Automatic segmentation of knee cartilage using quantitative MRI data

Marcus Lind
Master of Science Thesis in Electrical Engineering

Automatic segmentation of knee cartilage using quantitative MRI data

Marcus Lind
LiTH-ISY-EX--17/5041--SE

Supervisor: Marcus Wallenberg
ISY, Linköping University

Marcel Warntjes
SyntheticMR AB

Examiner: Maria Magnusson
ISY, Linköping University

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping, Sweden

Copyright © 2017 Marcus Lind
Abstract

This thesis investigates if support vector machine classification is a suitable approach when performing automatic segmentation of knee cartilage using quantitative magnetic resonance imaging data. The data sets used are part of a clinical project that investigates if patients that have suffered recent knee damage will develop cartilage damage. Therefore the thesis also investigates if the segmentation results can be used to predict the clinical outcome of the patients.

Two methods that perform the segmentation using support vector machine classification are implemented and evaluated. The evaluation indicates that it is a good approach for the task, but the implemented methods needs to be further improved and tested on more data sets before clinical use.

It was not possible to relate the cartilage properties to clinical outcome using the segmentation results. However, the investigation demonstrated good promise of how the segmentation results, if they are improved, can be used in combination with quantitative magnetic resonance imaging data to analyze how the cartilage properties change over time or vary between knees.
Acknowledgments

I would like to take this opportunity to thank my supervisor Marcus Wallenberg for his support during the work of this thesis. I would also like to thank my examiner Maria Magnusson for her support and interest in my work, my supervisor Marcel Warntjes for his support and for giving me the opportunity to do my thesis work at SyntheticMR AB and Per Widholm who has given me valuable input.

Finally, I would like to thank some of the people who are, and always have been, there for me. So thank you Johanna, Jan-Olof, Annette and Erik for your support through everything.

Linköping, June 2017
Marcus Lind
# Contents

1 Introduction 1

1.1 Aim 2
1.2 Approach 2
1.3 Limitations 2

2 Theory 5

2.1 Magnetic Resonance Imaging 5
  2.1.1 Underlying physics 5
  2.1.2 Excitation and relaxation 6
  2.1.3 Spatial encoding 7
  2.1.4 Pulse sequences and image contrast 8
  2.1.5 Quantitative Magnetic Resonance Imaging 9

2.2 Anatomy of the knee 10
  2.2.1 Bones, cartilage and tissues 10

2.3 Theoretical background of SVM and SLIC 12
  2.3.1 Support vector machine 12
  2.3.2 Simple linear iterative clustering 13

2.4 Theoretical background of the evaluation metrics 14
  2.4.1 Precision 14
  2.4.2 Recall 14
  2.4.3 Specificity 15
  2.4.4 Matthews correlation coefficient 15
  2.4.5 Mutual information 15

3 Method 17

3.1 Classifiers 17

3.2 Generating data for the classifiers 17
  3.2.1 Annotation of data 17
  3.2.2 Femur segmentation 20
  3.2.3 Region of interest 21
  3.2.4 Features 21

3.3 Method 1 - Classification of individual voxels 25

3.4 Method 2 - Classification of SLIC regions 25
3.5 Training ......................................................... 26

4 Results ......................................................... 27
  4.1 Segmentation methods and classifiers ............................... 27
    4.1.1 Method 1 .................................................. 28
    4.1.2 Method 2 .................................................. 30
    4.1.3 Classifier with reduced number of features ................. 32
  4.2 Comparison of cartilage in healthy and damaged knees ........ 34

5 Discussion .................................................... 39
  5.1 Results ....................................................... 39
    5.1.1 Segmentation methods and classifiers ....................... 39
    5.1.2 Comparison of cartilage in healthy and damaged knees .... 40
  5.2 Method ....................................................... 41

6 Conclusions and future work ........................................ 43
  6.1 Conclusions .................................................. 43
  6.2 Future Work ................................................ 44

Bibliography ...................................................... 45
1

Introduction

Magnetic resonance imaging (MRI) is a medical imaging technique that is used to visualize the interior of the human body. MRI is able to generate images by using physical properties of the body’s hydrogen atoms in combination with several magnetic fields. The usage of MRI is very suitable for imaging of soft tissues since they have an abundance of hydrogen atoms, and it is the method of choice for evaluation of cartilage due to the results it yields.[1]

At the Center for Medical Image Science and Visualization (CMIV)\(^1\) in Linköping, a clinical project has begun that includes 120 patients with recent knee damage. The knees of these patients have been scanned using MRI and will be scanned again later on to see if the patients develop cartilage damage. This will result in a lot of data, and an automatic segmentation method that is able to extract the cartilage could therefore be helpful in the process of investigating if the cartilage properties change over time. The company SyntheticMR AB\(^2\) is interested in knowing if it is possible to perform the automatic segmentation of knee cartilage using quantitative magnetic resonance imaging (qMRI) data, which is the motivation behind this thesis.

The data used for this thesis was qMRI data sets of both left and right knees from four patients that are part of the clinical project. The imaging sequence used to generate the provided data was developed by SyntheticMR AB. The sequence differs from conventional MRI sequences in such way that the result of it is an absolute quantification of the tissue parameters \(T1\), \(T2\) and proton density (PD). The results acquired from the qMRI sequence are contained in three individual maps, one for each tissue parameter, for each slice. The advantage of qMRI compared to conventional MRI is that the intensities of the qMRI maps are directly related to the tissue properties, whereas the intensities of conventional

---

\(^1\)CMIV is a multidisciplinary research center within Linköping University [2].

\(^2\)SyntheticMR AB is a company situated in Linköping that develops software for MRI [3].
MRI images have no such property. This is because the intensities in conventional MRI images are dependent on the tissue parameters and machine parameters.[4]

1.1 Aim

The aim of the master’s thesis is to answer the following questions:

• How should an automatic segmentation algorithm for segmentation of knee cartilage be implemented when using qMRI data?

• How do the implemented segmentation methods perform compared to the ground truth?

• How can the properties of the cartilage be related to clinical outcome?

1.2 Approach

Two similar methods that aim to perform automatic segmentation of the cartilage were implemented and evaluated. For both methods the supervised learning method support vector machine (SVM) was used for classification. The difference between the two implemented methods was that one of the methods classified separate voxels as either cartilage or not, whereas the other method classified whole regions. The method that classified regions was implemented to investigate how classifying regions would affect the segmentation results, as well as the performance in terms of execution speed, compared to if individual voxels were classified. The regions that were classified were acquired by using simple linear iterative clustering (SLIC).

The results of the methods were evaluated by comparing them to a ground truth set, and the evaluation metrics that were calculated were precision, recall, specificity and Matthews correlation coefficient.

To investigate how the cartilage properties could be related to clinical outcome, the average $R1$, $R2$ and $PD$ were calculated using the segmentation results and compared between the healthy and damaged knees of two patients.

1.3 Limitations

When the work related to this thesis was carried out there was no ground truth available for the provided data sets. The ground truth that was used for both training of the SVM classifiers and the evaluation was therefore created using a semi-automatic segmentation algorithm. The created ground truth was deemed to be a reasonable representation of what was cartilage in the data sets by a radiologist at CMIV, but it was not entirely correct. This means that the results presented in this thesis serve as more of an indication of how the implemented methods perform when performing knee cartilage segmentation rather than a final verdict.
1.3 Limitations

To perform the automatic segmentation of knee cartilage the methods were limited to only using the qMRI maps. This meant that no synthetically produced MRI images were used as input throughout the work.

