MACHINE LEARNING APPLIED TO THE STUDY OF KNEE OSTEOARTHRITIS AND ASSOCIATED PAIN

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2019

By
Luca Minciullo
School of Computer Science

Contents

Word Count: 25421

List of Tables

List of Figures

Abstract

Background: Knee osteoarthritis is one of the leading causes of disability worldwide, affecting 3.8 million people around the globe. Despite its prevalence it is still a poorly understood condition for which limited treatments are currently available. Osteoarthritis(OA) usually starts by affecting the articular cartilage covering the surface of bones in the knee joint, ending up involving all tissue types together with the synovial fluid. The way clinicians diagnose this disease is by looking at radiographic images and assigning a severity score from 0 to 4, where 0 stands for a healthy knee and 4 is late stage OA. Current methodologies in automated study and diagnosis of OA have been applied to small datasets made of a few hundred images and tend to involve only the analysis of Posterior-Anterior(PA) view radiographs. In addition, the relationship between structural changes in the joint and symptoms has not been well understood.

Aim: The aim of this work was to improve the performance of computer aided diagnosis techniques available when studying knee OA from medical images. These techniques have the chance of helping the experience of people affected by this very common disease by supporting and speeding up the diagnosis so that appropriate counter measures can be taken. There are two main improvements that we propose: first, the incorporation of additional informative data and second the refinement of the machine learning model. Furthermore, we wanted to contribute to the understanding of the relation between what can be seen in medical images and the symptoms that people experience when affected by OA. This has the potential to deepen our understanding of osteoarthritic sources of pain and ultimately can affect the direction of focus in clinical trials.

Methods: Random forest based landmark point detectors have been built to find the outlines of the bones in knee joint radiographs. Separate models were built for lateral and PA view images. The found annotations allowed the automatic extraction of measurements associated with the shape, texture and appearance of the bones and their spatial relation within radiographic images. We used these features to perform several OA related classification tasks, including automated diagnosis of structural changes. We proposed an improvement over the classification model previously used with the introduction of what we called “Indecisive Forests” together with ways of optimising

such forests once they have already been trained. Finally, a comprehensive exploration on radiographic sources of pain and investigation on whether it is possible to find a clearer relation between images and symptoms was performed.

Results: Our lateral knee model was able to perform as well as the PA model in the first experiments and showed high discriminative ability considering that it is not used by clinicians to perform the grading. The combination of features from the two views only marginally improved performance, with the full knee model using appearance features achieving the best overall results. Using the indecisive forest further reduced the number of classification errors on two classification tasks, while the results of the experiments on the proposed optimisation routine did not allow us to conclude on its effectiveness. The radiographic structural changes that can be seen as a source of pain were a combination of lateral and PA manual features. Consistent knee pain showed an improved correlation with manual scores compared to what has been reported in the literature.

Conclusions: The results of our work suggest that features extracted from the lateral view are informative and that using multiple views in general helps performance, though not always by a large margin. Predicting future knee pain is the hardest task for the automated models we used. Our indecisive forest based experiments achieved the state of the art on the tasks of interest, though at the expense of a higher computational costs. We presented the highest correlation between radiographic features and frequent knee pain when evaluated with AUC.

Declaration

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

Copyright

Figure 1: Neural Networks painting works following the style of famous artists. (image downloaded from [2])

> “Patience is something you admire in the driver behind you, but not in the one ahead. ”
>
> — Anonymous

Abbreviations

AAM Active Appearance Model
AAN Artificial Neural Network
AI Artificial Intelligence
ASM Active Shape Model
AUC Area Under the Curve
BMI Body Mass Index
CAD Computer Aided Diagnosis
CAM Combined Appearance Models
CLM Constrained Local Models
CNN Convolutional Neural Network
CV Cross Validation
D-T Discriminative Training
EEG Electro Encephalo Gram
FCN Fully Convolutional Network
fMRI functional MRI
FS Fractal Signature
GANs Generative Adversarial Networks
HOG Histograms of Oriented Gradients
HOT Hurst Orientation Transform
INRIA Institute for Research in Computer Science and Automation
JSN Joint Space Narrowing
JSW Joint Space Width
KL Kellgren & Lawrence
MOAKS MRI Osteoarthritis Knee Score
MOST Multicenter Osteoarthritis Study
MRI Magnetic Resonance Imaging
OA Osteoarthritis
OAI Osteoarthritis Initiative
OARSI Osteoarthritis Research Society International
PA Posterior-Anterior
PCA Principal Component Analysis
RA Rheumatoid Arthritis
ReLU Rectified Linear Unit
RF Random Forest

