.0172	0.0169
1	$\oplus$	179	0.1136	0.0162	0.0279
1	$\oplus$	298	0.1133	0.0144	0.0202
2	$\times$	0	0.1177	0.0149	0.0225
2	$\times$	122	0.1447	0.0200	0.0204
2	$\times$	321	0.1157	0.0072	0.0140
3	$-$	0	0.1138	0.0163	0.0231
3	$-$	119	0.1087	0.0078	0.0090

Table 5.5: Data table for longitudinal rescans. This data is plotted in Figure 5.9.

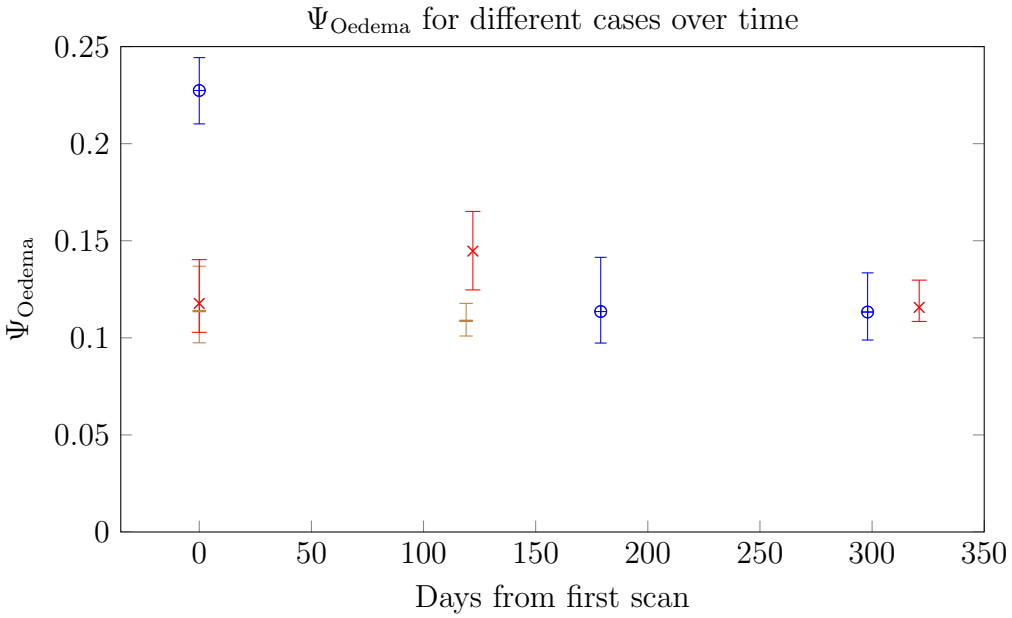
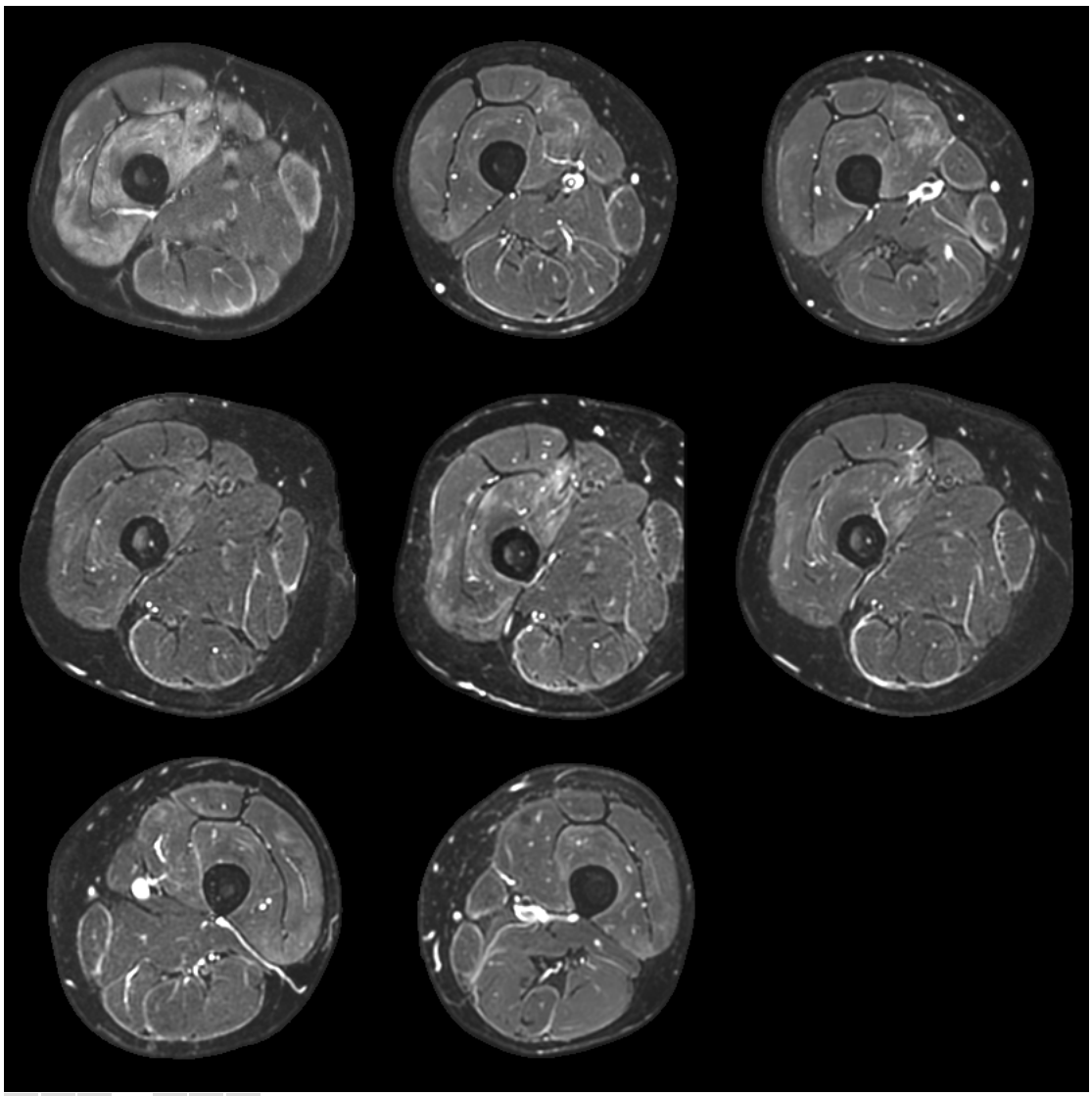


Figure 5.9: Oedema metric for longitudinal cases. Cases are marked ($\oplus, \times, -$) for cases 1, 2, and 3, respectively. The lower error bar is the minimum. The upper error bar is the maximum. The mark shows the mean of all 6 slices.

5.6.4 Discussion

In section 5.1 it was found that an exponential distribution may be used to model the upper tail of the STIR distribution, where acute disease is expressed. The automatic selection of the diseased region $\mathbf{z}_1$ was detailed (section 5.2) and an improvement to the region suggested in section 5.3. The Ψ_{Oedema} metric was formed (section 5.4) and found to have positive correlation with expert grading (section 5.5) and the best refined region was found to be $\mathbf{z}_9$. In this section (section 5.6), results have been presented that strongly suggest that the Ψ_{Oedema} metric is sufficiently robust to track changes in disease state. With consideration to



a	b	c	⊕	⊕	⊕
d	e	f	×	×	×
g	h		−	−	

Figure 5.10: Examples for multiple time-points for longitudinal cases. (a)-(c) are a single slice from time-points of the case marked ⊕ (in order, left to right). Similarly, (d)-(f) are single slices for case marked ×. (g) and (h) are single slices for case marked −. The visual assessment for each subject through time is commensurate with the time-series profile of Figure 5.9. Notice: (a) is clearly the worst case and has the highest (average) metric value of 0.2274. By inspecting Table 5.5, it is reasonable to assume that the first scan for case 3 (mark −) would have similar disease to either of the last two scans for case 1 (⊕), and indeed (g) may be assessed with (b) and (c).

the validation tests, it is now assumed that the metric is, on average, representative of acute disease. This can only be said when certain conditions of the data are met. One such condition is that adequate intensity separation exists in the STIR slice between fat and muscle tissues. This is normally expressed as adequate "fat suppression" in the literature. Slices that are blurred or have poor SNR may be acceptable so long as the prior condition is met. Importantly, the metric is only valid once major blood vessels have been excluded from analysis.

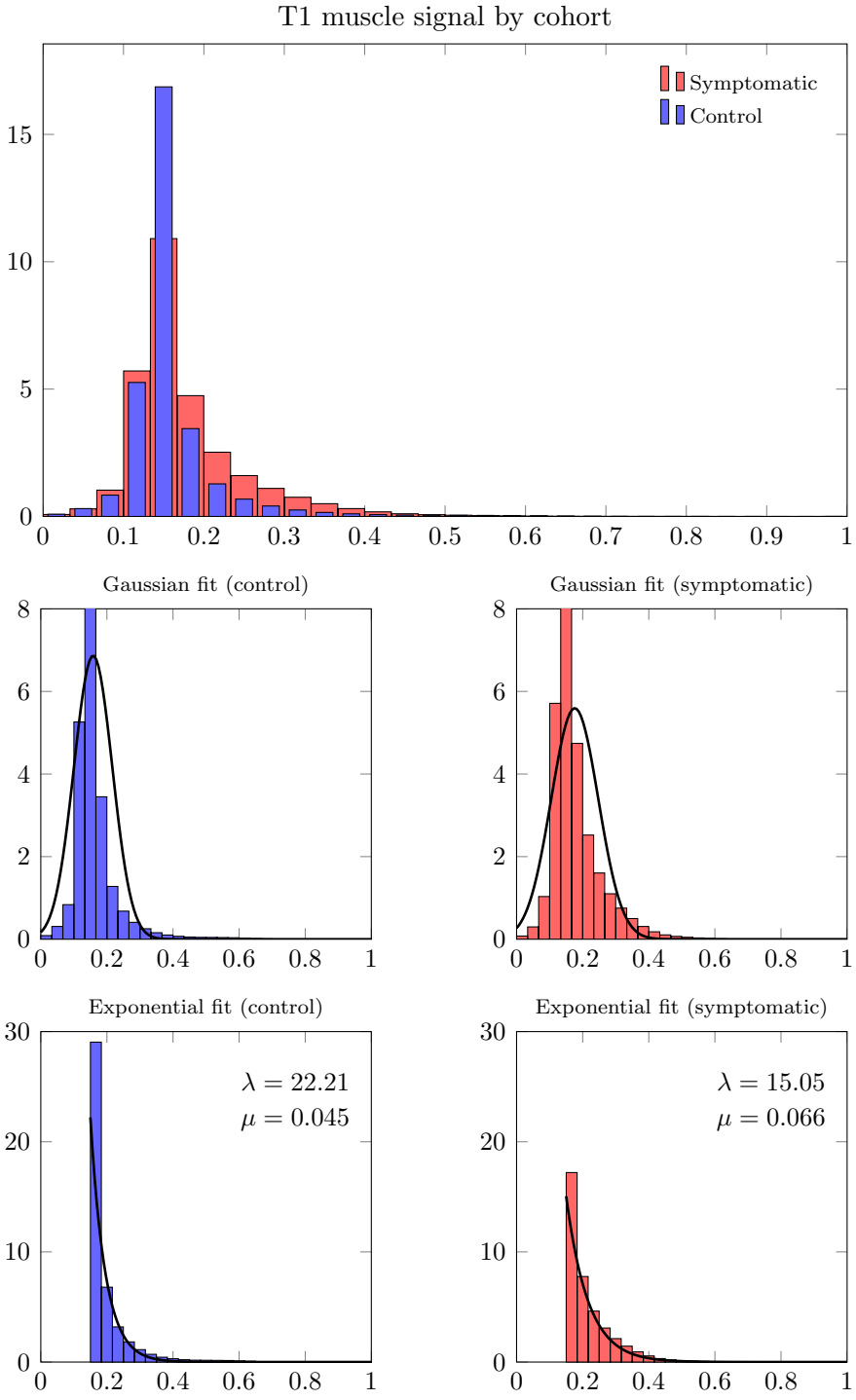
Some consideration must be made to the small number of test cases available. Given a larger database, it would be possible to calculate confidence intervals for the metric. In the absence of enough suitable data, the test approach has been conducted as rigorously as is possible.