The possibility of drawing any conclusions from the property comparison was limited, since there was no information available regarding if any of the patients in the clinical project had developed cartilage damage yet.
This chapter provides some of the theory behind magnetic resonance imaging in order to understand the data that has been used, as well as a brief description of the relevant anatomy of the knee. It also contains some theoretical background of support vector machine classifiers and simple linear iterative clustering, which are used in the implemented segmentation methods, and some theoretical background of the measurements used for evaluation.

2.1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a medical imaging technique that is used to visualize the interior of the human body. It uses magnetic fields in combination with the spin property of the hydrogen atoms in the body to generate images. The theory in the following subsections, not including section 2.1.5, are explained with the help of [5].

2.1.1 Underlying physics

The signal that is used in MRI comes from hydrogen atoms. More specifically it is the nucleus of the hydrogen atom, a proton, that generates the signal. The proton has a spin where it rotates around its own axis, and it also has both mass and an electrical charge. These properties give the proton a magnetic moment which can be affected by an external magnetic field. The magnetic field of one hydrogen nucleus is very small, but since the total number of hydrogen atoms in the body is large, it is possible to generate a larger magnetic field. By default the net magnetization of the body's hydrogen nuclei is zero. This is because all of the magnetic fields of individual hydrogen nuclei are pointing in different directions and cancelling each other out. However, if an external magnetic field ($B_0$) is
applied, the protons will align with $B_0$ either parallel or anti-parallel, with a small majority aligning in the parallel direction. This majority causes a resulting net magnetization, which is represented by a vector denoted $M$. The applied magnetic field also causes the magnetic moment of the protons to precess around the direction of $B_0$. The rate at which this precession is conducted is known as the Larmor frequency, which is directly proportional to $B_0$. The Larmor frequency is defined as

$$\omega = \gamma B_0,$$

where $\omega$ is the Larmor frequency in the unit MHz, $\gamma$ is the gyromagnetic ratio in the unit MHz/T and $B_0$ is, as previously mentioned, the applied external magnetic field in the unit T. The value for $\gamma$ is different depending on the atom under consideration, but for hydrogen atoms the value is 42.58 MHz/T.

### 2.1.2 Excitation and relaxation

Since both $M$ and $B_0$ are pointing in the same direction when $B_0$ is applied, it is impossible to measure $M$ in this state. If $B_0$ is considered to be along a $z$-axis, then $M$ must be tipped away from $B_0$ and its initial state ($M_z$) into the transverse $xy$-plane ($M_{xy}$) in order to be measured. To accomplish this a radio frequency (RF) pulse that has the same frequency as the Larmor frequency of the hydrogen atom is used. By using the same frequency as the spinning hydrogen nuclei, the RF pulse is able to excite the hydrogen nuclei through resonance. The excitation causes $M$ to be flipped into the $xy$-plane, where it rotates around the $z$-axis (see figure 2.1). In this state $M$ is measurable, since it works as an electrical generator that is able to induce an alternating voltage in a receiver coil, which is the signal used in MRI.

![Figure 2.1: The effect of applying an RF pulse. The left image describes the orientation of M when only B_0 is applied, whereas the right image describes the orientation of M when the RF pulse is applied.](image-url)
When the RF pulse is turned off $M$ returns to its previous stable state of being aligned with $B_0$ along the z-axis. This process is known as relaxation, which itself can be divided into two independent processes known as $T_1$-relaxation and $T_2$-relaxation.

- $T_1$-relaxation describes the longitudinal relaxation, i.e. how $M_z$ is restored after the RF pulse is turned off. It is caused by the nuclei transferring the excess energy they got from being excited to surrounding molecules. $T_1$ itself is a time constant describing the time it takes for $M_z$ to recover to 63% of its maximum value.

- $T_2$-relaxation describes the transverse relaxation, i.e. how $M_{xy}$ decays after the RF pulse is turned off. Immediately after the RF pulse is turned off the spins are precessing in phase. However, with no RF pulse present, energy transfers start to occur between spins. These energy transfers cause the spins to dephase and cancel each other out, resulting in a loss of transverse magnetization. $T_2$ itself is a time constant that describes the time it takes for $M_{xy}$ to decay to 37% of its maximum value.

Both $T_1$ and $T_2$ are intrinsic features that vary depending on the tissue and the external magnetic field. Sometimes the relaxation rates $R_1$ and $R_2$ are used instead of the relaxation times $T_1$ and $T_2$ respectively, and the relation between them is that they are each others inverse.

### 2.1.3 Spatial encoding

In order to generate an image from the MRI signal, two important concepts used to locate where the signals originate from are slice selection and spatial encoding. Slice selection is used to make it so that the RF pulse only excites a single slice along $B_0$. Spatial encoding is used to make it possible to know from which spatial position an MRI signal originates from in the excited slice. With information regarding the signals’ origin it is possible to generate an image.

In order to accomplish slice selection, a gradient coil that makes $B_0$ inhomogeneous is applied (see figure 2.2). The gradient causes each slice of the object that is being affected by $B_0$ to correspond to a specific Larmor frequency. The thickness of the slices depends on the bandwidth of the RF pulse, and also if the applied gradient is shallow or steep, which generates thick or thin slices respectively.

The spatial encoding consists of phase encoding and frequency encoding, which are performed in the y- and x-direction respectively. The phase encoding is accomplished by applying a gradient after the spins have been excited by the RF pulse. This gradient alters the precession frequency of the protons. When the gradient is turned off, the precession frequency returns to that of the Larmor frequency, but the phase of the spins are then different and can be related to their location along the y-axis. The frequency encoding is accomplished by applying a gradient when the signal is being registered. This gradient alters the Larmor frequency of the spins so that they can be related to their position along the x-axis.
Figure 2.2: The effect that a gradient coil has on the external magnetic field. The left image shows how the external magnetic field is initially constant along the z-axis, whereas the right image shows how each location along the z-axis corresponds to a specific magnetic field strength when a gradient coil is applied.

2.1.4 Pulse sequences and image contrast

When the RF pulse and the different gradients are to be activated, as well as when the signal is to be registered, is described by a pulse sequence. There are many different pulse sequences available in MRI, and the image that is being generated have different tissues highlighted depending on which one is used. Two important parameters that are are present in pulse sequences, that have a large impact on the contrast of MRI images, are the repetition time (TR) and the echo time (TE).

TR is the time that passes between two RF pulses and TE is the time between a RF pulse and the signal being measured. Depending on how they are chosen an image can be weighted based on the different tissue parameters of the body, which are the previously mentioned $T_1$ and $T_2$, as well as the proton density ($PD$).

- By choosing a short TR, tissues with different $T_1$ will recover different amount of $M_z$. The signal that the tissues give off will then be related to how much $M_z$ has recovered, emphasizing their $T_1$ values. If a longer TR is used this effect becomes less prominent.

- By choosing a long TE, tissues with different $T_2$ will have dephased differently, causing them to have different $M_{xy}$. The signal that the tissues give off will then be related to how much $M_{xy}$ has decayed, emphasizing their $T_2$ values. If a short TE is used, this effect becomes less prominent.

Table 2.1 describes the different image weightings acquired depending on the used TR and TE. By looking at the table it is observed that if TR and TE are chosen so that the image is neither $T_1$ nor $T_2$ weighted, the image becomes $PD$ weighted.
Table 2.1: The different image weightings that are acquired depending on how TR and TE are chosen.