RFCLM Random Forest Constrained Local Model

ROC Received Characteristic Curve

ROI Region Of Interest

SI Similarity Index

SAQ Self Assessed Questionnaire

SOM Self Organising Maps

SVM Support Vector Machine

TScreen Telephonic Screening Interview

VOT Variance Orientation Transform

WOMAC Western Ontario & McMaster Universities Osteoarthritis Index

WS WideSpread

YLDs Years Lived with Disability

Acknowledgements

I would like to especially thank my supervisor Tim Cootes for the continuous and tailored support he has provided me through the last years. His attitude towards our work has been essential. Secondly, I would like to thank David Felson, who has been a constant reference for me and a source of valuable feedback and energy. I would also like to thank my co-authors Jessie Thomson, Matthew Parkes and Paul Bromiley, with whom it was a real pleasure to collaborate.

Further, I would like to thank the following:

- My officemates Raja Ebsim and Luke Chaplin, my colleagues and all the people working in the division, whose social and academic support made my days much better.
- Michael Lee, for his generosity in helping me proofread this work.
- My family, for adapting to the distance though effort and unconditional support.
- The University of Manchester, for this life opportunity and for all the tools and resources provided to me.
- My funding body, the Engineering and Physical Sciences Research Council, without which none of this would have been possible.
- All my friends in Manchester and all the people that I shared moments with.
- Toyota Motor Europe for offering me a summer internship, during which I better understood what I wanted.
- 婧然, for being a constant source of inspiration.

A special mention goes to Manchester itself, whose vibrant and multicultural life showed me a new world and helped me understand who I was and who I wanted to become.

One final thanks goes to all the people that I forgot to thank. You deserved better.

About the Author

The author, Luca Minciullo, completed a BSc in Mathematics at "La Sapienza" University of Rome in 2012, and an MSc in Applied Mathematics at the "La Sapienza" University of Rome in 2014. The PhD was started in 2014 as part of the School of Computer Science CDT programme, during which the author has produced the following publications relevant to the project:

- Minciullo Luca and Timothy F. Cootes "OA Classification Using Lateral Knee Radiographs." International Workshop on Osteoarthritis Imaging 2016.
- Minciullo Luca, and Timothy F. Cootes. "Fully automated shape analysis for detection of osteoarthritis from lateral knee radiographs." Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016.
- Minciullo Luca, Jessie Thomson, and Timothy F. Cootes. "Combination of lateral and PA view radiographs to study development of knee OA and associated pain." Medical Imaging 2017: Computer-Aided Diagnosis. Vol. 10134. International Society for Optics and Photonics, 2017.
- Minciullo Luca, Paul A. Bromiley. Davod T. Felson, Timothy F. Cootes "Indecisive Trees for Classification and Prediction of Knee Osteoarthritis." International Workshop on Machine Learning in Medical Imaging. Springer, Cham, 2017.
- Minciullo Luca, Matthew Parkes, David T. Felson, and Timothy F. Cootes. "Comparing Image Analysis Approaches vs Expert Readers: the Relation of Knee Radiograph Features to Knee Pain", 2018 (Submitted to Annals of Rheumatic Diseases)

Alternative Format

This thesis has been written following the "Alternative" or "Journal" format. At The University of Manchester a PhD Thesis allows the author to present the outcome contributions of the project as list of papers. All papers presented in this work have been either published already or under review. Furthermore, our submissions have consistently been peer reviewed by international conferences or journals. This format was chosen as given the nature of our work and the flow that our papers follow, the story telling would have not been affected. Furthermore, choosing this format enabled the author to spend significantly more time on research, leading to his final journal publication.

This thesis reports four papers in total.Two of which are presented in chapter 4. One paper is discussed in chapter 5 and a final journal paper can be found in chapter 6. The final paper is presented in an expanded version compared to the one submitted in order to give more details and help its readability.

Chapter 1

Thesis Overview

1.1 Introduction

Knee Osteoarthritis (OA) is a disease whose main symptoms are stiffness and pain. This inflammatory disease affects all tissues in the joint, starting with the erosion of articular cartilage which cushions the joint, followed by the onset of bony spurs called Osteophytes. The joint capsule is also affected by these changes, making the synovial fluid which helps reduce friction in the joint less concentrated and thus causing more attrition and consequently pain. Knee OA is usually an asymmetrical disease, affecting mostly one side of the joint. That causes joint space narrowing and malalignment.