It must be expressed that all acquisitions were performed in a clinical setting and therefore the quality of scans is representative of what could normally be expected in most hospitals. Therefore, this is an encouraging sign suggesting it could be implemented and used in a practical setting.

5.7 Analysis of T1 intensity distribution for fat infiltration by cohort

Fat infiltration is where muscle cells have been replaced by fat as a result of cell damage. The damage is caused by the build up of inflammation, as a result of the disease. Therefore, the acute disease (oedema) may reduce in severity, but the chronic disease (fat infiltration) is not expected to reduce. It is assumed that fat cells replacing muscle cells will lead to a slightly elevated T1 muscle signal. The explanation for this is that, for a given voxel in the acquisition matrix, the number of fat cells (out of the total number of cells) has increased and that their summed contribution elevates the signal. The intensity is raised as the fat appears brighter than the muscle on T1 scans. Investigations of the joint intensity space appear to confirm this. Indeed, fat infiltration lying within the muscle does not have the same brightness as fat around the leg. Visual inspection of the joint intensity space for subjects with confirmed infiltration does suggest a broadening of the T1 signal around the muscle cluster (see Figure 5.2 on page 59). This forms the theoretical basis for the selection procedure.

Figure 5.11 shows the distribution for T1 muscle pixels. The Gaussian distribution fits the data moderately better than anticipated and therefore the Jarque-Bera test was employed for verification. The test rejected the null hypothesis for both control ($p < 0.001$) and symptomatic ($p < 0.001$) cohorts. Table 5.6 shows the goodness of fit tests for the Gaussian and exponential distributions.



a
b c
d e

Figure 5.11: T1 muscle tissue histograms in the average case grouped by cohort. All histogram data has been normalised. (a) is a comparison for all T1 muscle signal. (b) and (c) show a Gaussian fit of the data for control and symptomatic cohorts, respectively. (d) and (e) show the right-hand tail of the distribution ($t'' > 0.15$ in this example) extracted and modelled with an exponential distribution.

R^2	Control	Symptomatic
Gaussian (whole)	0.580	0.698
Exponential (tail)	0.775	0.908

Table 5.6: Comparison of goodness of fit for Guassian and exponential distributions for T1 muscle intensity distributions.

Modelling the upper tail of the distribution with the exponential model is satisfactory.

To select a suitable region of suspected fat infiltration, Equation 5.12, Equation 5.13, and Equation 5.14 are modified for the T1 signal:

$$\mathbf{g}_1 = \left\{(x, c(x)) : c(x) = 1 \text{ if } x \in M \wedge t''(x) > \gamma_T, \text{ otherwise } c(x) = 0\right\} \tag{5.18}$$

where γ_T is the intensity threshold, set to 0.15, and $\mathbf{t}''$ is the standardised T1 image. The rate parameter may be calculated with

$$\lambda = \frac{\sum_x \mathbf{g}_1(x)}{\sum_x \left[\left(\mathbf{t}''(x) - \gamma_T\right) \mathbf{g}_1(x)\right]} \tag{5.19}$$

and the region of suspected fat infiltration is

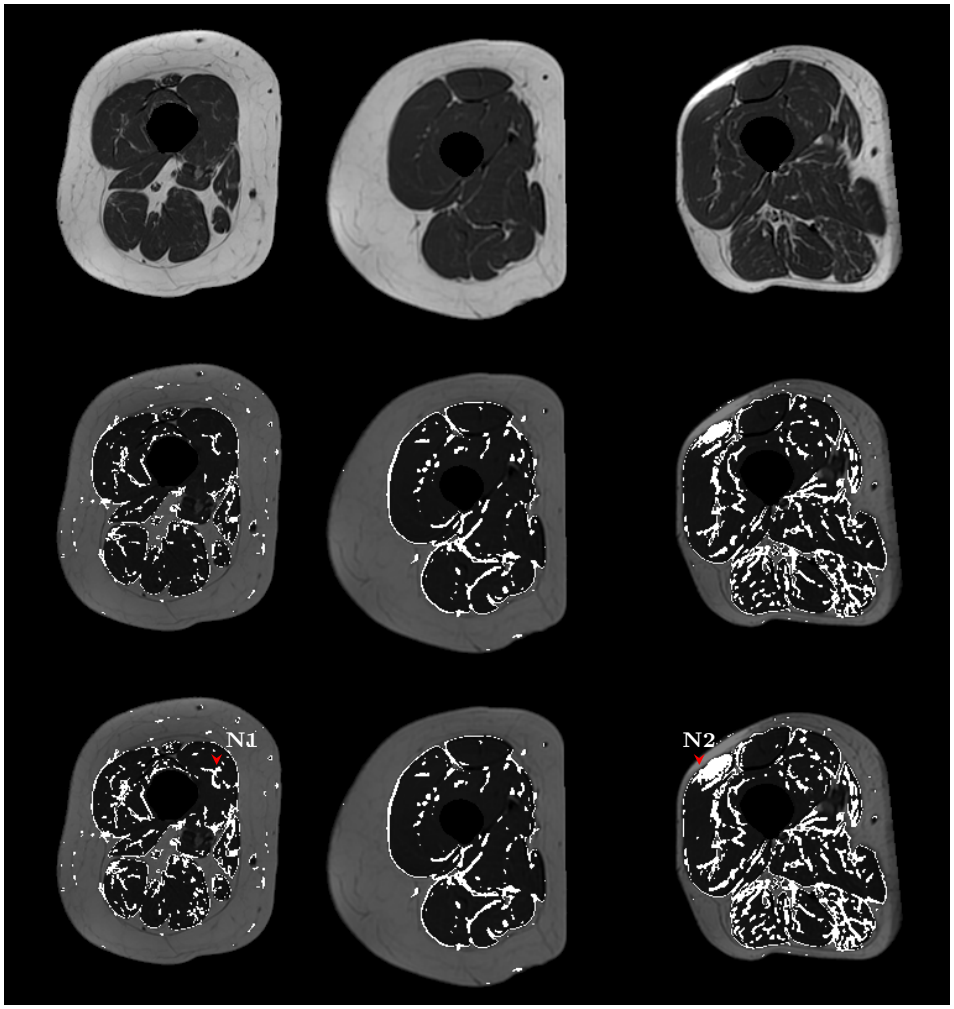
$$\begin{aligned} \mathbf{z}_1 &= \left\{(x, c(x))\right\}, \\ c(x) &= \begin{cases} 1 & \text{if } g_1(x)[\lambda\left(\mathbf{t}''(x) - \gamma_T\right) - 1] > g_1(x)\gamma_Z \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.20}$$

where γ_Z is a threshold value for the z-score, set to -0.31.

The above method is applied in a selection of cases in Figure 5.12. The difference between using an exponential distribution and a fixed threshold, is smaller than for the oedema metric (c.f. Figure 5.5). The intensity of T1 slices are generally more reliable than STIR. However, a fixed threshold method did not yield satisfactory results for the entire training database.

5.8 Definition of the infiltration metric

As with the oedema metric, it is clear that different subjects have a range of physical leg size and therefore the infiltration metric must account for the cross-sectional area of muscle present in the slice.



a	b	c
d	e	f
g	h	i

Figure 5.12: A T1 slice from 3 separate cases in the training database are shown in the first row. (a) is a control subject. (b) and (c) have mild and moderate fat infiltration, respectively. The middle row shows the application of the method from section 5.7 used on the slice above. The white region indicates $\mathbf{z}_1$, with the faded slice behind. The last row illustrates a fixed threshold method, which may be compared to the prior result. **Notations**: Arrow N1 shows slightly too much selection for the control case. Arrow N2 shows the effect of the bias field on the selection method.

Therefore, the infiltration metric is formulated

$$\Psi_{\text{Infil}} = \frac{\sum \mathbf{z}_1}{\sum M} \tag{5.21}$$

where "Infil" is short for infiltration, M is the set of muscle pixels, as before, and $\mathbf{z}_1$ is found with Equation 5.20.

5.9 Correlation of the infiltration metric with chronic disease

Linear regression of the infiltration metric against expert grading is shown in Figure 5.13. The correlation coefficient was 0.821, linear regression slope $m = 0.056$ with intercept $c = 0.147$, and the fit had an $R^2 = 0.6744$.

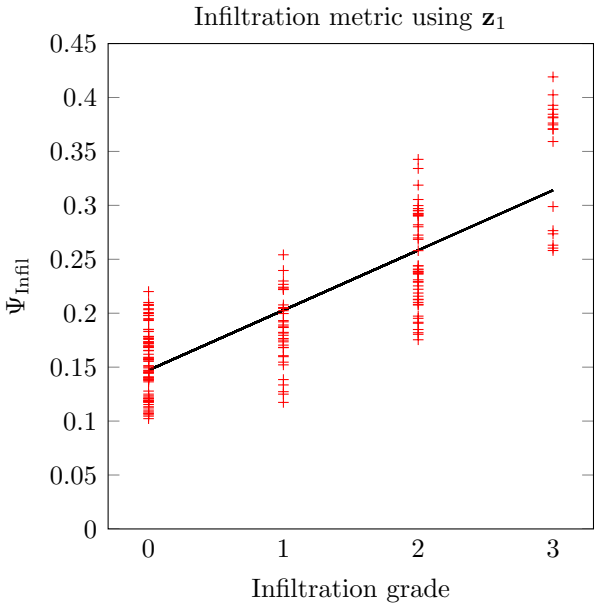


Figure 5.13: Linear regression between the Ψ_{Infil} metric and expert grade using region $\mathbf{z}_1$.

Differentiating between normal and mild levels of fat infiltration is difficult, as shown by the small increase in average value between these grades in the figure. It is expected (e.g. in control subjects) that a thin layer of fat separates some of the muscles, and therefore detecting the difference slightly above the "normal" baseline is subject to error.

Very severe fat infiltration may not be estimated correctly by the method detailed. A voxel that is entirely occupied by fat cells will have hyperintense signal, and therefore not expected to be a member of M. In these cases, the original muscle shape is hidden, and therefore it is extremely difficult to estimate the quantity of muscle that "should" be present.

5.10 Validation for the infiltration metric: rescanned control

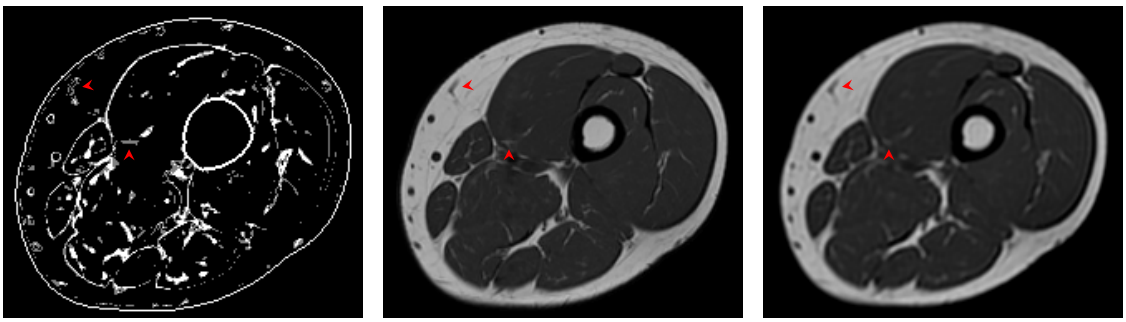
As with the oedema metric, further testing will help indicate the level of robustness for the infiltration metric. The conditions for the test were replicated as in

subsection 5.6.1. Table 5.7 shows the data from the validation test and Figure 5.14 shows the worst match in the case.

I_1[a]	I_2[a]	T_1[b]	T_2[b]	Ψ_1[c]	Ψ_2[c]	$\lvert\Psi_1 - \Psi_2\rvert$	Jaccard	Dice
3867	4839	41197	41470	0.0938	0.1166	0.0228	0.4609	0.6310
4802	4799	40082	39886	0.1198	0.1203	0.0005	0.4160	0.5876
4211	5574	38481	38395	0.1094	0.1451	0.0357[d]	0.4683	0.6379
4476	4288	36258	36136	0.1234	0.1186	0.0047	0.4042	0.5757
4637	4870	35247	34985	0.1315	0.1392	0.0076	0.4829	0.6513
3410	3542	33661	33421	0.1013	0.1059	0.0046	0.4384	0.6096
				0.1132[e]	0.1243[e]			

[a] Quantity of infiltration, $I_i = \sum \mathbf{z}_1$, subscripts denote scan 1 and 2.
[b] Quantity of muscle tissue, $T_i = \sum M$, for scan 1 and 2.
[c] $\Psi_i = \Psi_{\text{Infil}}$, for scan 1 and 2.
[d] Denoted as the worst match in the set. Shown in Figure 5.14.
[e] Mean of the values above.