<table>
<thead>
<tr>
<th>Short TE</th>
<th>Long TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short TR</td>
<td>T1-weighting</td>
</tr>
<tr>
<td>Long TE</td>
<td>N/A</td>
</tr>
</tbody>
</table>

2.1.5 Quantitative Magnetic Resonance Imaging

Quantitative magnetic resonance imaging (qMRI) is a technique that can be used to get an absolute quantification of the tissue parameters $T_1$, $T_2$ and $PD$ for each voxel. To acquire qMRI data, specific pulse sequences that are able to quantitatively measure the tissue parameters are used. An example of such a sequence is the Quantication of Relaxation Times and Proton Density by Multiecho acquisition of a saturation-recovery using Turbo spin-Echo Readout (QRAPMASTER).[4]

The qMRI maps differ from images that are generated from conventional MRI. In conventional MRI, different types of weighted images are generated depending on how the scanner parameters, such as TR and TE, are chosen. Not only does this cause the generated images to be dependent on the scanner parameters, but it also has the effect that the intensities of the images have no direct meaning. The results acquired with qMRI on the other hand are directly related to the tissues that are being imaged and are independent of scanner parameters. As a result of this, the qMRI maps’ visual appearance is quite different compared to conventional MRI images, which sometimes limit their clinical use. However, since the qMRI maps contain the tissue parameters, along with the fact that the scanner parameters are chosen manually, it is possible to synthesize the signal of any kind of conventional image weighting from the qMRI maps if it is needed.[4]

An example of what the qMRI maps look like when imaging a knee is shown in figure 2.3.

Figure 2.3: The $PD$ (left), $R_1$ (center) and $R_2$ (right) maps of a knee. This image is a screenshot from the SyntheticMR software SyMRI.
2.2 Anatomy of the knee

Since the work of this thesis consisted of segmenting out the cartilage, some knowledge of the anatomy in the knee is necessary. Therefore, this section explains some of the relevant and prominent parts in the knee in order to provide some very basic knowledge of the anatomy.

2.2.1 Bones, cartilage and tissues

The knee is the region of the human body where the femur and tibia, which are also known as the thighbone and shinbone respectively, meet. Aside from these bones the patella, which is also known as the kneecap, is also present in the region. On the surface of the three mentioned bones, at the locations where interaction between them might occur, there is cartilage present. Cartilage is a connective tissue that serves the function of enabling the bones to slide over each other with ease. Other than bones, there is a lot of muscle and fat tissue present in the area. In figure 2.4 and 2.5, PD maps from two different slices of the same knee are shown in the sagittal plane with some relevant annotations.

Figure 2.4: PD map from a central slice of the knee with annotations.
Figure 2.5: PD map from a peripheral slice of the knee with annotations.
2.3 Theoretical background of SVM and SLIC

When studying literature related to the subject of this thesis it was found that automatic segmentation of cartilage had previously been performed on conventional MRI images by using methods such as supervised machine learning and texture analysis [6]. Since the methods had been performed on conventional MRI data and not qMRI data, they could not be directly applied, but could be used as a basis for performing the segmentation in this work. It was found that [7] and [8] had approached the problem by using support vector machine (SVM, see section 2.3.1) classifiers to varying extents, where some of the used features came from MRI images acquired from different sequences. This was somewhat comparable to how the qMRI maps each contain different information, which could be used to generate features. Therefore it was decided that the segmentation of cartilage that was to be performed in the work of this thesis would also use SVM classification.

The input to the SVM classifiers represented either individual voxels or small regions. When the input represented regions, the regions were created using simple linear iterative clustering (SLIC, see section 2.3.2).

This section explains the basic concepts of SVM classification and SLIC. The information provided in these sections can also be found in [9] and [10], respectively.

2.3.1 Support vector machine

SVM is a supervised machine learning model that can be used to perform either regression analysis or binary classification of data. When used for classification, as it was in this work, it takes data points that each have an n-dimensional feature vector as input. The SVM then predicts which class each data point belongs to with the help of a separating hyperplane. The hyperplane itself is found by letting the SVM train on data points that, apart from the previously mentioned n-dimensional feature vector, have a target value indicating which class they belong to.

It is possible to create different types of SVM classifiers. The most basic type is hard margin SVM classification, in which all training data points can be linearly separated. However, some of the more complicated classifiers are ones that work on non-linearly separable data, allows for some misclassification, or those that can do both.

When using hard margin SVM classification the data points $D$ used for training can be described by the following equation

$$ D = (x_i, y_i) \quad i = 1, ..., m, $$

where $x_i$ represents the n-dimensional feature vector, and $y_i$ is either equal to -1 or 1 depending on which class $x_i$ belongs to. Using these notations, the hyperplane $f$ can be defined as

$$ f(x) = w \cdot x + b, $$
where $w$ is a weight vector and $b$ is a bias. Both of these variables can be acquired in the process of training. For each point in $D$, the hyperplane separates the training data so that

$$f(x) = \begin{cases} 
\geq 1 & \text{if } y_i = 1 \\
\leq -1 & \text{if } y_i = -1
\end{cases}.$$  

(2.4)

In order to find the hyperplane that best separates the two classes, the distance between it and the closest data points is maximized. This distance is called the margin, and the data points closest to the hyperplane are called support vectors. The margin is calculated as $1/||w||$, which means that maximizing the margin is equal to minimizing $||w||$. The training is therefore an optimization problem, where $||w||$ is to be minimized while still satisfying equation (2.4). This problem can be solved using Lagrange multipliers, and solving it results in equation (2.3) becoming rewritten as

$$f(x) = \sum_t \alpha_t y_t x_t \cdot x - b,$$  

(2.5)

where $\alpha_t$ is a Lagrange multiplier, $x_t$ is a support vector, $y_t$ is the class of the support vector and $x$ is the data that is to be classified.

When working with non-linearly separable data, the training data points are mapped to a higher-dimensional space. By performing this operation the data can become linearly separable, and it is then possible to find a separating hyperplane and maximize its margin. When dealing with this type of data, a kernel function $K$ is used in order to simplify some of the calculations that come with having to map the data points into a higher dimensional space. The function that describes how non-linearly separable data is classified is defined as

$$f(x) = \sum_t \alpha_t y_t K(x_t, x) - b.$$  

(2.6)

Some commonly used kernel functions are the radial basis function (RBF) and the polynomial kernel functions.

### 2.3.2 Simple linear iterative clustering

SLIC is an algorithm which is able to create superpixels in an image based on the pixels' color similarities and proximity. The algorithm starts off by sampling a number of cluster centers that are regularly spaced throughout the image. The number of cluster centers are specified by the user as an input parameter $K$. This means that for an image with $N$ pixels, the size of each cluster will approximately be $N/K$. The centers are then moved to the position where the gradient is the lowest within an $n \times n$ neighborhood. This operation is performed in order to avoid placing the centers on for instance noise or edge pixels. All of the pixels that are not centers are then related to a cluster center based on its distance and similarity to them.
For each of the clusters created from this process, a new center is computed based on the belonging pixels’ color values and position. The process of relating the non-centers to a cluster center, followed by computing new cluster centers, is repeated until the distance between the new and the previous centers are below a set threshold.

### 2.4 Theoretical background of the evaluation metrics

In order to evaluate how the implemented segmentation methods performed, precision, recall, specificity (see section 2.4.1, 2.4.2 and 2.4.3, respectively) and Matthews correlation coefficient (MCC, see section 2.4.4) were used as evaluation metrics. These measures are calculated from the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), all of which can be acquired by comparing the results of a segmentation method with its corresponding ground truth (see table 2.2).