Knee OA is the most common form of arthritis and one of the leading causes of disability globally, affecting 3.8% of the global population [24]. Its relative prevalence is shown in Figure 1.1. Among musculoskeletal diseases and pathologies, knee OA is the third most prevalent, only preceded by low back pain and neck pain. When compared to diseases in general, OA was also found to be one of the main causes of disability, ranked 11th overall [24], immediately after diabetes. Disability can be quantified using the YLDs (Years Lived with Disability) measure. Figure 1.2 shows the ranking of the most prevalent musculoskeletal diseases according to this factor.

Despite the availability of more modern imaging modalities such as CT and MRI, radiographic assessment is still the gold standard in Osteoarthritis imaging, mainly due to the acquisition costs and speed. There are a number of grading schemes to quantify the severity of structural OA, most of which are based on posterior-anterior radiographic images. The most used grading scheme is semi-quantitative and was designed by Kellgren and Lawrence in 1957 [59].

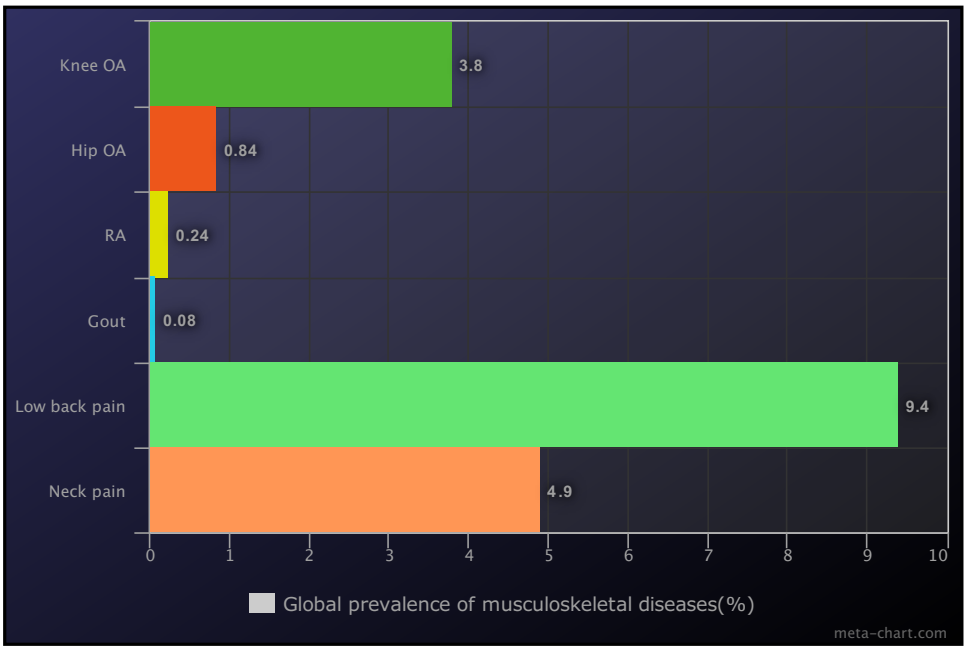


Figure 1.1: Knee Osteoarthritis is the third most prevalent musculoskeletal disease found in the global burden of disease 2010 study [24, 23, 52, 53, 99].

One of the main issues with manual grading of radiographic images is the subjectivity of the assessment. Several studies [118, 46] have shown high inter- and intra-rater variability, partially due to the arguably ambiguous definition of the individual grades. This can lead to significant inefficiency when running clinical trials. Most often, in order to minimise the effect of subjectivity, more clinicians are needed to assign grades to images and more participants need to be recruited. This increases both costs and the time spent running clinical trials.

In medical imaging we have seen an increasing interest in Computer Aided Diagnosis (CAD) systems. These systems automate the study of medical images, performing the same tasks that trained clinicians would do. The obvious advantage is in the repeatability of such measurements and the time efficiency (it takes seconds to have a response from a machine vs. days or weeks for a clinician). Improving the reliability and accuracy of such methodologies is crucial to optimise clinical trials and provide quick and cheap evaluation of radiographic images. Early diagnosis of OA can help slow down disease development and reduce symptoms by suggesting changes to people's lifestyles and adopting other precautions.