Table 5.7: Data table for rescanned control (infiltration test). The six rows in the table correspond to the six slice locations (identical for both scans).



a b c

Figure 5.14: Diagram of worst match in rescanned control case for the infiltration metric. (a) is an overlap digram showing the two $\mathbf{z}_1$ regions. The solid white contours show the leg mask and femur. Intersection of the two regions is shown in white. (b) is the first scan. (c) is the second scan. The view window has been cropped to show detail. Note that (c) is blurred compared to (b).

5.11 Conclusion

Two disease-specific metrics have been developed to measure the key clinical presentations of inflammatory muscle disease: oedema and fat infiltration. The metrics were designed from the expected distribution of diseased muscle tissue in the joint histogram. Based on the validation tests, they have been shown to have

a good level of repeatability. Moderately strong correlation with expert grade suggests that the computed quantities are in line with expert medical opinion. Class overlap may be explained by the effect of rater variance and the fact that only a single expert grade is available per case. Given that all scans were acquired clinically, the methods presented are suitable for application in a clinical environment.

In the following chapter, the disease metrics will be used in conjunction with statistical measures for the task of automatic classification.

Chapter 6

Classification of disease severity

The disease metrics from Chapter 5 are based on the hyperintense signal of muscle tissue for T1 and STIR modalities. In this chapter, properties of the joint distribution are utilised, together with the oedema and infiltration metrics, for automatic classification on a per slice basis.

6.1 Premise for the approach

As known, the joint histogram represents the distribution of tissue intensities. The shape of the distribution appears to represent the disease state, as shown in Figure 5.2 (page 59). It is assumed that the change in distribution is caused by physical changes occurring in the tissues. This is supported by the clinical reason for imaging the patient in the first instance: to show the severity of oedema and fat infiltration. In the previous chapter, two disease metrics, Ψ_{Oedema} and Ψ_{Infil}, were designed to model the distribution for these disease characteristics. These metrics are based on the observed differences in distribution above a threshold value for muscle tissue. However, the whole joint intensity space contains a great deal of data that is not used in either of the disease metrics. While these metrics may be the most important from a clinical point of view, it is assumed that other regions in the joint intensity space may carry "information" on the disease that could also be harnessed.

Knowledge of the most important regions in the joint intensity space would be of benefit to help design a CAD system, where the output classification is expected to match the expert grade. The two grades (oedema and fat infiltration) are treated as separate classification problems, since it is known that theses disease characteristics can be independent. It is expected that a system may be devised with the training database and tested on the out of sample database for realistic clinical results.

6.2 Selected statistical measures

Measures were selected in order to attempt to describe the broad shape and characteristics of the underlying data. The measures are: standard deviation, skewness, kurtosis, and gini coefficient.

Given a sequence $x = x_1, x_2, \ldots, x_n$ with the number of elements n, then the standard deviation is defined as

$$\sigma = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \right)^{\frac{1}{2}}. \tag{6.1}$$

This measure is used on the presumption that deviations in the distribution may have some correlation with disease severity.

Skewness is defined as

$$s = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^3}{\sigma^3}, \tag{6.2}$$

which is negative for long left-tail distributions, 0 for symmetric, and positive for long right-tail distributions. The same property may be measured more robustly with another statistic called the medcouple, which may be computed in $O(n \log n)$ time [9].

The kurtosis is defined as

$$k = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^4}{\sigma^4}. \tag{6.3}$$

This measures how "pinched" the distribution is with respect to the standard normal distribution.

The Gini coefficient is a statistical measure of concentration [54; 11] and it is defined as

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1) x_{(i)}}{n^2 \mu}, \tag{6.4}$$

where $x_{(i)}$ denotes the ith element in the ordered sequence of x (ascending).

The measures described above may be calculated on any sequence x. However, since the joint intensity space is two dimensional, data from a given region must be represented by a sequence, in some manner. The approach taken here is to use a column or row sum over the region to yield a sequence on which to calculate the measures.

A formal definition of the joint histogram was presented earlier (Equation 4.4, page 39). Given that the joint histogram has already been computed, it is convenient to consider it as a matrix $H_{i,j}$ where i indexes rows and j indexes columns.

$$H_{i,j} = \begin{pmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,J} \\ H_{2,1} & H_{2,2} & \cdots & H_{1,J} \\ \vdots & \vdots & \ddots & \vdots \\ H_{I,1} & H_{I,2} & \cdots & H_{I,J} \end{pmatrix} \tag{6.5}$$

where $I = 128$, representing maximum STIR intensity, and $J = 128$, representing maximum T1 intensity. A rectangular ROI on $H_{i,j}$ will be denoted $\mathbf{R}(I_1, I_2, J_1, J_2)$ which specifies the region where $I_1 \leq i \leq I_2$ and $J_1 \leq j \leq J_2$. The column sum for a region may then be found with

$$x_{k+1} = \sum_{i=I_1}^{I_2} H_{i,k+J_1} \text{ for } k = 0, \ldots, J_2 - J_1 \tag{6.6}$$

where x_1 would be the first column from the region. The meaning for a single column is: the sum of frequencies for any joint intensity where the (discretised) T1 signal was fixed but the (discretised) STIR signal was between certain limits. The sequence x_i represents a contiguous range of column sums. Similar to Equation 6.6, the row sum may be found with

$$x_{k+1} = \sum_{j=J_1}^{J_2} H_{k+I_1,j} \text{ for } k = 0, \ldots, I_2 - I_1, \tag{6.7}$$

where x_1 would be the first row from the region. The meaning for a single row is: the sum of frequencies for any joint intensity where the (discretised) STIR signal was fixed but the (discretised) T1 signal was between certain limits. The sequence x_i represents a contiguous range of row sums.

Use of carefully selected regions, $\mathbf{R}(I_1, I_2, J_1, J_2)$, allow two-dimensional data to be aggregated into a sequence, where the univariate statistical measures (presented above) may be applied. Which regions to use are then of primary interest.

6.3 Method for scoring the intensity space

Initial work using the selected features on three broad regions of the joint histogram space (rectangular regions covering the muscle, the fat, and the whole intensity space) indicated that using the raw numbers was not powerful. As discussed in the previous chapter, it was recognised that the difference in distribution is important, particularly above the muscle centre value for each dimension. Therefore, it was assumed that a scoring system that encompassed some reference to the signal intensity would be useful. In this case, a straightforward scoring system is

applied to the joint histogram. A T1-scored feature is where the joint histogram matrix $H_{i,j}$ is multiplied by the square of the T1-bin index such that

$$\hat{H}_{i,j} = H_{i,j} \circ \begin{pmatrix} 1^2 & 2^2 & \cdots & 128^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1^2 & 2^2 & \cdots & 128^2 \end{pmatrix} \tag{6.8}$$

where $\circ$ is element-wise (Hadamard) multiplication. The selected measures are calculated on a region of $\hat{H}_{i,j}$. Similarly, STIR-scored features use the square of the STIR-bin index such that

$$\hat{H}_{i,j} = H_{i,j} \circ \begin{pmatrix} 1^2 & \cdots & 1^2 \\ 2^2 & \cdots & 2^2 \\ \vdots & \ddots & \vdots \\ 128^2 & \cdots & 128^2 \end{pmatrix}. \tag{6.9}$$

In both cases, the histogram is normalised such that volume under the surface (defined by region $\mathbf{R}$) is unity. Further testing indicated this scoring system to be more powerful than using raw frequencies.

6.4 Feature vector template

Table 6.1 shows the initial feature vector. Note that the features are only calculated over the region currently being considered, as shown by the re-indexing in Equation 6.6 and Equation 6.7.

Scoring:	T1-scored				STIR-scored				T1-scored				STIR-scored			
Sum type:	Row-sum								Column-sum							
Statistic:	k	s	G	σ	k	s	G	σ	k	s	G	σ	k	s	G	σ
Element:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Table 6.1: Layout of the feature vector.

6.5 Evaluation of region performance

To establish which regions in the joint histogram were the most suitable, a test was conducted. The aim of the test was to find a general trend in the regions being considered. Table 6.2 defines the regions that were used in the test.

Each region (1 through 16) was considered in turn. The feature vector (as detailed in section 6.4) was constructed by considering only the data from the

No.	Region	Area	Primary tissue
1	$\mathbf{R}(52, 77, 20, 45)$	676	Muscle
2	$\mathbf{R}(52, 94, 20, 62)$	1849	Muscle
3	$\mathbf{R}(52, 111, 20, 79)$	3600	Muscle
4	$\mathbf{R}(52, 128, 20, 96)$	5929	Muscle
5	$\mathbf{R}(26, 77, 13, 26)$	728	Muscle
6	$\mathbf{R}(18, 86, 9, 30)$	1518	Muscle
7	$\mathbf{R}(9, 94, 5, 35)$	2666	Muscle
8	$\mathbf{R}(1, 103, 1, 39)$	4017	Muscle
9	$\mathbf{R}(26, 52, 58, 84)$	729	Fat
10	$\mathbf{R}(26, 69, 41, 84)$	1936	Fat
11	$\mathbf{R}(26, 86, 24, 84)$	3721	Fat
12	$\mathbf{R}(26, 103, 7, 84)$	6084	Fat
13	$\mathbf{R}(20, 32, 77, 90)$	182	Fat
14	$\mathbf{R}(13, 39, 71, 96)$	702	Fat
15	$\mathbf{R}(7, 45, 64, 103)$	1560	Fat
16	$\mathbf{R}(1, 52, 58, 109)$	2704	Fat

Table 6.2: Definition of the regions used in the region importance test. Regions 1-4 increase in size but the top-left corner is located at the muscle cluster centre. Regions 5-8 expand with the muscle centre located in the middle. Regions 9-12 increase in size but the top-right corner is located at the fat cluster centre. Regions 13-16 expand with the fat centre located in the middle. Area is in units squared.

defined region. Using the training database, 186 instances of the feature vector were computed for the region. Since two expert grades are available per case (one for oedema, one for fat infiltration), two classification problems are to be solved and therefore two output datasets are generated per region. In total, there are 32 datasets (16 regions, 2 classifications).

The classifier for this test was chosen to be Naive Bayes. This classifier has been used extensively in the literature and it has been shown to work well in a variety of domains [10]. It is popular due to its robustness [73] and it generally gives better predictive accuracy, as compared to more sophisticated methods [7]. The dataset is loaded and a subset selection algorithm [24] is applied. 10-fold cross validation is used with Naive Bayes. The weighted average of the AUROC is used to judge the overall performance of the region, along with TPR, FPR, and the F-measure. The F-measure is defined as the harmonic mean of TPR and precision.

The above test was repeated 10 times and the average performance taken. In order to make the test more robust, each run was randomised (e.g. fold location).

Region performance per test type is summarised in Table 6.3. The category of the features selected by the evaluator is shown in Table 6.4.

Region	Test type	TPR	FPR	F-measure	Weighted ROC area
1	infiltration	0.7891	0.0704	0.7890	0.9502
2	infiltration	0.7716	0.0762	0.7705	0.9406
3	infiltration	0.7436	0.0855	0.7422	0.9362
4	infiltration	0.7321	0.0893	0.7303	0.9328
5	infiltration	0.7146	0.0951	0.7085	0.9356
6	infiltration	0.7303	0.0900	0.7298	0.9334
7	infiltration	0.7382	0.0872	0.7351	0.9327
8	infiltration	0.7282	0.0906	0.7240	0.9274
9	infiltration	0.5191	0.1603	0.5057	0.7858
10	infiltration	0.6558	0.1146	0.6478	0.8926
11	infiltration	0.7859	0.0714	0.7846	0.9403
12	infiltration	0.8240	0.0588	0.8219	0.9483
13	infiltration	0.4501	0.1833	0.4150	0.7396
14	infiltration	0.5174	0.1607	0.4949	0.7576
15	infiltration	0.4983	0.1673	0.4742	0.7881
16	infiltration	0.5075	0.1642	0.4867	0.8025
1	oedema	0.7250	0.0915	0.7274	0.9070
2	oedema	0.7264	0.0912	0.7271	0.9234
3	oedema	0.7461	0.0846	0.7451	0.9288
4	oedema	0.7321	0.0892	0.7302	0.9076
5	oedema	0.5679	0.1439	0.5736	0.8272
6	oedema	0.5863	0.1379	0.5930	0.8616
7	oedema	0.6119	0.1294	0.6167	0.8568
8	oedema	0.6689	0.1103	0.6708	0.9004
9	oedema	0.5919	0.1360	0.5816	0.8298
10	oedema	0.6246	0.1251	0.6202	0.8589
11	oedema	0.6533	0.1156	0.6449	0.8777
12	oedema	0.6946	0.1019	0.6948	0.8955
13	oedema	0.4532	0.1822	0.4296	0.7464
14	oedema	0.5712	0.1429	0.5660	0.8153
15	oedema	0.5761	0.1414	0.5738	0.8195
16	oedema	0.5924	0.1358	0.5881	0.8371

Table 6.3: Region performance data: all figures are an average of 10 runs.