Since several features were going to be used by the SVM classifiers, it was thought to be interesting to analyze how the features fared compared to each other. Therefore the mutual information (MI, see section 2.4.5) was calculated for each feature and used as a comparable score.

**Table 2.2: Confusion matrix explaining how the number of TP, FP, TN and FN are acquired.**

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Prediction</th>
<th>Relevant element</th>
<th>Irrelevant element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant element</td>
<td>TP</td>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>Irrelevant element</td>
<td>FP</td>
<td>TN</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.4.1 Precision

The precision measure describes how many of the elements classified as relevant by a segmentation method that actually are relevant when compared to the ground truth. It is defined as

\[
Precision = \frac{TP}{TP + FP}.
\]  

#### 2.4.2 Recall

The recall measure describes how well a segmentation method is able to correctly classify the relevant elements present in the ground truth, regardless of the number of FP. It is defined as

\[
Recall = \frac{TP}{TP + FN}.
\]
2.4 Theoretical background of the evaluation metrics

Recall

\[ Recall = \frac{TP}{TP + FN}. \]  

(2.8)

2.4.3 Specificity

The specificity measure describes how well a segmentation method is able to correctly classify the irrelevant elements present in the ground truth, regardless of the number of FN. It is defined as

\[ Specificity = \frac{TN}{TN + FP}. \]  

(2.9)

2.4.4 Matthews correlation coefficient

MCC is a measure that is used in machine learning to describe the performance of a classifier. MCC is calculated using all of the information in the confusion matrix, and it is defined as

\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}. \]  

(2.10)

The possible values of MCC ranges from \(-1\) to \(1\). Getting a MCC equal to \(1\) corresponds to correct classification of all data points, whereas \(-1\) corresponds to the opposite. If MCC is equal to zero the performance of the classifier is no better than random prediction.[11]

2.4.5 Mutual information

MI is a measure that describes how much information one variable contains about another. In machine learning it can be used as a feature scoring method to get an indication of how important the used features are compared to each other, since it can describe how much information each feature contains about the sought after class. It is defined as

\[ MI = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right), \]  

(2.11)

where, in machine learning, \(x\) denotes the feature values, \(y\) denotes the class labels and \(p\) denotes probability. A high MI-score indicates that a feature contains a lot of information about the class, whereas a score of zero indicates that the feature is independent of the class label.[12]
This chapter contains a description of the used classifiers, how the data used by the classifiers was generated, the two methods that were implemented and the training of the classifiers.

3.1 Classifiers

In this work non-linear SVM classifiers were used to perform the classification. All of the used SVM classifiers were trained using the MatLab function `fitcsvm`, and an RBF kernel function was used.

3.2 Generating data for the classifiers

In total, ten qMRI data sets that had been acquired with a 3 T scanner were used. Eight of the ten data sets consisted of the left and right knee of four different patients. The remaining two data sets came from two of these patients who have had their damaged knee imaged again. Each data set consisted of 26 slices in the sagittal plane and the spacing between the slices was 3.3 mm. The pixel resolution varied between $512 \times 512$, $560 \times 560$ and $576 \times 576$, for which the pixel spacing was $0.3125$ mm, $0.2857$ mm and $0.2767$ mm, respectively. The $PD$ map of these data sets displayed the $PD$ as a value between zero and 180, where higher values corresponded to a higher proton density.

3.2.1 Annotation of data

Since manual annotation can be very time consuming, the annotated data was produced by creating and applying a semi-automatic segmentation method on
the qMRI data sets in the sagittal plane. A pipeline describing the implemented method is seen in figure 3.1, but the following bullet list contains a more detailed description:

- The normalized PD and R2 map (see figure 3.2a and 3.2b, respectively) are passed as input.

- A foreground mask is created by thresholding all non-zero values of the PD map. In the foreground mask the imaged leg are represented by ones and the background of the image is represented by zeros (see figure 3.2c).

- Using this mask, the foreground of the R2 map is then segmented using the MatLab function `imbinarize`, which performs adaptive thresholding (see figure 3.2d). The input parameter to `imbinarize` that specifies the sensitivity of the adaptive thresholding is given by the user.

- Noise and small clusters in the mask created in the previous step are removed using the MatLab function `bwareaopen` (see figure 3.2e). The minimum size of the objects that are to be kept is specified by the user.

- The MatLab function `roipoly` is then used on the PD map to let the user specify a region of interest (ROI) that includes the cartilage and excludes the muscle tissue (see figure 3.2f). The overlap between this ROI and the previous segmentation results is kept as a refined mask.

- The PD map is then segmented in the area specified by the refined mask using the MatLab function `imbinarize`, which performs adaptive thresholding (see figure 3.2g). The input parameter to `imbinarize` that specifies the sensitivity of the adaptive thresholding is given by the user.

- In a final step, noise and clusters that are present in the mask from the previous step are removed using the MatLab function `bwareaopen` (see figure 3.2h). The minimum size of the objects that are to be kept is specified by the user.

- The mask that remains after the previous operation is passed as output.

Some of the operations used in the annotation method, such as the adaptive thresholding, require input parameters. The parameters that were used differed, since the anatomy of the knee and the amount of noise differed between the slices and data sets that were used. How specific the ROI had to be placed around the cartilage also differed between slices and data sets because of the same reasons.

\[\text{Figure 3.1: Pipeline describing the how cartilage masks were created.}\]
Figure 3.2: Images from different parts of the annotation method.
3.2.2 Femur segmentation

Segmentation of the bones in the knee area was described as a common step by [6] when performing automatic segmentation of cartilage. It is performed because it comes with several benefits. Therefore it was decided to try to implement a bone segmentation method in this work.

Due to lack of previous literature on how bone segmentation is to be accomplished when using qMRI data, as well as time constraints, unfortunately only the femur bone was successfully segmented out. However, the femur was enough to come with benefits on its own. One of the benefits was that the femur could be used as an indication if cartilage was present at all in an image, since cartilage is only present on the surface of bones. Several images could therefore be dismissed from further classification by simply checking if the femur bone was present or not. A second benefit was that the femur could serve as a landmark, which helped narrow down the area of where cartilage could potentially be present. This is because the location of the cartilage is quite close to the surface of the femur. A third benefit was that it introduced the opportunity to use features related to the voxels’ distance to the femur.

The implemented femur segmentation method was inspired by the method described in [13], in which the bones in the knee are segmented out from conventional MRI images. The implemented method is performed on all slices in the sagittal plane and it is described as follows:

- The normalized $PD$, $T1$ and $R2$ map (see figure 3.3a, 3.3b and 3.3c, respectively) are passed as input.

- The $T1$ and $R2$ map are subtracted from the $PD$ map in order to get an image where the edges of some of the present objects are emphasized (see figure 3.3d).

- The MatLab function `edge` is used to perform canny edge detection on the results of the previous step in order to acquire a binary mask of the edges (see figure 3.3e).

- The edge mask is then subtracted from the $PD$ map, and all values that go below zero are set to zero (see figure 3.3f).

- The result from the previous step is subtracted by the $R2$ map, and all values that are still above zero are set to one, whereas those that are below are set to zeros, resulting in a mask (see figure 3.3g).

- The mask is then subtracted by a thresholded version of the $T1$ map, which removes some irrelevant parts from the mask (see figure 3.3h).

- In order to separate objects that are still connected, the mask is eroded using the MatLab function `imerode`, and some of the top and bottom rows of the mask that are of no interest and might prevent the femur segmentation are removed (see figure 3.3i).
3.2 Generating data for the classifiers

- Based on the assumption that the object representing the femur will have its bounding box filled to a large extent, be located in the top half of the binary mask, and that its centroid is close to the middle of the top half, the object that best fulfill these constraints is marked as being the femur (see figure 3.3j).