Current methodologies in automated study and diagnosis of OA [108, 109, 97, 6] have been applied to small datasets made of a few hundred images and tend to involve

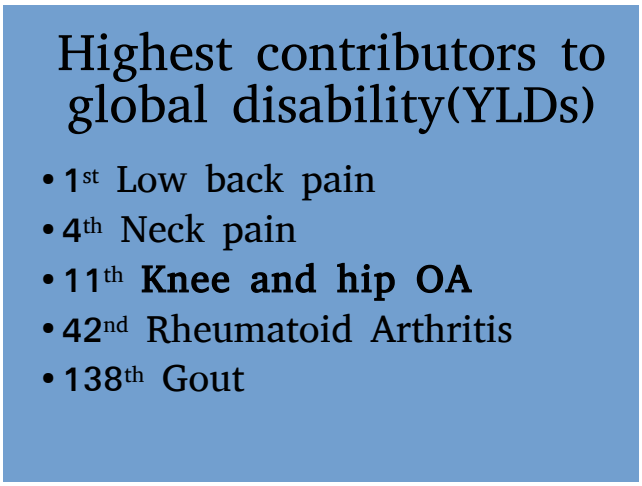


Figure 1.2: Among musculoskeletal diseases knee and hip OA are the third biggest global contributor of disability. This becomes the 11th if we consider all diseases regardless of their type.

only the analysis of the PA view, while other views are often available and as easy to acquire. Furthermore, little evidence is available in terms of what images contain specific disease related information. Finally, the relation between radiographic OA and symptomatic OA has always been hard to grasp. The way to reduce disease symptoms is to first understand what is causing them.

In this project we created a model to analyse lateral view radiographic images. A segmentation model allowed us to localise landmark points associated to the bone shape. Different sets of features were extracted using these coordinates and their discriminative ability at various OA related classification tasks was evaluated. The comparison and combination with features associated to different radiographic views and physiological and demographic data were also performed. We extended the work to the largest datasets available in the field, increasing by two orders of magnitude the number of subjects. In addition, we designed a novel machine learning model for classification to reduce the number of mistakes made by the system. Furthermore, we explored in more depth the relation between symptoms and radiographic features, by taking into account pain measures acquired at different time points with respect to radiographic acquisition.

1.2 Aims and Objectives

The aim of this work was to improve the performance of computer aided diagnosis techniques available when studying knee OA from medical images. These techniques have the chance of helping the experience of people affected by this very common disease, by supporting and speeding up the diagnosis so that appropriate counter measures can be taken. There are two main improvements: first by incorporating additional informative data and second by refining the machine learning model. In addition, we wanted to give supplementary evidence on some of the commonly believed concepts in clinical practice. Given the nature of our work the kind of evidence we can provide is both quantitative and qualitative and this can support or challenge conjectures in other domains. Furthermore, we wanted to contribute to the understanding of the relation between what can be seen in medical images and the symptoms that people experience when affected by OA. This has the potential to deepen our understanding of osteoarthritic sources of pain and ultimately can affect the direction of focus in clinical trials.

1.3 Contributions

The main contributions reported in this thesis are as follow:

- The development of a **fully automated landmark point detector for lateral knee radiographs**
- The extraction of radiographic features from those images and their use to **automatically diagnose the presence of structural OA**
- The comparison between the two main radiographic views from a machine learning stand point and the exploration on the benefits of **combining radiographic features from multiple views** to solve three OA related classification tasks
- The implementation and evaluation of a **novel machine learning classifier called "Indecisive Tree"**. The exploration of optimisation methods of an already trained indecisive forest based on back-propagation.
- The exploration of the **relationship between manually graded radiographic features and symptoms** experienced by the participants of a OA study. The

development of a model based on automatically extracted features capable of performing as well as manual grades.

- In general, our work was the first one as far as we are aware to use large datasets (around 20k images) in imaging for OA.

1.4 Outline of the Thesis

Chapter 2 provides an overview on knee osteoarthritis, its features and risk factors together with the available treatments.

Chapter 3 presents a literature review on machine learning techniques in computer vision and a more detailed description of the main approaches to study knee radiograph when trying to determine the presence of OA.