Regions	Test type	T1-scored	STIR-scored	Row-sum	Column-sum
1-8	infiltration	330	210	170	370
9-16	infiltration	260	370	270	360
1-8	oedema	380	380	470	290
9-16	oedema	330	440	490	280

Table 6.4: Tally of selected features by region and category. Note that the evaluator may select any number of features.

Careful analysis of these tables reveals some important details that should be discussed before moving on.

In general, Table 6.4 shows that, on average, the evaluator preferred row-summed features for classifying the oedema grade, irrespective of the exact region used. The reverse is true for classifying infiltration grade. STIR-scored features were generally preferred for regions 9-16 (over the fat cluster). Interestingly, T1-scored features over the muscle cluster (regions 1-8) were preferred for classifying infiltration, which supports the method used in section 5.7 to help calculate Ψ_{Infil}. Such preference for classifying oedema is not so clear for STIR-scored features over the muscle (regions 1-8), however the overall preference for row-summed features does give strong support for the method developed in section 5.2 to compute Ψ_{Oedema}.

Table 6.3 indicates that oedema classification is generally a more difficult task than infiltration, as shown by the generally higher FPR. Regions 1-8 for oedema suggest that larger regions for oedema classification are useful, given that the average ROC area correlates with the region size. The lower value for region 4 could be caused by the relative paucity of data-points in the upper portion of the distribution for some image pairs. Region 3 has the best performance. For the classification of infiltration, the results suggest more than one region is applicable. Regions 1-8 perform well, but the weighted ROC area is negatively correlated with region size, suggesting the most useful data is near the cluster centre.

6.6 Correlation as a disease feature

With careful consideration, it is possible to hypothesise a change in correlation between T1 and STIR signal for control subjects versus symptomatic subjects. In general, correlation estimates the degree of dependence of one variable on another variable. When the relationship is dependent, the correlation may be strong, but when the relationship is independent, the correlation is expected to be close to zero. Since corresponding elements in the T1 and STIR signals are expected to be a sample of the same physical space, and therefore the same material, it is expected that some baseline level of correlation exists between the two signals. However, due to the difference in acquisition sequences, the relationship is not expected to be linear. Indeed, to consider a range of tissue types (ordered by their expected signal intensity) for the two modalities, it is clear that their order is not simply reversed. If the samples are limited to a single tissue class, however, then the relationship between the two modalities should be more clear and a baseline level of correlation may be obtained for the control case. In the presence of the disease, a divergence from the baseline level of correlation is expected. This can

easily be seen due to the presence of hyperintense STIR signal for inflamed muscle tissue, with limited to no change in the corresponding T1 signal. As a possibly confounding issue, it can also be seen that a slightly elevated T1 signal, in the case of mild to moderate fat infiltration, would also affect the correlation statistic. However, conceding that possibility, it is still expected to represent a divergence from the control case, but may be non-specific to the disease state.

Given the above hypothesis, and a fundamental reason for its function, the Spearman rank correlation coefficient is included in the feature vector. It is calculated on corresponding samples for muscle tissue. The Spearman correlation coefficient is calculated with two equal length sequences such that

$$\rho = 1 - \frac{6 \sum[(x_{(i)} - y_{(i)})^2]}{n(n^2 - 1)} \tag{6.10}$$

where $x_{(i)}$ is the rank for the ith element in a sequence x, $y_{(i)}$ is the rank for the ith element in another sequence y, and n is the cardinality of x. In this case, the disease specific feature was chosen to operate on a subset of muscle pixels,

$$q = \{m : m \in M \wedge s''(m) > 0.4 \wedge t''(m) > 0.15\}, \tag{6.11}$$

where M is the set of muscle pixel locations, as before. ρ is then calculated on the two sequences for locations in q.

Some caution may be warranted when using statistical measures and, in the end case, it has been advised that a visible relationship is established [2].

6.7 Configuration for classification

Table 6.5 and Table 6.6 show the layout of the oedema and infiltration feature vectors, respectively.

Data:	Region 3	Muscle pixels M	Muscle pixels M
Feature:	{Table 6.1}	ρ	Ψ_{Oedema}
Element:	1-16	17	18

Table 6.5: The oedema feature vector.

Data:	Region 4	Region 8	Muscle pixels M	Muscle pixels M
Feature:	{Table 6.1}	{Table 6.1}	ρ	Ψ_{Infil}
Element:	1-16	17-32	33	34

Table 6.6: The infiltration feature vector.

Feature no.	Feature	Region	Sum type	Scoring
1	k, kurtosis	3	row	T1
10	s, skewness	3	column	T1
12	σ, std.dev.	3	column	T1
13	k, kurtosis	3	column	STIR
16	σ, std.dev.	3	column	STIR
17	ρ, correlation	M	-	-
18	Ψ_{Oedema}	M	-	-

Table 6.7: Features selected from the oedema vector after dimensionality reduction.

Feature no.	Feature	Region	Sum type	Scoring
13	k, kurtosis	4	column	STIR
25	k, kurtosis	8	column	T1
29	k, kurtosis	8	column	STIR
34	Ψ_{Infil}	M	-	-

Table 6.8: Features selected from the infiltration vector after dimensionality reduction.

Region 3 was chosen for the oedema vector as this region showed the best performance in section 6.5. In an initial test, region 4 (alone) was selected for the infiltration vector, however poor classification performance resulted. In visualisation of the feature space, it became apparent that the majority of features from this region were not distinct. Therefore, region 8 was appended, which also had good performance in section 6.5.

Note that both vectors have the rank correlation of muscle signal, ρ (section 6.6). The oedema vector has disease metric Ψ_{Oedema} (section 5.4), and the infiltration vector has Ψ_{Infil} (section 5.8). It is assumed that the combination of standard statistical measures and disease specific metrics will result in a powerful classifier.

Feature vectors were constructed for the training database and dimensionality was reduced using the subset selection evaluator [24], as before. Features were scored out of 10 folds (k-fold validation). Features scoring less than 100% were dropped from further analysis. Table 6.7 and Table 6.8 show the selected features for oedema and infiltration, respectively. Note that, with "competition" from other features, the Ψ_{Oedema} and Ψ_{Infil} metrics developed in the previous chapter were both chosen. Correlation (ρ) was chosen for oedema but not for infiltration. A total of 11 features were chosen out of a possible 52.

Figure 6.1 shows the feature space for feature 1 (based on kurtosis) against feature 18 (Ψ_{Oedema}). Since it is clear that feature 1 generally decreases with increasing Ψ_{Oedema} then negative correlation is present. Note that, in general,

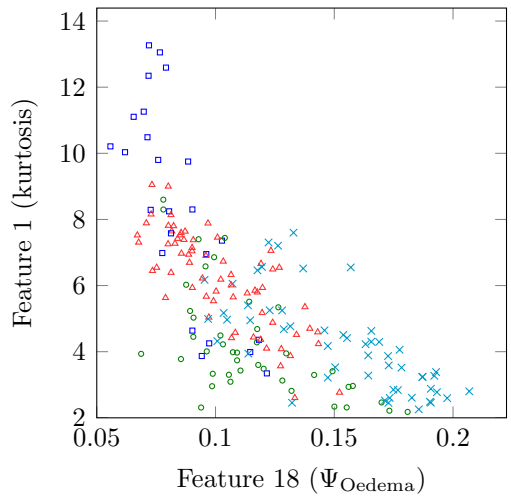


Figure 6.1: Oedema feature space for features 1 and 18.

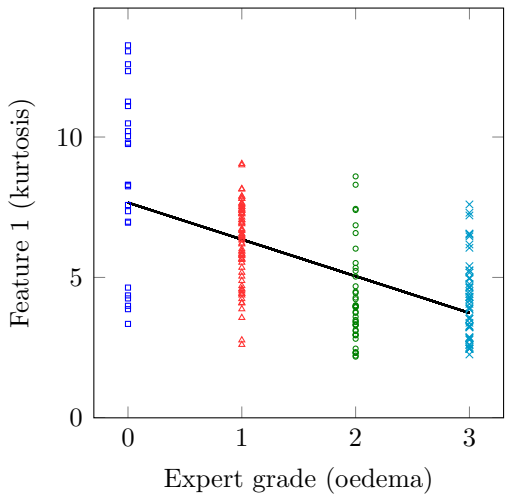


Figure 6.2: Oedema feature 1 against expert grade with linear regression line.

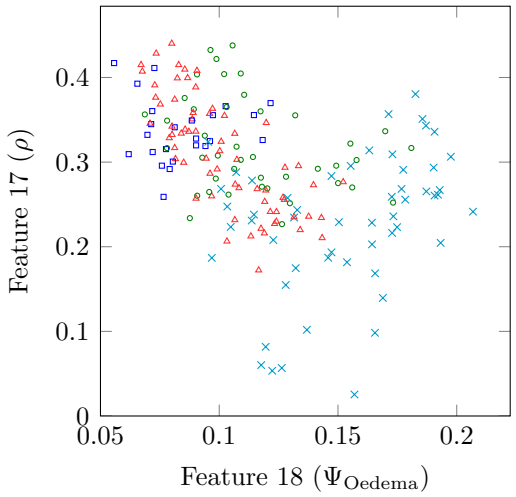


Figure 6.3: Oedema feature space for features 17 and 18.

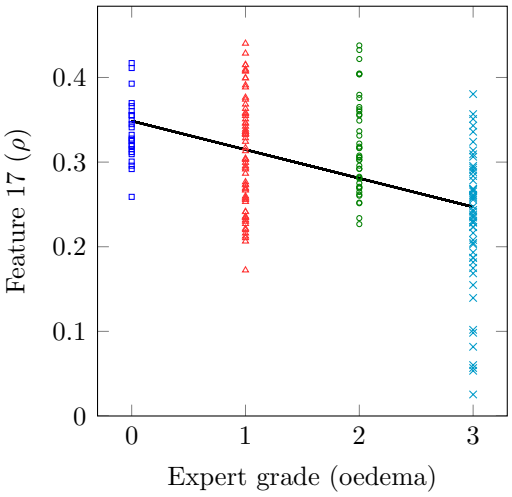


Figure 6.4: Oedema feature 17 against expert grade with linear regression line.

the "monotonicity" with respect to disease grade is maintained (i.e. disease grade order is maintained). Figure 6.2 shows feature 1 against the expert grade with a linear regression fit for comparison. Note that it appears to help separate the Normal and Mild grades.

Figure 6.3 shows the feature space for feature 17 (rank correlation) against feature 18 (Ψ_{Oedema}), and Figure 6.4 shows feature 17 against expert grade. As discussed in section 6.6, it appears that the rank correlation does reduce in the presence of oedema, however there is some overlap present between classes.

Since the most "visibly strong" features have been discussed, the other selected features from Table 6.7 are not detailed further. Many of the features have a moderate degree of overlap between grades, but some trend can be seen.

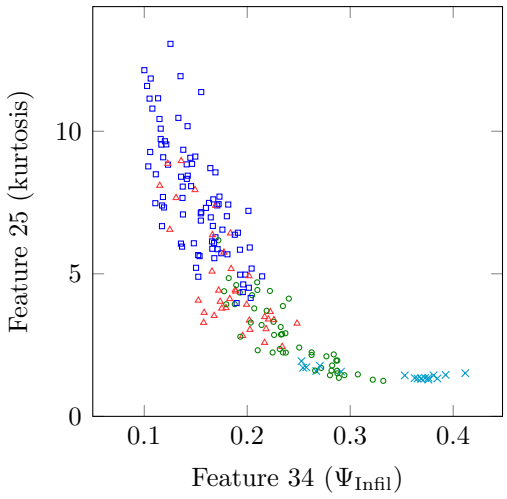


Figure 6.5: Infiltration feature space for features 25 and 34.