- The centroid of the marked object is used as a seed point for a region growing algorithm. The region growing algorithm is then performed on a cropped version of the PD map that has been subtracted by the T1 and R2 map (i.e. a cropped version of the image seen in figure 3.3d).

- The result of the region growing algorithm (see figure 3.3k), as well as the coordinates of the point located in the bottom middle of the femur’s bounding box, are passed as output.

The region growing algorithm that was used in the method was downloaded from Mathworks file exchange [14].

3.2.3 Region of interest

Not the complete images of the data sets were used when training and classifying images, but instead only the voxels within a ROI were used. For each image, the bottom-middle of the femur mask’s bounding box was used as an origin to divide the image into quadrants (see figure 3.4). The ROI then included voxels within 5 mm of the femur in the first quadrant, and voxels within 20 mm of the femur in the remaining quadrants. These distances were chosen after having created and observed the value distributions of the distance between the femur and the cartilage in each quadrant (see figure 3.5).

3.2.4 Features

For each voxel a total of 21 features that could be used by the SVM classifiers were calculated from the qMRI maps. All of the features were calculated from slices in the sagittal plane, which meant that the classification also had to be performed on slices in the sagittal plane. The features that were used and their MI-score (see section 2.4.5) are presented in decreasing order of rank in table 3.1.

The PD, R1 and R2 values were acquired directly from the qMRI maps. The means were acquired by convolving the qMRI maps with a 5 × 5 sized box filter. The variances, standard deviations and medians were calculated from the qMRI maps in 5 × 5 regions. The gradient magnitudes were acquired using the MatLab function `imgradient` on the qMRI maps. The feature describing the voxels’ distance from the femur bone was acquired using the MatLab function `bwdist`, which creates a distance map, on the femur mask. The origin, to which the x- and y-coordinates were related to, was set to be the bottom-middle of the femur mask’s bounding box in each slice (see the red circle in figure 3.4). An example of what three of the feature maps look like for one slice is shown in figure 3.6.
Figure 3.3: Images from different parts of the femur segmentation method.
3.2 Generating data for the classifiers

Figure 3.4: An example of how one of the slices was divided into quadrants. The red circle marks the bottom-middle of the femur mask’s bounding box and the green lines separates the quadrants.

Figure 3.5: Value distributions of the distance between the femur and the cartilage in the different quadrants.
Table 3.1: Features used in the classifier, their MI-score and corresponding rank when compared to each other.

<table>
<thead>
<tr>
<th>Feature</th>
<th>MI-score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_2$ median</td>
<td>0.14391</td>
<td>1</td>
</tr>
<tr>
<td>$R_2$ mean</td>
<td>0.14236</td>
<td>2</td>
</tr>
<tr>
<td>$R_2$ value</td>
<td>0.14183</td>
<td>3</td>
</tr>
<tr>
<td>$PD$ median</td>
<td>0.11032</td>
<td>4</td>
</tr>
<tr>
<td>$PD$ mean</td>
<td>0.10849</td>
<td>5</td>
</tr>
<tr>
<td>$R_2$ variance</td>
<td>0.10769</td>
<td>6</td>
</tr>
<tr>
<td>$PD$ mean</td>
<td>0.10352</td>
<td>7</td>
</tr>
<tr>
<td>$R_2$ standard deviation</td>
<td>0.08440</td>
<td>8</td>
</tr>
<tr>
<td>Distance to the femur</td>
<td>0.07460</td>
<td>9</td>
</tr>
<tr>
<td>$R_1$ variance</td>
<td>0.06222</td>
<td>10</td>
</tr>
<tr>
<td>$R_2$ gradient magnitude</td>
<td>0.05443</td>
<td>11</td>
</tr>
<tr>
<td>y-coordinate related to local origin</td>
<td>0.04378</td>
<td>12</td>
</tr>
<tr>
<td>$R_1$ standard deviation</td>
<td>0.03580</td>
<td>13</td>
</tr>
<tr>
<td>$R_1$ mean</td>
<td>0.03341</td>
<td>14</td>
</tr>
<tr>
<td>$R_1$ median</td>
<td>0.03096</td>
<td>15</td>
</tr>
<tr>
<td>$R_1$ value</td>
<td>0.02717</td>
<td>16</td>
</tr>
<tr>
<td>$R_1$ gradient magnitude</td>
<td>0.02244</td>
<td>17</td>
</tr>
<tr>
<td>x-coordinate related to local origin</td>
<td>0.01224</td>
<td>18</td>
</tr>
<tr>
<td>$PD$ variance</td>
<td>0.00354</td>
<td>19</td>
</tr>
<tr>
<td>$PD$ gradient magnitude</td>
<td>0.00239</td>
<td>20</td>
</tr>
<tr>
<td>$PD$ standard deviation</td>
<td>0.00236</td>
<td>21</td>
</tr>
</tbody>
</table>

Figure 3.6: The feature maps for the $PD$ median (left), the $R_2$ gradient magnitude (middle), and the distance to the femur (right).
3.3 Method 1 - Classification of individual voxels

Method 1 takes the qMRI maps as input and starts off by performing the femur segmentation. If no femur is found a matrix of zeros is returned as output, which corresponds to no cartilage being present in the image. This is because the femur is needed in order for cartilage to be present, and also because some of the features are calculated using the femur. If, on the other hand, the femur found, the features needed by the classifier are calculated for each voxel within the ROI defined in section 3.2.3. The data points are then passed to the classifier and the results are returned in a binary mask where the cartilage voxels are represented by ones and the non-cartilage voxels by zeros. The pipeline of method 1 can be seen in figure 3.7, where it follows the black and red arrows.

3.4 Method 2 - Classification of SLIC regions

Method 2 functions in the exact same way as method 1 until after the ROI has been created. After the ROI is created, method 2 proceeds to perform SLIC on a modified version of the PD map. The modification consists of setting every voxel outside of the ROI to zero. The features are then calculated for every voxel that belongs to a SLIC region within the ROI. With all features calculated for these voxels, the method proceeds to average the individual feature values over each region. The averaging makes it so that each of the data points being passed to the classifier correspond to one of the SLIC regions. Just as for method one, the classification results are returned in a binary mask where the cartilage voxels are represented by ones and the non-cartilage voxels by zeros. The pipeline of method 2 can be seen in figure 3.7, where it follows the black and blue arrows.

Figure 3.7: Pipeline depicting how the two methods are performing the segmentation. Method 1 follows the black and red arrows, whereas method 2 follows the black and blue arrows.
3.5 Training

Initially eight SVM classifiers were trained and used in the two segmentation methods. This came from having four permutations of used features and two amounts of data when training the classifiers. One classifier used all of the 21 features, and the remaining three each did not use the six features acquired from the PD, R1 and R2 map respectively. The amount of training data was either 50% of the available training data for each class in each slice, or all of it. The reason for training these eight classifiers was to further analyze the significance of the features from the different qMRI maps, and also to analyze how the amount of training data affected the results.

Based on the results of the eight classifiers, as well as some additional calculations that are further described in section 4.1.3, one last classifier that aimed to reduce the classification time was trained and used in the two segmentation methods. This classifier did not use the features that appeared to be insignificant, which reduced the number of used features to seven.