Chapter 4 is the first chapter containing experimental results. We introduce the novel lateral knee landmark point detector and show its effectiveness to extract structural features to discriminate OA affected knees from healthy ones. In addition, we present experiments on combining features from both Lateral and PA radiographs. The classification tasks involved prediction of onset of both structural and symptomatic OA.

Chapter 5 introduces a novel classification method called "Indecisive Tree". This techniques generalises the behavior of a classic random forest by adding an indecisive region to the space where optimal binary splits are made. Pros and cons of using this technique are presented.

Chapter 6 describes the work on the relation between radiographic features and frequent knee pain. We compare manual and automated assessments looking at multiple pain scores acquired at different time points and calculated from participants' responses.

Chapter 7 sums up the contributions of this work and what can be concluded from our findings. We propose possible natural directions for future investigation.

Chapter 2

Knee Osteoarthritis

This chapter includes an overview of knee osteoarthritis, describing its prevalence and how it affects the daily life of patients. The review provides an introduction to principles of physiology of the knee joint, a detailed description of the main characteristics of Osteoarthritis, its development in time, the factors that seem to increase the chance of future onset of the disease and the treatments that are currently employed to reduce the symptoms and increase mobility.

2.1 Structure of the knee joint

An illustration of knee anatomy can be found in Figure 2.1. In what follows we will use the standard notation of *medial* side, meaning the one towards the centre of the body and *lateral* side, facing away from the centre of the body. The structure of the knee joint can be divided into several components. The first component is bone and it is made of four different elements: proximal tibia, distal femur, fibula and patella. The second component is cartilage, made of meniscus (medial and lateral, both located between the tibia and the femur is an example of fibrocartilage) and articular cartilage (tibial, femoral and patellar, covering the surface of each bone is the main example of hyaline cartilage). The third component is the ligaments 2.2 (the cruciate and collateral ligaments) that keep the other components of the joint together. The last component is the synovial membrane (also called joint capsule) containing the synovial fluid, a non-Newtonian fluid responsible for reducing attrition during movement.

The knee joint is the largest joint in the human body. It consists of what can be regarded as two joints, the joint of the tibia and the femur (tibiofemoral joint) and the one comprising patella and femur (patellofemoral joint). The tibiofemoral joint is the