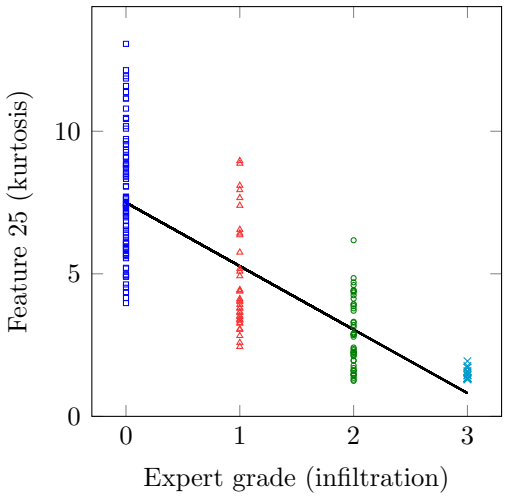


Figure 6.6: Infiltration feature 25 against expert grade with linear regression line.

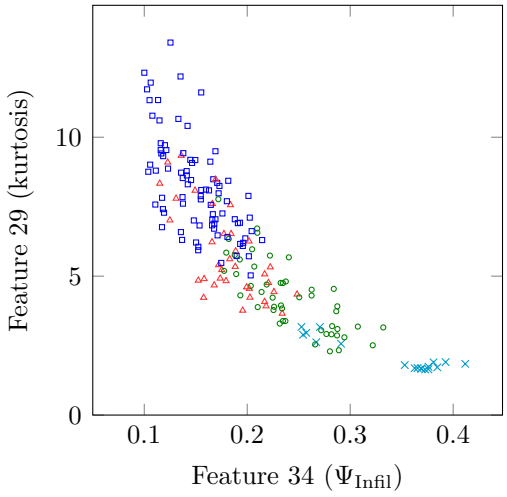


Figure 6.7: Infiltration feature space for features 29 and 34.

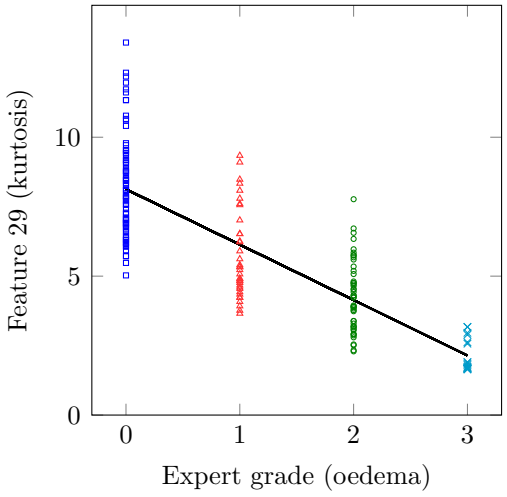


Figure 6.8: Infiltration feature 29 against expert grade with linear regression line.

Figure 6.5 shows the relationship between feature 25 from the infiltration vector and feature 34 (Ψ_{Infil}). Despite overlap between the grades, it can be seen that the feature tends to decrease while the percentage of fat infiltration increases. Figure 6.6 shows feature 25 against the infiltration grade, with a linear regression line fitted. Figure 6.7 shows feature 29 against feature 34 (Ψ_{Infil}), again exhibiting the same trend, although the spread is mildly different when compared to Figure 6.5. Figure 6.8 shows feature 29 against infiltration grade, with linear regression line, for reference.

6.8 Classification results for the training database

Instances were 10-fold cross validated on the training database, comprised of 186 instances. For comparison, two different classifiers have been used: Naive Bayes [10] and AdaBoost [22].

The Naive Bayes classifier was chosen since the selected features from the evaluator (for both oedema and infiltration) generally had the appearance of slowly-varying probabilities across the feature space (for different classes). As discussed in section 6.5, Naive Bayes is known to perform well in a variety of domains. Despite the naive assumption of conditional independence between features, it is this property that reduces the variance of its predictions [7, p. 1782].

The AdaBoost classifier was chosen as an alternative to Naive Bayes. AdaBoost is a popular ensemble method which has been empirically shown not to over-fit the training data [73], which was considered important for robustness, especially with regards to out of sample data. In this case, the tree classifier J48 (also known as C4.5 [60]) was used as the weak classifier for AdaBoost. J48 was chosen because it was clear that in some feature spaces, some division would allow adequate separation of one or more of the classes. It was assumed that AdaBoost would outperform Naive Bayes.

For classificaiton of oedema, Table 6.9 shows the confusion matrix for each classifier, and Table 6.10 shows a comparison of their performance. Both classifiers perform well, however AdaBoost achieves equal or better performance in all classes. Careful inspection of Table 6.9 reveals that each classifier predicts the class correctly on average, but there is moderate "jitter" or overlap. In the event that the classifier predicts the wrong class, and considering that the output class is nominal, then it would be particularly desirable to measure the rate with which a given classifier predicts "within one" of the real class. For instance, it is preferable that a subject with Moderate disease be classified as Severe, or possibly Mild, but not Normal. By this measure, the Naive Bayes classifier scored 177 (95%), and AdaBoost scored 171 (92%).

For classification of infiltration, Table 6.11 shows the confusion matrix for each classifier, as before, and Table 6.12 shows a comparison of their performance. Clearly, both classifiers have difficulty with the Mild infiltration class. AdaBoost achieves a greater TPR but slightly worse FPR as compared to Naive Bayes. The "within one" score is 185 (99%) for Naive Bayes and 178 (96%) for AdaBoost.

Note that the "within one" score is generally high compared to the TPR, as found in Table 6.10 and Table 6.12. To explain this difference note that the TPR is calculated based on 1 possible correct outcome and 3 possible incorrect

Real	Naive Bayes				AdaBoost			
	0	1	2	3	0	1	2	3
0 (Normal)	17	6	1	0	22	2	0	0
1 (Mild)	5	42	12	7	2	54	3	7
2 (Moderate)	0	9	31	2	0	5	35	2
3 (Severe)	0	1	7	46	0	8	3	43

Table 6.9: Confusion matrices for oedema classification using the training database.

outcomes, and therefore is lower than expected, whereas the "within one" score incorporates the importance of "close" classifications on the ordinal scale. The low FPR indicates that, in general, the other instances are not incorrectly labelled as belonging to the class.

6.9 Classification results for the out of sample database

The classification models constructed in the previous section were re-evaluated on the out of sample database, using 10-fold cross validation.

Table 6.13 shows the confusion matrix for both classifiers for oedema. Table 6.14 shows the associated performance. The distinction between Normal and Mild subjects is particularly subtle, and both classifiers labelled Normal subjects as Mild. The "within one" score was 183 (98%) for Naive Bayes and 181 (96%) for AdaBoost.

Similarly, Table 6.15 shows the confusion matrix for both classifiers for infiltration. Table 6.16 shows the associated performance. AdaBoost performed poorly on Mild cases, with a TPR of 0.133. Naive Bayes performed well, with higher TPR and comparable FPR, as compared to AdaBoost. The "within one" score was 185 (99%) for Naive Bayes and 181 (96%) for AdaBoost.

ROC curves are shown for the Naive Bayes classifier using the out of sample database in Figure 6.15 and Figure 6.16.

Overall, the results suggest that classification of oedema is more difficult than infiltration (with the currently defined feature vectors). The figures on page 97 compare some of the feature spaces between the training database and the out of sample database. Figure 6.10 and Figure 6.12 suggest that the general trend for features 1 and 17 from the oedema vector were maintained between databases. Figure 6.14 shows features 25 and 24 from the infiltration vector allow quite specific clusters to form in the feature space, maintained between databases.

Class	TPR	FPR	F-measure	ROC area
Naive Bayes				
0 (Normal)	0.708	0.031	0.739	0.968
1 (Mild)	0.636	0.133	0.677	0.877
2 (Moderate)	0.738	0.139	0.667	0.918
3 (Severe)	0.852	0.068	0.844	0.960
Weighted average	0.731	0.102	0.731	0.922
AdaBoost				
0 (Normal)	0.917	0.012	0.917	0.997
1 (Mild)	0.818	0.125	0.800	0.925
2 (Moderate)	0.833	0.042	0.843	0.947
3 (Severe)	0.796	0.068	0.811	0.976
Weighted average	0.828	0.075	0.828	0.954

Table 6.10: Performance of oedema classification using the training database.

Real	Naive Bayes				AdaBoost			
	0	1	2	3	0	1	2	3
0 (Normal)	65	19	0	0	65	15	4	0
1 (Mild)	10	14	12	0	11	18	7	0
2 (Moderate)	1	10	27	10	4	5	36	3
3 (Severe)	0	0	4	14	0	0	7	11

Table 6.11: Confusion matrices for infiltration classification using the training database.

Class	TPR	FPR	F-measure	ROC area
Naive Bayes				
0 (Normal)	0.774	0.108	0.813	0.930
1 (Mild)	0.389	0.193	0.354	0.714
2 (Moderate)	0.563	0.116	0.593	0.882
3 (Severe)	0.778	0.060	0.667	0.979
Weighted average	0.645	0.122	0.653	0.881
AdaBoost				
0 (Normal)	0.774	0.147	0.793	0.869
1 (Mild)	0.500	0.133	0.486	0.726
2 (Moderate)	0.750	0.130	0.706	0.880
3 (Severe)	0.611	0.018	0.688	0.946
Weighted average	0.699	0.128	0.701	0.851

Table 6.12: Performance of infiltration classification using the training database.

Real	Naive Bayes				AdaBoost			
	0	1	2	3	0	1	2	3
0 (Normal)	0	5	0	1	0	6	0	0
1 (Mild)	3	17	4	0	1	17	2	4
2 (Moderate)	0	7	11	0	0	6	12	0
3 (Severe)	0	2	9	13	0	1	10	13

Table 6.13: Confusion matrices for oedema classification using the out of sample database.

Class	TPR	FPR	F-measure	ROC area
Naive Bayes				
0 (Normal)	0.000	0.045	0.000	0.321
1 (Mild)	0.708	0.292	0.618	0.780
2 (Moderate)	0.611	0.241	0.524	0.797
3 (Severe)	0.542	0.021	0.684	0.871
Weighted average	0.569	0.168	0.565	0.776
AdaBoost				
0 (Normal)	0.000	0.015	0.000	0.467
1 (Mild)	0.708	0.271	0.630	0.781
2 (Moderate)	0.667	0.222	0.571	0.760
3 (Severe)	0.542	0.083	0.634	0.875
Weighted average	0.583	0.175	0.564	0.781

Table 6.14: Performance of oedema classification using the out of sample database.

In some cases it is difficult to ascertain if a relationship is maintained across databases, in part due to the limited quantity of data-points. Allowance must be made for the single-sample expert grading, which will have some unknown variance. However, it is felt that in general, the rough location of clusters in the feature space are maintained for each class.

Real	Naive Bayes				AdaBoost			
	0	1	2	3	0	1	2	3
0 (Normal)	10	2	0	0	11	1	0	0
1 (Mild)	9	16	5	0	21	4	5	0
2 (Moderate)	1	5	10	2	5	0	13	0
3 (Severe)	0	0	0	12	0	0	4	8

Table 6.15: Confusion matrices for infiltration classification using the out of sample database.

Class	TPR	FPR	F-measure	ROC area
Naive Bayes				
0 (Normal)	0.833	0.167	0.625	0.878
1 (Mild)	0.533	0.167	0.604	0.775
2 (Moderate)	0.556	0.093	0.606	0.889
3 (Severe)	1.000	0.033	0.923	1.000
Weighted average	0.667	0.126	0.661	0.858
AdaBoost				
0 (Normal)	0.917	0.433	0.449	0.874
1 (Mild)	0.133	0.024	0.229	0.676
2 (Moderate)	0.722	0.167	0.650	0.903
3 (Severe)	0.667	0.000	0.800	0.989
Weighted average	0.500	0.124	0.466	0.818

Table 6.16: Performance of infiltration classification using the out of sample database.