The training data used was five of the qMRI data sets (see section 3.2), provided by two different patients. This approach was chosen so that the training data would consist of both left and right knees, as well as one damaged knee that had been imaged three months later.
In this chapter the results of the implemented segmentation methods and the comparison of properties between healthy and damaged knees are presented.

The data sets used to obtain results and evaluate the two methods, as well as to compare properties, were the five data sets that had not been used to train the classifiers. These data sets consisted of the left and right knee of two patients, as well as a rescan of one of the patients’ damaged knee, which added up to a total of 130 slices.

### 4.1 Segmentation methods and classifiers

As mentioned in section 3.5, eight classifiers were trained initially and one final classifier was trained based on the results of the initial eight. The results obtained using method 1 and 2 with the eight initially trained classifiers are presented in section 4.1.1 and 4.1.2, respectively. The results obtained using method 1 and 2 with the final classifier are presented in section 4.1.3. Section 4.1.3 also details the subset of seven features used by the final classifier, as well as the reasons for this selection.

In addition to the evaluation metrics, two other values are presented for each method and classifier. The first is the number of times that the predictions contain cartilage when it is not present in the ground truth, or vice versa. This is referred to as complete misses (CM). The second is the time it takes to classify all data sets. This is referred to as classification time.

In the tables of the following subsections the precision, recall and specificity are denoted P, R and S respectively. The presented classification times were obtained using a laptop that has 8 GB of RAM and an Intel(R) Core(TM) i7 – 2670QM CPU, which has four cores and a clock rate of 2.20 GHz per core.
4.1.1 Method 1

Table 4.1 and 4.2 present the results of method 1 with the classifiers that were trained on 50% and 100% of the training data, respectively. Figure 4.1 displays the TP, FP, TN and FN obtained when using method 1, with the classifiers trained on 100% of the training data, on three slices.

As can be seen in table 4.1 and 4.2, the MCC is higher when all features or no R1 features are being used compared to if no PD features or no R2 features are being used. It is also seen that there is a big difference in classification time depending on the amount of training data used.

Table 4.1: Results of method 1 with classifiers that were trained on 50% of the training data.

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>MCC</th>
<th>#CM</th>
<th>Classification time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>0.7433</td>
<td>0.7148</td>
<td>0.9980</td>
<td>0.7198</td>
<td>7</td>
<td>11160</td>
</tr>
<tr>
<td>No PD features</td>
<td>0.7103</td>
<td>0.6883</td>
<td>0.9976</td>
<td>0.6806</td>
<td>7</td>
<td>11240</td>
</tr>
<tr>
<td>No R1 features</td>
<td>0.7340</td>
<td>0.7248</td>
<td>0.9979</td>
<td>0.7176</td>
<td>7</td>
<td>8849.8</td>
</tr>
<tr>
<td>No R2 features</td>
<td>0.7267</td>
<td>0.5468</td>
<td>0.9985</td>
<td>0.6213</td>
<td>7</td>
<td>12468</td>
</tr>
</tbody>
</table>

Table 4.2: Results of method 1 with classifiers that were trained on 100% of the training data.

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>MCC</th>
<th>#CM</th>
<th>Classification time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>0.7425</td>
<td>0.7180</td>
<td>0.9980</td>
<td>0.7211</td>
<td>7</td>
<td>22087</td>
</tr>
<tr>
<td>No PD features</td>
<td>0.7103</td>
<td>0.6915</td>
<td>0.9976</td>
<td>0.6866</td>
<td>7</td>
<td>21538</td>
</tr>
<tr>
<td>No R1 features</td>
<td>0.7321</td>
<td>0.7235</td>
<td>0.9979</td>
<td>0.7160</td>
<td>7</td>
<td>17173</td>
</tr>
<tr>
<td>No R2 features</td>
<td>0.7261</td>
<td>0.5489</td>
<td>0.9984</td>
<td>0.6221</td>
<td>7</td>
<td>23422</td>
</tr>
</tbody>
</table>
4.1 Segmentation methods and classifiers

(a) Results from using the classifier that uses all features.

(b) Results from using the classifier that does not use any features related to PD.

(c) Results from using the classifier that does not use any features related to R1.

(d) Results from using the classifier that does not use any features related to R2.

Figure 4.1: The TP (red), FP (yellow), TN (black) and FN (yellow) obtained when using method 1, with the classifiers that were trained on 100% of the training data, on three slices.
### 4.1.2 Method 2

Table 4.3 and 4.4 present the results of method 2 with the classifiers that were trained on 50% and 100% of the training data, respectively. Figure 4.2 displays the TP, FP, TN and FN obtained when using method 2, with the classifiers trained on 100% of the training data, on three slices.

For all of the presented results, the input parameter to SLIC, i.e. the number of cluster centers, was specified to 20,000.

As can be seen in table 4.3 and 4.4, just as for method 1 the MCC is higher when all features or no $R_1$ features are being used compared to if no $PD$ features or no $R_2$ features are being used. It is also seen that there is a big difference in classification time depending on the amount of training data used, but it is not as big as it is for method 1.

**Table 4.3: Results of method 2 with classifiers that were trained on 50% of the available training data.**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>MCC</th>
<th>#CM</th>
<th>Classification time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>0.7020</td>
<td>0.7089</td>
<td>0.9975</td>
<td>0.6958</td>
<td>7</td>
<td>4162.5</td>
</tr>
<tr>
<td>No $PD$ features</td>
<td>0.6832</td>
<td>0.6785</td>
<td>0.9972</td>
<td>0.6650</td>
<td>7</td>
<td>4137.3</td>
</tr>
<tr>
<td>No $R_1$ features</td>
<td>0.7033</td>
<td>0.7103</td>
<td>0.9975</td>
<td>0.6948</td>
<td>7</td>
<td>3947.6</td>
</tr>
<tr>
<td>No $R_2$ features</td>
<td>0.6960</td>
<td>0.5229</td>
<td>0.9982</td>
<td>0.5935</td>
<td>7</td>
<td>4162.8</td>
</tr>
</tbody>
</table>

**Table 4.4: Results of method 2 with classifiers that were trained on 100% of the training data.**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>MCC</th>
<th>#CM</th>
<th>Classification time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>0.6992</td>
<td>0.7135</td>
<td>0.9975</td>
<td>0.6969</td>
<td>7</td>
<td>4791.9</td>
</tr>
<tr>
<td>No $PD$ features</td>
<td>0.6827</td>
<td>0.6841</td>
<td>0.9972</td>
<td>0.6678</td>
<td>7</td>
<td>4636.1</td>
</tr>
<tr>
<td>No $R_1$ features</td>
<td>0.6995</td>
<td>0.7125</td>
<td>0.9975</td>
<td>0.6940</td>
<td>7</td>
<td>4351.0</td>
</tr>
<tr>
<td>No $R_2$ features</td>
<td>0.7006</td>
<td>0.5263</td>
<td>0.9982</td>
<td>0.5974</td>
<td>7</td>
<td>4747.1</td>
</tr>
</tbody>
</table>
4.1 Segmentation methods and classifiers

(a) Results from using the classifier that uses all features.

(b) Results from using the classifier that does not use any features related to PD.

(c) Results from using the classifier that does not use any features related to R1.

(d) Results from using the classifier that does not use any features related to R2.

Figure 4.2: The TP (red), FP (yellow), TN (black) and FN (yellow) obtained when using method 2, with the classifiers that were trained on 100% of the training data, on three slices.
4.1.3 Classifier with reduced number of features

As previously mentioned in section 3.5, this classifier was built with the intention of getting a shorter classification time than those of the eight initially trained classifiers. This was to be accomplished by removing the features that appeared to be insignificant and thus mainly added to the complexity of the classifiers.