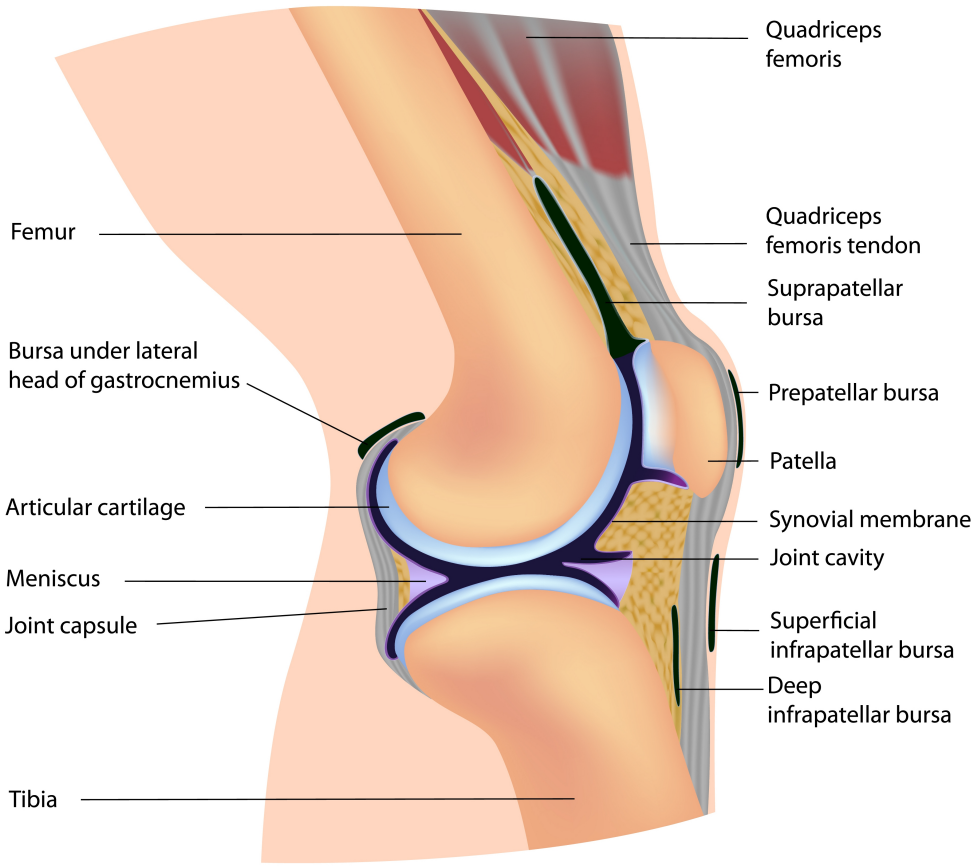


Figure 2.1: An illustration of the anatomy of the knee joint in its different components: bone, articular cartilage and meniscus [115].

weight-bearing joint whose movement enables the leg to bend. Its stability is assisted by the presence of cruciate and collateral ligaments. The patella is attached to the quadriceps femoral muscles at the front of the tibiofemoral joint via the quadriceps tendon.

There are two types of bone tissue in the human body: cortical bone and trabecular bone. Cortical bone is the outer layer of any bone and makes around 80% of bone mass. Its functions are organ protection, stability of structure and storage of calcium. Trabecular bone is the inner layer of bone and can be mostly found at the extremities of long bones such as the femur and tibia. It is a softer kind of tissue compared to cortical bone and its composition makes it flexible to adapt to different load distributions.

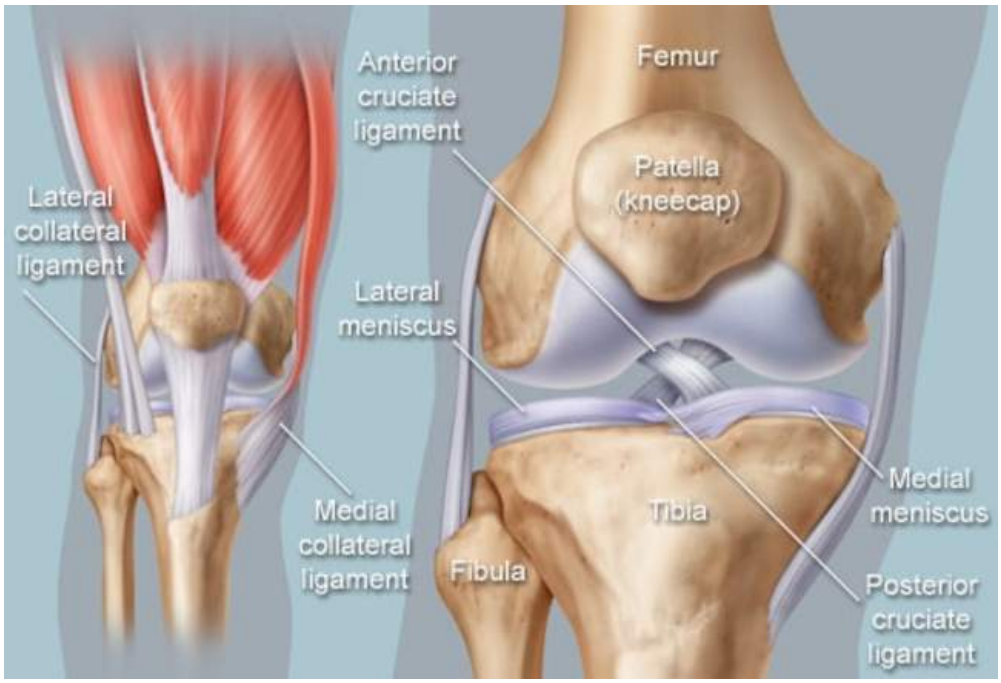


Figure 2.2: An illustration of the anatomy of the knee joint with particular emphasis on ligaments. [51]

2.2 The Disease

Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world. The disease is associated with pain, stiffness, and loss of function. It has been reported [36] that by 2030 around 20% of the American population will be above the age of 65, and that half of them (35 million patients) will be at high risk of developing OA, hence requiring huge amounts of public money for treatments and surgery [15].

Despite its prevalence and severity, OA still remains poorly understood and a condition for which there are limited effective treatments available [54], most of them just reducing the severity of the symptoms. Researchers are as yet unsure what initiates Osteoarthritis or in which tissues the pathology originates [9]. Originally, it was attributed to a deterioration of the cartilage, although it is now known to be a disease that affects all joint tissues, causing both degeneration and malformed restoration of the joint.