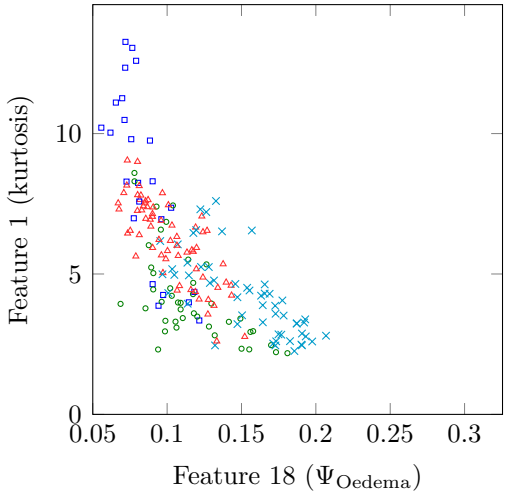


Figure 6.9: Oedema features 1 and 18 *from the training database.*

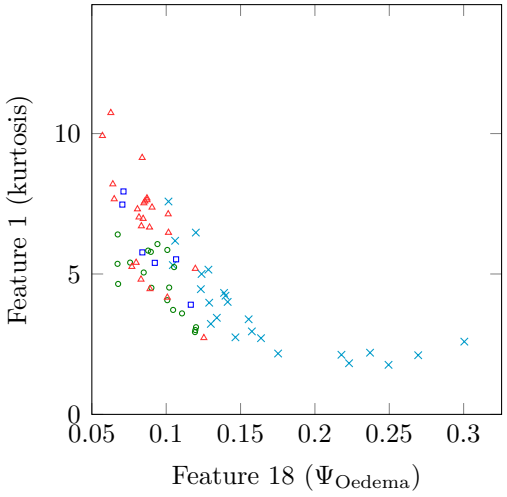


Figure 6.10: Oedema features 1 and 18 *from the out of sample database.*

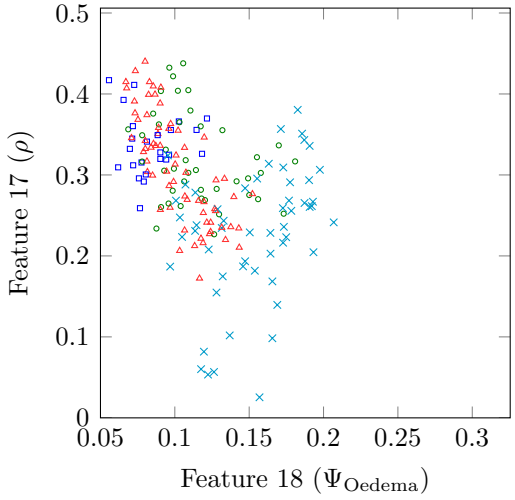


Figure 6.11: Oedema features 17 and 18 *from the training database.*

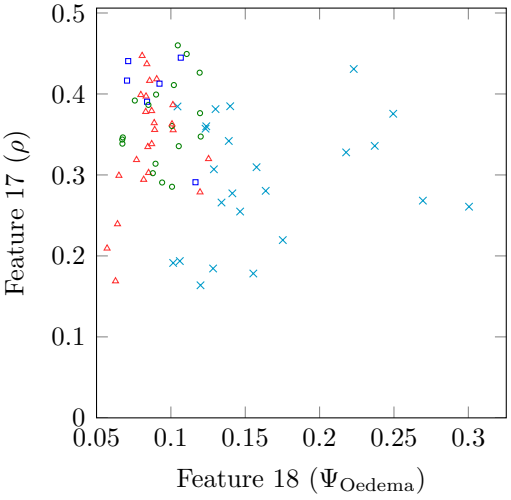


Figure 6.12: Oedema features 17 and 18 *from the out of sample database.*

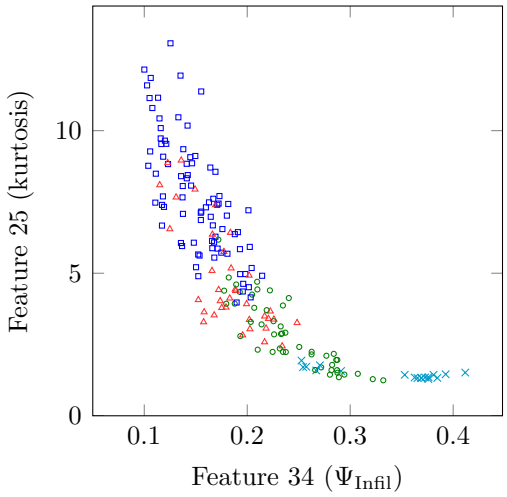


Figure 6.13: Infiltration features 25 and 34 *from the training database.*

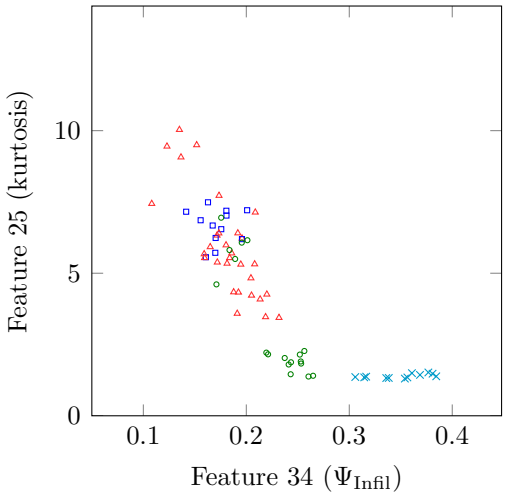


Figure 6.14: Infiltration features 25 and 34 *from the out of sample database.*

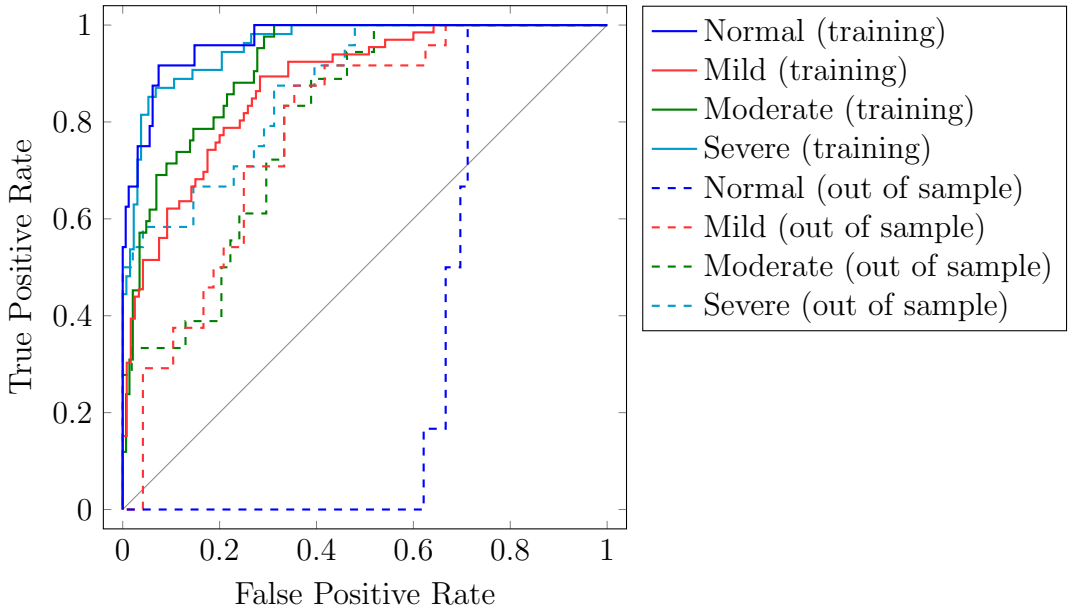


Figure 6.15: ROC curves per class for Naive Bayes (oedema classification).

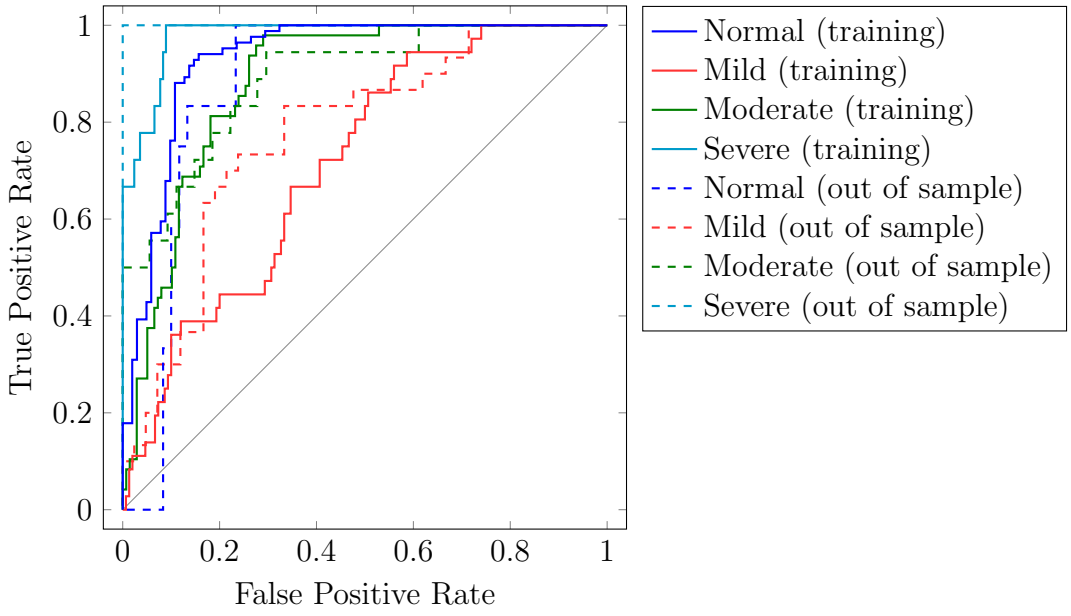


Figure 6.16: ROC curves per class for Naive Bayes (infiltration classification).

6.10 Conclusion

In this chapter, two feature vectors were designed with features from the joint intensity space and the disease metrics (which were defined in Chapter 5). Automatic classification using the out of sample database was successful with these vectors. The results for the Naive Bayes classifier were marginally superior to AdaBoost, especially with respect to the "within one" score. From these results, it is clear that the joint intensity space of T1 and STIR signal contains sufficient data to automatically classify disease state, both for oedema and fat infiltration.

Chapter 7

Conclusions

Very little literature exists that use computer-aided methods to approach disease quantification or automatic diagnosis for inflammatory muscle disease (IIM). One previous work attempted to fit a Gaussian distribution to the STIR signal intensity histogram. However, when inflammatory muscle disease is suspected, the standard protocol dictates that both T1 and STIR sequences be acquired such that the level of oedema and fat infiltration may be assessed. When these sequences are acquired with identical slice locations then it is known that the joint-intensity for muscle tissue deviates in the presence of disease. Therefore, it is natural to assume (with little to no alteration to the standard protocol) that both acquisitions may be combined for superior analysis. However, with very little previous research to build upon, the quantification methods presented in this work have been designed with very few preconceptions, and therefore they are largely based on the expected shape of the joint distribution. Confidence in the proposed methods is increased for two reasons. First, the expected joint distribution is seen to deviate more for subjects with increased disease, both for oedema and fat infiltration. Second, the fundamental reason for signal intensity change is a result of composition changes in the muscle tissue, with associated changes in physical properties, and hence the altered signal response from the MRI scanner.

Since all scans in the databases have been obtained clinically, the methods have been constructed with appreciation for the quality level expected from routine clinical scans. In the hypothetical scenario where very high-quality MRI scans have been obtained with no MR artefacts, it is assumed that the disease metrics would be highly sensitive, if blood vessels are removed from consideration. However, this is not practical in a clinical setting. The applicability of the quantification methods proposed are much greater given that they have been constructed with imperfect acquisitions and the fact that multiple MRI units have been utilised.

It is clear that any computer-aided analysis of MR data, for any pathology, must be appropriately pre-processed due to a plethora of interacting physical phe-

nomena present during acquisition. Without intensity correction methods, results cannot be trusted as reliable or robust. In the work presented in this thesis, the MRI scans have been very carefully pre-processed, especially with regards to over-standardisation of intensity. The intensity scaling method presented was designed directly from the need to compare joint histograms between subjects. Data which are not useful for diagnosis, or may otherwise be confounding, have been removed. For instance, a specialised tool has been designed to extract the femur bone semi-automatically, where only one seed point is required per bone for the entire set of slices. Although the method has only been tested on sets of 6 slices over the upper leg region, it is expected that the method would work adequately for a greater number of slices, or for a different body region, e.g. to extract the humerus bone from scans of the upper arm. It may not work sufficiently for more complex shapes such as the hip bone, since only one bone region is expected per slice.

Within the research presented in this thesis, quantification methods for oedema and fat infiltration have been designed and shown to have good positive correlation with expert medical opinion. The oedema selection method used empirical tests to find the approximate threshold size of contiguous objects in the initial region selection that were not meaningful to the overall disease value. In validation testing, the quantification methods have shown a good level of repeatability. However, in general, quantification had considerable class overlap for both oedema and fat infiltration problems. Class overlap may be exacerbated by the effect of a single medical expert rater (with unknown variance) and the fact that only a single grade is available per case of 6 slices. It is expected that some slices contain less disease than others due to "patches" of disease, and therefore a range of disease quantities per case is expected.