The seven features used by the classifier were the $R^2$ and PD median, $R^2$ variance, $R^2$ standard deviation, distance to femur, $R^2$ gradient magnitude, and y-coordinate related to local origin. The reason for not using any of the features related to $R^1$, is because the results of the eight initially trained classifiers indicates that they are insignificant. These features being insignificant also suggests that the four features with worse MI than all $R^1$ features could be insignificant as well, which is why they are not used. The reason for not using the voxel and mean value from either of the PD or $R^2$ map is because they, for each individual map, are heavily correlated with each other and the median feature. The correlation indicates that they provide the same information, and that using only one of them from each map should be sufficient. Therefore only the median features from each map was kept since they have the highest MI.

The classifier was trained on 50% of the training data. This amount was used because the results of the eight initially trained classifiers showed that it heavily reduced the classification time, while achieving only slightly inferior results.

Table 4.5 presents the results of method 1 and 2 with the classifier that did not use the seemingly insignificant features. Figure 4.3 visualizes the TP, FP, TN and FN obtained when using method 1 and 2 with this classifier on three slices.

For the presented results of method 2, the input parameter to SLIC, i.e. the number of cluster centers, was specified to 20,000.

| | | | | | |
|---|---|---|---|---|
| P | R | S | MCC | #CM |
| **Method 1** | 0.7169 | 0.6766 | 0.9977 | 0.6823 | 7 |
| **Method 2** | 0.6898 | 0.6609 | 0.9974 | 0.6593 | 7 |
| **Classification time [s]** | | | | | 9683.3 |

Table 4.5: Results of method 1 and 2 when using the classifier with reduced number of features.
4.1 Segmentation methods and classifiers

(a) Results from using the classifier with seven features in method 1.

(b) Results from using the classifier with seven features in method 2.

**Figure 4.3:** The TP (red), FP (yellow), TN (black) and FN (yellow) obtained when using method 1 and 2, with the classifier that only used seven features, on three slices.
4.2 Comparison of cartilage in healthy and damaged knees

The diagrams in figure 4.4, 4.5 and 4.6 present the average PD, R1 and R2 respectively, of the cartilage for each patient and knee. The presented values in these figures are calculated from the qMRI maps, using the segmentation results of method 1 with the classifier that use all features and is trained on 100% of the training data. The results of this method and classifier are used since it has the highest MCC.

For patient 1, the cartilage in the damaged knee has a lower average PD, R1 and R2 than the cartilage in the healthy knee. For patient 2, the cartilage in the damaged knee initially has a lower average R1 and R2, but a higher PD, than the cartilage in the healthy knee. After three months the cartilage in the damaged knee has a lower PD than the cartilage in the healthy one and R1 is the same for the cartilage in both knees. However, the R2 of the cartilage in the damaged knee is still lower than the cartilage in the healthy knee, but the gap is smaller than it was three months prior. Thus the only difference between the cartilage in the healthy and damaged knees that is consistent for both patients, is that R2 is lower in the damaged knees than it is in the healthy ones.

In order to know if the values presented in figure 4.4, 4.5 and 4.6 are close to the actual values, the average PD, R1 and R2 are also calculated using the created ground truth. These values are presented in figure 4.7, 4.8 and 4.9.

It is observed that the values presented in figure 4.7, 4.8 and 4.9 are different compared to the ones presented in figure 4.4, 4.5 and 4.6. Due to this, some of the previously observed property differences between the cartilage in healthy and damaged knees are the opposite of the actual differences. According to the property values that are calculated using the ground truth, the cartilage in the damaged knee of patient 1 has a slightly higher PD than the cartilage in the healthy knee. Also, after three months the cartilage in the damaged knee of patient 2 has a higher R1 and R2 than the cartilage in the healthy knee.
4.2 Comparison of cartilage in healthy and damaged knees

**Figure 4.4:** Diagram presenting the average PD in the knees of the two patients. The values within the parentheses are the standard deviations. All of the presented values are calculated using the best segmentation results.

**Figure 4.5:** Diagram presenting the average R1 in the knees of the two patients. The values within the parentheses are the standard deviations. All of the presented values are calculated using the best segmentation results.
Figure 4.6: Diagram presenting the average $R_2$ in the knees of the two patients. The values within the parentheses are the standard deviations. All of the presented values are calculated using the best segmentation results.

Figure 4.7: Diagram presenting the average PD in the knees of the two patients. The values within the parentheses are the standard deviations. All of the presented values are calculated using the created ground truth.
4.2 Comparison of cartilage in healthy and damaged knees

Figure 4.8: Diagram presenting the average $R_1$ in the knees of the two patients. The values within the parentheses are the standard deviations. All of the presented values are calculated using the created ground truth.

Figure 4.9: Diagram presenting the average $R_2$ in the knees of the two patients. The values within the parentheses are the standard deviations. All of the presented values are calculated using the created ground truth.
In this chapter the results and the method of the work are discussed.

5.1 Results

This section discusses the results acquired from using the two methods and the different SVM classifiers, as well as the comparison between the cartilage in healthy and damaged knees. When comparing the results of the different methods and SVM classifiers it is mostly done by comparing the MCC values, since it is a single value that describes the overall performance.

5.1.1 Segmentation methods and classifiers

When comparing the classification performance of the two segmentation methods it becomes apparent that method 1 is better than method 2, since all classifiers that use method 1 outperform their method 2 counterpart. However, when comparing the classification time of the methods, it becomes apparent that the classification time of method 2 is significantly shorter than that of method 1. The difference in classification performance is probably explained by the fact that when grouping voxels into regions, more voxels become prone to misclassification since the regions can contain voxels of both classes. The reduced classification time of method 2 is also explained by the fact that regions are used. By letting the data points represent regions instead of voxels, the total number of data points that are to be classified are reduced, which in turn reduces the classification time.

The impact that the different amounts of training data have on the classification performance is seemingly small. The classification performance becomes marginally better for most of the classifiers when 100% of the training data is
used instead of 50%, but the classification times are heavily increased. The results therefore suggest that using 50% of the training data could be sufficient, but it depends on the variability of the used data.

When comparing the results of the eight initial classifiers, the importance of the features from the different qMRI maps becomes apparent. The classifiers that do not use the features related to $R_2$ have the worst MCC of all classifiers. This indicates that the $R_2$ features are good features. The second to worst MCC is achieved by the classifiers that do not use the features related to $PD$. This indicates that the $PD$ features are good as well, although not as good as the $R_2$ features. On the other hand, it is noted that the classifiers that do not use the features related to $R_1$ achieve a similar MCC to the classifiers that use all features, but the classification time is reduced significantly. The impact that the features related to $R_1$ have on the results suggests that they mainly add to the classifiers complexity. Therefore the features related to $R_1$ might not be suitable to use in the classifiers. These observations of the $PD$, $R_1$ and $R_2$ features’ importance are in accordance with how the features rank in table 3.1. According to table 3.1, the $R_2$ features are ranked high, followed by some of the $PD$ features and the $R_1$ features are ranked low.

The classifier that was built based on the results of the initial eight classifiers, with the purpose of reducing the classification time, did not perform as expected. The expectation was that the classification time would be shorter and the results only slightly worse than those of the classifier that was trained on 50% of the training data and used no $R_1$ features. As it turned out, the classification time became longer and the results worse than the expected outcome. This outcome indicates that some of the features that were assumed to be insignificant should in fact not be removed, since they are obviously useful. Which of the removed features that actually are useful was not investigated further due to time constraints.