It can be challenging to separate the structural changes due to aging from pathological changes as a result of OA [17]. Up to a few decades ago, OA was regarded only as disease related to aging, but now it has been established that the disease can

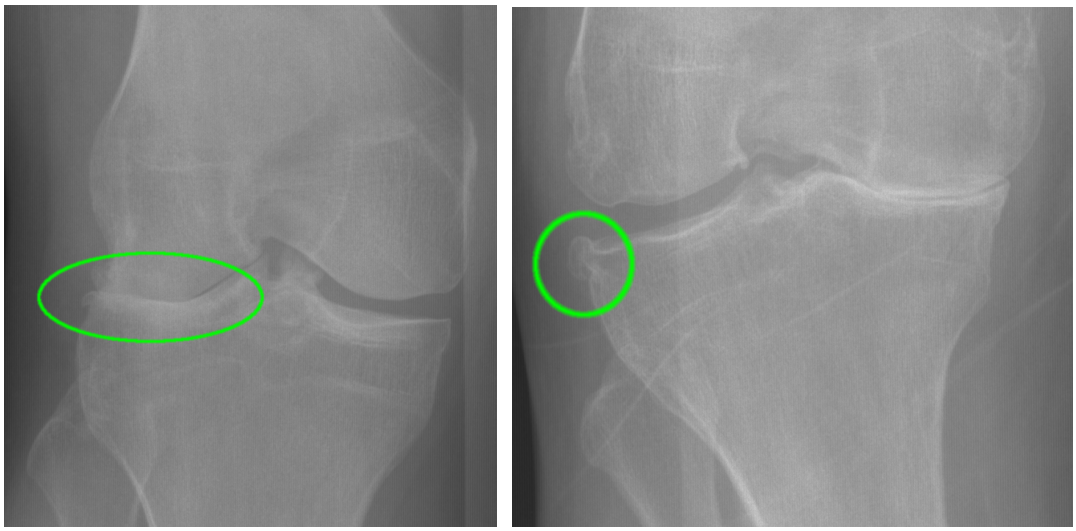


Figure 2.3: Examples of severe JSN and tibial osteopyte.

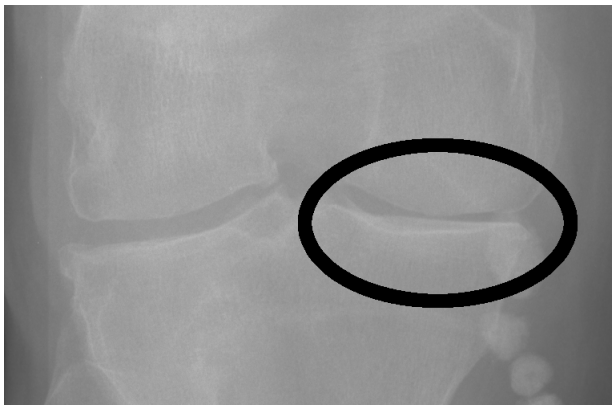


Figure 2.4: Example of a OARSI grade 2 sclerosis.

be developed at any age. It is often considered to be a condition with multiple causes, making it harder to define and to establish a relationship between cause and effect.

There are a number of structural features of the knee joint that are related to OA. The most relevant are *joint space narrowing* (JSN), which implies that the distance between a pair of bones is not symmetrical in the medial and lateral side and is therefore reduced, causing increased attrition (Figure 2.3, left). The commonest form of JSN appears in the medial tibiofemoral area. A second structural sign of OA is the development of *Osteophytes* (Figure 2.3, right), small bony spurs that form around the joint. Other common features are calcium deposits, sclerosis (Figure 2.4) and cysts, all of which are linked to mechanical stress.

Osteoarthritis usually develops in the following way[1]: first, articular cartilage of both the femur and the tibia starts to break down, losing smoothness and causing discomfort in movement. Osteophytes begin to develop on the edge of the joint. Osteophytes are considered to be a natural response of our body to the loss of cartilage. Second, the articular cartilage begins to erode causing narrowing between the joints and increased pain in flexion. Hyaluronic acid in the joint capsule loses density reducing its lubricative function together with the synovial fluid. At this stage subchondral bone, located right below the articular cartilage has possibly been affected as well. This type of bone is responsible for transferring oxygen and hydration to cartilage and the attempt of the body to repair releases proteins and cytokines in the synovial fluid, further lowering its density and efficacy, giving pain even in rest position. Third, osteophytes increase in number and size and the cartilage deteriorates up to the point where the bones touch each other, causing severe pain and impacting movement and lifestyle.