The proposed automatic classification of expert grade was highly successful, particularly with regards to the out of sample results. Different regions of the joint intensity space were useful in this respect and the results support the notion that specific regions have diagnostic value. Naive Bayes performed marginally better than AdaBoost (with J48), which may be due to the exact location of classes in feature space that are inappropriately segmented by the J48 tree. In general, classification of oedema was more difficult than classification of infiltration. This is almost certainly due to higher SNR in T1 slices and the confounding nature of blood vessels with MR artefacts and lower SNR in STIR slices.

It was shown that the semi-automatic quantification methods and automatic classifiers will allow the systematic assessment of patients with suspected inflammatory muscle disease in a clinical setting. These methods will provide doctors with robust and sensitive tools to measure disease state. In this way, it is ex-

pected that the overall variation in disease assessment will be reduced. With time, it is expected that the methods will improve and their use will become more widespread.

The research conducted within the scope of this thesis has contributed to the following publications. Further publications are planned in due course, as stated below.

Published:

- James Jack, Terence Jones, Charles Hutchinson, Helmut Bez, and Eran Edirisinghe, 'Automatic diagnosis of inflammatory muscle disease for MRI using computer-extracted features of bivariate histograms'. Proceedings of the International Society for Optics and Photonics, March 2015, pages 941402–941402-7.

Planned:

- James Jack, Charles Hutchinson, and Eran Edirisinghe, 'Computer-aided quantification and automatic classification of inflammatory muscle disease using clinically acquired MRI'. European Journal of Radiology. November 2015.

- James Jack, Eran Edirisinghe, and Charles Hutchinson, 'Quantification of oedema and fat infiltration for inflammatory muscle disease'. IEEE Transactions on Medical Imaging. December 2015.

With respect to the above discussions, it can be seen that the aims set out at the beginning have been achieved. However, some additional discussion on possible enhancements and conclusions is warranted.

The joint histogram space potentially contains more powerful features than those found with the current feature vectors. A thorough testing strategy over hundreds of regions could have produced interesting results. It is assumed that any feature vector will, on average, perform better over some regions of the joint intensity space. However, this rigorous level of testing can have a negative impact due to the possibility of over-fitting the data, and therefore it can only realistically be conducted with a substantially larger dataset.

In order to train the system more thoroughly, it would be advantageous to have multiple experts grade individual slices for oedema and fat infiltration. The average grade could then be taken as a decimal figure. However, this would be extremely laborious due to the total number of slices involved. Nevertheless, a small scale trial would provide better validation of the proposed quantification methods, particularly with regards to class overlap.

Further validation using a much larger set of re-scanned subjects would be highly beneficial, since it would allow the computation of confidence intervals for each disease metric. Scans of control subjects on different MRI scanners would be particularly useful.

The expert grade is a direct indication of the severity of oedema and fat infiltration found in the scan. This was done in order that the quantification and classification methods could be tested fairly. However, in reality, there may be other factors which affect the level of apparent disease. For instance, it may be true that the elderly typically have a higher baseline level of fat infiltration, as compared to the young adult. If this were the case, the original metrics could be adjusted to account for the quantity of disease above the expected baseline level, based on the age of the subject. For classification, it is possible that additional, clinically-relevant features would increase the usefulness of the prediction. Note that a new definition of "Normal" would be required, since it would no longer reflect an absolute level.

The infiltration metric may not give accurate results when used on very advanced fat infiltration cases. This is a result of the segmentation method, based on the joint histogram, and the physical changes which have taken place. Estimating the original quantity of muscle in the leg prior to fat infiltration is a difficult research question which should be addressed.

None of the databases included patients with juvenile myositis. Although the method presented would in theory work for this type of disease, it has not been tested. Ideally, another database of scans would be required to ascertain this possibility.

The manual process of masking out blood vessels from the input scans (normally found on STIR sequences) could be replaced by an automatic process. Since both blood vessels and oedema produce a hyperintense signal on STIR sequences, the removal of all blood vessels prior to analysis would clearly be beneficial. However, it was found that automatic methods from the literature were particularly prone to mislabelling oedema, or were otherwise erroneous due to the variability in presentation or interaction of MR artefacts. Vessel pulsation, for instance, often causes hyperintense signal extending horizontally or vertically away from the blood vessel, which may be selected as oedema. Therefore, blood vessel extraction is a major confounding issue and requires further research.

The semi-automatic method to extract the femur bone from transverse slices could be improved by being fully-automatic. It is assumed that an initial set of thresholds with shape analysis could adequately identify an initial seed point.

References

[1] U. E. Aladl and T. Peters. 'Medical image registration'. In: *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Ed. by A. S. El-Baz et al. Vol. 2. Springer New York, 2011, pp. 227–245. DOI: 10.1007/978-1-4419-8204-9_9.

[2] F. J. Anscombe. 'Graphs in statistical analysis'. In: *The American Statistician* 27.1 (Feb. 1973), p. 17. DOI: 10.2307/2682899.

[3] M. L. Bartlett et al. 'Quantitative assessment of myositis in thigh muscles using magnetic resonance imaging'. In: *Magnetic Resonance Imaging* 17.2 (Feb. 1999), pp. 183–191. DOI: 10.1016/S0730-725X(98)00092-7.

[4] P. Bearcroft. 'Imaging modalities in the evaluation of soft tissue complaints'. In: *Best Practice & Research Clinical Rheumatology* 21.2 (Apr. 2007), pp. 245–259. DOI: 10.1016/j.berh.2006.11.003.

[5] B. Belaroussi et al. 'Intensity non-uniformity correction in MRI: existing methods and their validation'. In: *Medical Image Analysis* 10.2 (Apr. 2006), pp. 234–246. DOI: 10.1016/j.media.2005.09.004.

[6] M. Boesen et al. 'MRI quantification of rheumatoid arthritis: current knowledge and future perspectives'. In: *European Journal of Radiology* 71.2 (Aug. 2009), pp. 189–196. DOI: 10.1016/j.ejrad.2009.04.048.

[7] H. Brighton and G. Gigerenzer. 'The bias bias'. In: *Journal of Business Research* 68.8 (Aug. 2015), pp. 1772–1784. DOI: 10.1016/j.jbusres.2015.01.061. (Visited on 21/09/2015).

[8] M. Brown and R. Semelka. *MRI: basic principles and applications*. 3rd ed. John Wiley & Sons, 2003.

[9] G. Brys, M. Hubert and A. Struyf. 'A robust measure of skewness'. In: *Journal of Computational and Graphical Statistics* 13.4 (Dec. 2004), pp. 996–1017. DOI: 10.1198/106186004X12632.

[10] J. Cheng and R. Greiner. 'Comparing Bayesian network classifiers'. In: *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 101–108.

[11] C. Damgaard and J. Weiner. 'Describing inequality in plant size or fecundity'. In: *Ecology* 81.4 (2000), pp. 1139–1142. DOI: 10.1890/0012-9658(2000)081[1139:DIIPSO]2.0.CO;2.

[12] B. M. Dawant, A. P. Zijdenbos and R. Margolin. 'Correction of intensity variations in MR images for computer-aided tissue classification'. In: *Medical Imaging, IEEE Transactions on* 12.4 (1993), pp. 770–781. DOI: 10.1109/42.251128.

[13] L. R. Dice. 'Measures of the amount of ecologic association between species'. In: *Ecology* 26.3 (July 1945), p. 297. DOI: 10.2307/1932409.

[14] M. M. Dimachkie. 'Idiopathic inflammatory myopathies'. In: *Journal of Neuroimmunology* 231.1-2 (Feb. 2011), pp. 32–42. DOI: 10.1016/j.jneuroim.2010.10.013.

[15] K. Doi. 'Computer-aided diagnosis in medical imaging: historical review, current status and future potential'. In: *Computerized Medical Imaging and Graphics* 31.4-5 (June 2007), pp. 198–211. DOI: 10.1016/j.compmedimag.2007.02.002.

[16] H. Ellis, B. M. Logan and A. Dixon. *Human sectional anatomy: atlas of body sections, CT and MRI images*. 3rd ed. London: Hodder-Arnold, 2007.

[17] A. Elnakib et al. 'Medical image segmentation: a brief survey'. In: *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Ed. by A. S. El-Baz et al. Vol. 2. Springer New York, 2011, pp. 1–39. DOI: 10.1007/978-1-4419-8204-9_1.

[18] T. Fawcett. 'An introduction to ROC analysis'. In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[19] A. Fischmann et al. 'Quantitative MRI can detect subclinical disease progression in muscular dystrophy'. In: *Journal of Neurology* 259.8 (Aug. 2012), pp. 1648–1654. DOI: 10.1007/s00415-011-6393-2.

[20] J. Flusser and B. Zitová. 'A comment on "a novel approach for the registration of weak affine images"'. In: *Pattern Recognition Letters* 34.12 (Sept. 2013), pp. 1381–1385. DOI: 10.1016/j.patrec.2013.04.024.

[21] A. F. Frangi et al. 'Multiscale vessel enhancement filtering'. In: *Medical image computing and computer-assisted interventation—MICCAI'98*. Springer, 1998, pp. 130–137. DOI: 10.1007/BFb0056195.

[22] Y. Freund and R. E. Schapire. 'Experiments with a new boosting algorithm'. In: *Machine Learning: Proceedings of the Thirteenth International Conference*. Vol. 96. 1996, pp. 148–156.

[23] A. A. Goshtasby. *Image registration principles, tools and methods*. London: Springer London, 2012.

[24] M. A. Hall. 'Correlation-based feature subset selection for machine learning'. PhD thesis. Hamilton, New Zealand: University of Waikato, 1998.

[25] J. A. Hanley and B. J. McNeil. 'The meaning and use of the area under a receiver operating characteristic (ROC) curve.' In: *Radiology* 143.1 (1982), pp. 29–36. DOI: 10.1148/radiology.143.1.7063747.

[26] R. H. Hashemi. *MRI: the basics*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2010.

[27] P. Hellier. 'Consistent intensity correction of MR images'. In: *Proc. International Conference on Image Processing 2003*. Vol. 1. IEEE, 2003, pp. 1109–1112. DOI: 10.1109/ICIP.2003.1247161.

[28] R. E. Hendrick. 'The AAPM/RSNA physics tutorial for residents. Basic physics of MR imaging: an introduction.' In: *RadioGraphics* 14.4 (July 1994), pp. 829–846. DOI: 10.1148/radiographics.14.4.7938771.

[29] J. W. Hennel. *A primer of magnetic resonance imaging*. London : River Edge, NJ: Imperial College Press; Distributed by World Scientific Pub. Co, 1997.

[30] S. Herlidou et al. 'Comparison of automated and visual texture analysis in MRI: characterization of normal and diseased skeletal muscle'. In: *Magnetic Resonance Imaging* 17.9 (1999), pp. 1393–1397. DOI: 10.1016/S0730-725X(99)00066-1.

[31] G. Hermosillo et al. 'Image registration in medical imaging: applications, methods, and clinical evaluation'. In: *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Ed. by A. S. El-Baz et al. Vol. 2. Springer New York, 2011, pp. 263–313. DOI: 10.1007/978-1-4419-8204-9_11.

[32] D. Hill et al. 'Medical image registration'. In: *Physics in Medicine and Biology* 46.3 (2001), R1–R45. DOI: 10.1088/0031-9155/46/3/201.

[33] Z. Hou. 'A review on MR image intensity inhomogeneity correction'. In: *International Journal of Biomedical Imaging* 2006 (2006), pp. 1–11. DOI: 10.1155/IJBI/2006/49515.

[34] C. Hutchinson. *A model radiologist*. Inaugural lecture. University Hospital Coventry and Warwickshire, Jan. 2013. URL: http://www2.warwick.ac.uk/fac/med/news/events/inaugural/archive/2013/chutchinson/ (visited on 19/08/2013).

[35] K. Hynynen et al. 'Temperature monitoring in fat with MRI'. In: *Magnetic Resonance in Medicine* 43.6 (June 2000), pp. 901–904. DOI: 10.1002/1522-2594(200006)43:6<901::AID-MRM18>3.0.CO;2-A.

[36] P. Jaccard. 'The distribution of the flora in the alpine zone'. In: *New Phytologist* 11.2 (1912), pp. 37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

[37] J. Jack et al. 'Automatic diagnosis of inflammatory muscle disease for MRI using computer-extracted features of bivariate histograms'. In: *Proc. SPIE*. Vol. 9414. Orlando, Florida, United States: International Society for Optics and Photonics, Mar. 2015, pp. 941402–941402-7. DOI: 10.1117/12.2076757.