### 5.1.2 Comparison of cartilage in healthy and damaged knees

Since the values presented in figure 4.4, 4.5 and 4.6 are quite different from the ones that are presented in figure 4.7, 4.8 and 4.9, the values that are calculated using the segmentation results are not suitable to use to draw any conclusions. The reason why there are differences is because even though the method and classifier used to calculate the values was the combination that achieved the best MCC, the evaluation shows that its precision and recall are fairly low. The precision and recall are 0.7425 and 0.7180 respectively, which means that approximately a quarter of the voxels used to calculate the properties were not cartilage, and that more than a quarter of all cartilage voxels were not included in the calculations. This of course affects the outcome when calculating the average of each property, causing the presented property values to differ from the actual values.

When looking at the differences between the properties of cartilage in healthy and damaged knees that are calculated using the ground truth, the differences are quite inconsistent between the two patients. However, if the rescan of the second patient’s damaged knee is ignored, then the cartilage of both patients has a lower $R_1$ and $R_2$ value, and a higher $PD$ value in the damaged knee than in
the healthy one. A decrease of $R_2$ (i.e. an increase of $T_2$) is actually commonly associated with cartilage damage according to [15], but the comparison also indicates a possible relation between $PD$ and cartilage damage, as well as $R_1$ and cartilage damage. On the other hand, if the rescan of the second patient is taken into account, these possible relations become less likely since the cartilage in the damaged knee then has a higher $R_1$ and a lower $PD$ than in the healthy knee.

Using the calculated property values when trying to relate the property differences of the cartilage in healthy and damaged knees to clinical outcome is very problematic. Not only is it problematic because there is only results from two patients, which makes it difficult to draw any general conclusions, but because of two other reasons. The first reason is that there is currently no information available regarding if the patients in the clinical project have actually developed any cartilage damage in their damaged knees yet. This makes it impossible to relate any property differences to cartilage damage since there is a possibility that any differences could be natural variations. The second reason is that there is a possibility that some of the voxels that are annotated as cartilage might suffer from partial volume effects. Partial volume effect means that a voxel contains several kinds of tissues, which causes the voxel’s signal strength to be an average of the contained tissues. If a lot of cartilage voxels suffer from this, it can of course generate misleading results. Partial volume effect could possibly explain the fairly large standard deviations of the calculated property values. However, the effect that such voxels might have had on the property values that are calculated using the segmentation results are probably much smaller than the effect of the imprecise segmentation.

5.2 Method

The parts of the method that are identified as areas where other approaches or improvements could be made are the choice of segmentation method, the annotation of data, the femur segmentation and the features used by the classifier.

As previously mentioned in section 2.3, the choice of using a supervised machine learning method to perform the segmentation was made after having read literature related to the subject. Not only had it been successful when using conventional MRI data, but it was also thought to be an interesting opportunity to use features related to the different qMRI maps. No other automatic segmentation methods were tried out in this work because of time constraints. This means that using a SVM classifier to solve the task might not be the optimal choice. More methods should be implemented and evaluated so that it becomes possible to analyze how using SVM classification fares compared to other methods.

The method used to create the annotated data should have been approached in another way. The implemented method requires input parameters and a manually placed ROI. This causes the results to vary depending on the user who is performing the annotation, but this is a problem with manual annotations as well. The produced annotations in this work were deemed to be a reasonable representation of the cartilage by a radiologist, although he did say that they were not
entirely correct. The incorrect annotations have probably had an impact on the results, since they were used both when training the classifiers and evaluating the results. The preferred approach would have been to use manually annotated data, created by someone with proper knowledge of the anatomy. This is in fact going to be produced for the used data sets, but unfortunately it did not happen within the time frame of this work. It would have been interesting to see how such annotations could have affected the results.

The femur segmentation is a part that suffers from some problems. One problem is that it could only be evaluated visually since there was no corresponding ground truth available. Therefore it is impossible to objectively tell if it performed well or not. Upon visual inspection the method produced good results on most of the used data, but on some slices the femur was not detected at all, or suffered from either over- or undersegmentation. Another problem was that not the whole femur could be segmented out due to a lack of prominent edges around it in the qMRI data. The lack of prominent edges was also the reason why neither of the patella or tibia could be segmented out at all. If it had been possible to segment out all bones, a more specific ROI that describes where the cartilage is present could have been created. Such a ROI might also have made it possible to identify to which of the bones that the cartilage belongs to. The second problem suggests that other bone segmentation methods should be tested to see if it is possible to generate better results.

Regarding the classification, it would have been interesting to further investigate which of the features used that were the most significant and how the amount of training data affected the results. In this work only a few permutations of features and used amount of training data were tried out due to time constraints. As a result of this, only general conclusions could be drawn, such as that the features related to $R^2$ are good features. Two other things that would have been interesting to do would be to improve the current features, as well as to implement new ones.
Conclusions and future work

This chapter presents the findings of this work and answers the questions in section 1.1. It also gives suggestions on what could be done if the segmentation methods are to be worked further upon.

6.1 Conclusions

The two segmentation methods implemented in the work of this thesis are able to perform automatic segmentation of knee cartilage using qMRI data with varying results. Method 1 and 2 were able to achieve a MCC of 0.7211 and 0.6969, respectively. Regarding the question of how an automatic segmentation algorithm for segmentation of knee cartilage should be implemented when using qMRI data, these results suggest that using a SVM classifier to perform the task can be a good approach. However, both methods have to be further improved and tested before clinical use.

When evaluating how the segmentation methods performed compared to the ground truth, which was one of the questions, the evaluation shows that method 1 achieves a better precision, recall, specificity and MCC than method 2. This was expected since method 1 classifies each voxel individually. The advantage of using method 2 over method 1, is that method 2 has a considerably shorter classification time. For the best results acquired with each method, method 2 classified the data sets in 22% of the time it took with method 1. Therefore, the choice of using either method 1 or 2 can be considered a trade-off, where method 1 generates better results, but method 2 is a much faster alternative.

If a short classification time is essential, the results in section 4.1 show that there is an efficient way of lowering it for both segmentation methods. By not using the features related to $R_1$, and by training the classifier using only 50% of the available training data points from each class and slice, the classification time
is reduced. Doing this only generates slightly worse results than if the classifiers used all features and were trained using 100% of the training data.

Regarding the question of how the properties of the cartilage can be related to clinical outcome, unfortunately no good answer can be given based on the performed comparison. The lack of information regarding whether or not the patients have developed cartilage damage, in combination with the segmentation results being quite imprecise, makes it impossible to draw any conclusions. Nevertheless, if the segmentation results are improved, the investigation showed that using the segmentation results in combination with the qMRI maps can be an efficient way of studying how the cartilage properties change over time or vary between knees.

6.2 Future Work

The implemented methods should be trained and evaluated using data sets that have manually annotated ground truth available. The annotations should be created by someone with the proper knowledge of the anatomy. This would generate results that are more representative of how the two methods actually perform.

The femur segmentation should be improved so that it yields better results. This could involve trying other methods than the one implemented in this work. The segmentation should also be extended so that it is able to segment out all of the bones in the knee and not just the femur. This would give rise to the possibility of using more features related to landmarks and creating a better ROI, both of which might improve the cartilage segmentation results.

Many of the features used could be modified so that they are calculated using three dimensions instead of two. Future work should also include a feature that describes the voxels’ location along the z-axis in relation to some landmark and see how it affects the results.
Bibliography