[38] F. Jäger and J. Hornegger. 'Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging'. In: *IEEE Transactions on Medical Imaging* 28.1 (Jan. 2009), pp. 137–150. DOI: 10.1109/TMI.2008.2004429.

[39] F. Jäger et al. 'A new method for MRI intensity standardization with application to lesion detection in the brain'. In: *Vision Modeling and Visualization*. Ed. by L. Kobbelt et al. Berlin: Aka GmBH, 2006, pp. 269–276.

[40] T. Jerman et al. 'Beyond Frangi: an improved multiscale vesselness filter'. In: *Proc. SPIE*. Vol. 9413. Orlando, Florida, United States: International Society for Optics and Photonics, 2015, 94132A–94132A-11. DOI: 10.1117/12.2081147.

[41] Z. Ji et al. 'Generalized rough fuzzy c-means algorithm for brain MR image segmentation'. In: *Computer Methods and Programs in Biomedicine* 108.2 (Nov. 2012), pp. 644–655. DOI: 10.1016/j.cmpb.2011.10.010.

[42] A. Kassner and R. E. Thornhill. 'Texture analysis: a review of neurologic MR imaging applications'. In: *American Journal of Neuroradiology* 31.5 (May 2010), pp. 809–816. DOI: 10.3174/ajnr.A2061.

[43] F. Khalifa et al. 'State-of-the-art medical image registration methodologies: a survey'. In: *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Ed. by A. S. El-Baz et al. Vol. 1. Springer US, 2011, pp. 235–280. DOI: 10.1007/978-1-4419-8195-0_9.

[44] H. K. Kim et al. 'T2 mapping in Duchenne muscular dystrophy: distribution of disease activity and correlation with clinical assessments'. In: *Radiology* 255.3 (2010), pp. 899–908. DOI: 10.1148/radiol.10091547.

[45] J. J. Kim et al. 'The utility of magnetic resonance imaging in inflammatory myopathy'. In: *Journal of Rheumatic Diseases* 20.5 (2013), p. 297. DOI: 10.4078/jrd.2013.20.5.297.

[46] T. König et al. 'Ultrasound texture-based CAD system for detecting neuromuscular diseases'. In: *International Journal of Computer Assisted Radiology and Surgery* 10.9 (Sept. 2015), pp. 1493–1503. DOI: 10.1007/s11548-014-1133-6.

[47] C. Köse, O. Gençalioğlu and U. Şevik. 'An automatic diagnosis method for the knee meniscus tears in MR images'. In: *Expert Systems with Applications* 36.2 (Mar. 2009), pp. 1208–1216. DOI: 10.1016/j.eswa.2007.11.036.

[48] C. Lo et al. 'Effect of magnetic resonance imaging on core body temperature in anaesthetised children'. In: *Anaesthesia and Intensive Care* 42.3 (2014), pp. 333–339. URL: http://www.aaic.net.au/Document/?D=20120359 (visited on 17/07/2015).

[49] T. Y. Lu, K. P. Ng and D. A. Isenberg. 'Inflammatory muscle disease assessment'. In: *Current rheumatology reports* 10.4 (2008), pp. 328–332. DOI: 10.1007/s11926-008-0052-3.

[50] A. Madabhushi and J. K. Udupa. 'Interplay between intensity standardization and inhomogeneity correction in MR image processing'. In: *IEEE Transactions on Medical Imaging* 24.5 (May 2005), pp. 561–576. DOI: 10.1109/TMI.2004.843256.

[51] J. B. A. Maintz and M. A. Viergever. 'A survey of medical image registration'. In: *Medical Image Analysis* 2.1 (Mar. 1998), pp. 1–36. DOI: 10.1016/S1361-8415(01)80026-8.

[52] A. C. Mamourian. *Practical MR physics: and case file of MR artifacts and pitfalls*. New York: Oxford University Press, 2010.

[53] D. Messineo et al. 'MRI in the study of distal primary myopathopies and of muscular alterations due to peripheral neuropathies: possible diagnostic capacities of MR equipment with low intensity field (0.2 T) dedicated to peripheral limbs'. In: *Magnetic Resonance Imaging* 16.7 (1998), pp. 731–741. DOI: 10.1016/S0730-725X(98)00080-0.

[54] B. Milanovic. 'A simple way to calculate the Gini coefficient, and some implications'. In: *Economics Letters* 56.1 (1997), pp. 45–49. DOI: 10.1080/1350485032000100279.

[55] L. G. Nyúl and J. K. Udupa. 'An approach to standardizing MR image intensity scale'. In: *Proc. SPIE*. Vol. 3658. San Diego, California, United States: International Society for Optics and Photonics, Feb. 1999, pp. 595–603. DOI: 10.1117/12.349472.

[56] L. G. Nyúl and J. K. Udupa. 'On standardizing the MR image intensity scale'. In: *Magnetic Resonance in Medicine* 42.6 (Dec. 1999), pp. 1072–1081. DOI: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[57] L. G. Nyúl, J. K. Udupa and X. Zhang. 'New variants of a method of MRI scale standardization'. In: *Medical Imaging, IEEE Transactions on* 19.2 (2000), pp. 143–150. DOI: 10.1109/42.836373.

[58] M. Petrén-mallmin et al. 'The effect of temperature on MR relaxation times and signal intensities for human tissues'. In: *MAGMA Magnetic Resonance Materials in Physics, Biology, and Medicine* 1.3-4 (Sept. 1993), pp. 176–184. DOI: 10.1007/BF01769420.

[59] R. A. Pooley. 'Fundamental physics of MR imaging'. In: *RadioGraphics* 25.4 (July 2005), pp. 1087–1099. DOI: 10.1148/rg.254055027.

[60] J. R. Quinlan. *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[61] B. Ramakrishna et al. 'An automatic computer-aided detection system for meniscal Tears on magnetic resonance images'. In: *IEEE Transactions on Medical Imaging* 28.8 (Aug. 2009), pp. 1308–1316. DOI: 10.1109/TMI.2009.2014864.

[62] G. X. Ritter and J. N. Wilson. *Handbook of computer vision algorithms in image algebra*. 2nd ed. Boca Raton: CRC Press, 2001.

[63] N. Robitaille et al. 'Tissue-based MRI intensity standardization: application to multicentric datasets'. In: *International Journal of Biomedical Imaging* 2012 (2012), pp. 1–11. DOI: 10.1155/2012/347120.

[64] A. Roche et al. 'The correlation ratio as a new similarity measure for multimodal image registration'. In: *Medical Image Computing and Computer-Assisted Interventation (MICCAI98)*. Springer, 1998, pp. 1115–1124. URL: http://dl.acm.org/citation.cfm?id=646921.709612 (visited on 09/02/2013).

[65] J. G. Sled, A. P. Zijdenbos and A. C. Evans. 'A nonparametric method for automatic correction of intensity nonuniformity in MRI data'. In: *Medical Imaging, IEEE Transactions on* 17.1 (1998), pp. 87–97. DOI: 10.1109/42.668698.

[66] Y. Tong et al. 'Interactive non-uniformity correction and intensity standardization of MR images'. In: *Proc. SPIE*. Vol. 9415. Orlando, Florida, United States: International Society for Optics and Photonics, Mar. 2015, 94151N–94151N-6. DOI: 10.1117/12.2082878.

[67] N. J. Tustison et al. 'N4ITK: improved N3 bias correction'. In: *IEEE Transactions on Medical Imaging* 29.6 (June 2010), pp. 1310–1320. DOI: 10.1109/TMI.2010.2046908.

[68] J. Van De Vlekkert et al. 'Combining MRI and muscle biopsy improves diagnostic accuracy in subacute-onset idiopathic inflammatory myopathy: combining MRI and muscle biopsy in IIM'. In: *Muscle & Nerve* 51.2 (Feb. 2015), pp. 253–258. DOI: 10.1002/mus.24307.

[69] D. J. Vining and G. W. Gladish. 'Receiver operating characteristic curves: a basic understanding.' In: *Radiographics* 12.6 (1992), pp. 1147–1154. DOI: 10.1148/radiographics.12.6.1439017.

[70] U. Vovk, F. Pernus and B. Likar. 'A review of methods for correction of intensity inhomogeneity in MRI'. In: *IEEE Transactions on Medical Imaging* 26.3 (Mar. 2007), pp. 405–421. DOI: 10.1109/TMI.2006.891486.

[71] S. Wang and R. M. Summers. 'Machine learning and radiology'. In: *Medical Image Analysis* 16.5 (July 2012), pp. 933–951. DOI: 10.1016/j.media.2012.02.005.

[72] D. Weishaupt, V. D. Köchli and M. Borut. *How does MRI work? An introduction to the physics and function of magnetic resonance imaging*. Translated from German. Berlin; New York: Springer, 2006.

[73] X. Wu et al. 'Top 10 algorithms in data mining'. In: *Knowledge and Information Systems* 14.1 (Jan. 2008), pp. 1–37. DOI: 10.1007/s10115-007-0114-2.

[74] T. S. Yoo. *Insight into images: principles and practice for segmentation, registration, and image analysis*. Wellesley, Mass.: A K Peters, 2004.

[75] Z. Zeng and R. Zwiggelaar. 'Joint histogram modelling for segmentation multiple sclerosis lesions'. In: *Computer Vision/Computer Graphics Collaboration Techniques*. Springer Berlin Heidelberg, 2011, pp. 133–144. DOI: 10.1007/978-3-642-24136-9_12.

[76] Y. Zhang et al. 'Medical image segmentation using new hybrid level-set method'. In: *Fifth International Conference BioMedical Visualization*. London: IEEE, July 2008, pp. 71–76. DOI: 10.1109/MediVis.2008.12.

[77] B. Zitová and J. Flusser. 'Image registration methods: a survey'. In: *Image and Vision Computing* 21.11 (Oct. 2003), pp. 977–1000. DOI: 10.1016/S0262-8856(03)00137-9.

Glossary

AD Alzheimer's Disease.

ADNI The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a large database of MRI scans and associated medical data.

AUC Area Under the Curve.

AUROC Area Under the Receiver Operating characteristic Curve.

CAD Computer Aided Detection or Computer Aided Diagnosis.

CADe Computer Aided Detection.

CADx Computer Aided Diagnosis.

CC Correlation Coefficient. A registration measure.

CK Creatine Kinase. A type of blood serum test.

CR Correlation Ratio. A registration measure.

CT Computed Tomography (X-ray).

DICOM The Digital Imaging and Communications in Medicine standard.

FNR False Negative Rate.

FOV Field of View.

FPR False Positive Rate.

FT Fourier Transform.

GLCM The Gray-Level Co-occurrence Matrix describes the number of transitions in gray-level between two pixels at a set distance and direction over an image.

IIM Idiopathic Inflammatory Myopathies. A group of systemic diseases with unknown cause.

infiltration Infiltration of muscle tissue (by fat).

IOI Intensity of Interest.

ITK Insight segmentation and registration ToolKit. An open source project for segmentation and registration. Available at: www.itk.org

MI Mutual Information. A registration measure.

MR Magnetic Resonance.

MRI Magnetic Resonance Imaging.

myopathy Muscle disease.

myositis A muscle disease of unknown cause, characterised by inflammation and muscle weakness.

oedema Inflammation or swelling; a build up of fluid in the body's tissues.

PDF Probability Density Function.

prevalence The prevalence is the proportion of a population that had a given disease, as measured at a certain point in time.

RF Radio Frequency. Part of the electromagnetic spectrum.

RLM Run-Length Matrix.

ROC Receiver Operating Characteristic.

ROI Region Of Interest.

sagittal The cardinal plane that goes from front to back, dividing a person into left and right halves.

sensitivity The percentage of real positive cases correctly identified as positive.

serum The component of blood without red blood cells, white blood cells, or clotting factor (fibrinogen).

SNR Signal to Noise Ratio. High signal to noise ratio is better.

specificity The percentage of real negative cases correctly identified as negative.

STIR A type of MRI imaging sequence: Short TI Inversion Recovery. The signal from fat is suppressed.

SVM Support Vector Machine.

T1 A type of MRI imaging sequence. It measures longitudinal relaxation.

T2 A type of MRI imaging sequence. It measures transverse relaxation.

TE Time to Echo. The amount of time to wait before measuring an echo from the RF pulse.

TPR True Positive Rate.

TR Time to Repeat. The amount of time to wait between RF pulses.

transverse The cardinal plane that divides the body into upper and lower parts.

US Ultrasound.