There are two recognized types of Osteoarthritis, depending on the cause: primary Osteoarthritis and secondary Osteoarthritis. Primary Osteoarthritis, also called "wear and tear", is the most commonly diagnosed Osteoarthritis. It is mainly associated with aging and with an excessive use of the joints. The increasing life expectancy worldwide is likely to cause a larger portion of the population to suffer from this.

Secondary Osteoarthritis has exactly the same symptoms but it can develop at any point in the life of an individual. The possible causes for secondary Osteoarthritis are reported as risk factors in the following section.

2.2.1 Risk factors

There are a number environmental and mechanical factors that are linked with an increased risk of developing OA.

Some of the more prominent risk factors are: *injuries* [33, 88], history of previous fractures or operation of the joint has been proven to increase the chance of developing the disease; *obesity* [34, 80], weight-bearing joints are at increased stress levels and this causes Osteoarthritis to occur and develop more rapidly; *hormonal factors* [86, 74], hormones affect the way and the speed at which cartilage regeneration and remodeling cycles happen; *collagen deficiency* [5], is a hereditary factor that causes

[1]While it is true that every individual experiences a different disease progression, the following serves as a description of causality in Osteoarthritis and gives both physiological and mechanical insights.

increased risk of OA due to collagen being a critical component in the bone and cartilage formation; *inactivity* [84], a sedentary lifestyle leads to weight gain, which can lead to osteoarthritis. Also, if you are inactive, you have weaker muscles and tendons that surround the joint. Strong muscles help keep joints properly aligned and stable; *Genetics* [101], having ancestors who had OA has shown to increase the chance to develop it; *inflammation from other diseases* [103], diseases that cause inflammation, such as rheumatoid arthritis, can increase your risk of onset of OA.

Other risk factors are currently being studied, but some of them are already widely accepted by the research community. It is worth mentioning in this context the relation between practicing sports at a professional level and the development of the disease [94].

2.3 Available treatments

As mentioned earlier there is no cure for Osteoarthritis but a number of treatments are available to reduce the symptoms, potentially slow down disease progression and therefore limit the impact of the disease on patients' lifestyles. The most prominent treatments are: *exercise*, taking care of fitness is a crucial aspect when aiming for healthy joints. There is evidence that having strong quadriceps can protect against knee pain and slow down disease progression [95]. On the other hand, people who have OA might tend to reduce their physical activities causing their quadriceps to weaken. This is known to increase instability of the joint and accelerate erosion of cartilage. Another treatment is adjusting *diet* in order to reduce the BMI. People who are overweight tend to stress more the weight-bearing joints like the knee and this, combined with OA, can deteriorate tissues faster.

Injection of hyaluronic acid, *Analgesics* and the *Injection of steroids* are three common ways to relieve pain symptoms caused by OA. Hyaluronic acid is naturally present in the knee capsule to reduce friction. Nowadays, it is often extracted from rooster combs. Analgesics are all-purpose pain killers but their effect is quite limited in time. Steroids have a much longer effect compared to analgesics. Their consumption is only advised in presence of severe pain. *Knee replacement* is the last resort and it is usually done when the last stage of OA is diagnosed. This procedure is highly invasive and expensive [89] and having an artificial knee does not solve the problem since surgery is required again every ten years [28].

2.4 Grading systems

Similarly to any other disease a branch of research is performed to explore reliable and repeatable ways of assigning severity scores to the stage of OA. This can be done by looking at different sorts of data: medical imaging modalities, physiological data, mechanical data etc. The most widely used grading systems of structural OA are based on the measurement of some of the osteoarthritic features described before from plain radiographic images. That is mainly due to the fast and cheap acquisition of radiographic images. Current methods for establishing the presence and measuring the severity of OA from radiographs are split into two groups: *quantitative*, where the grading makes use of specific measurements of the osteoarthritic features; and *semi-quantitative*, where the assessor has to compare X-rays against some typical reference representations of the different grades. For the semi-quantitative grading, the most prominently used methods are: Kellgren-Lawrence [59] grading and atlas grading methods, such as the Line Drawing Atlas [85] and Altman et al. [4] grading, also called OARSI atlas.

Atlas based grading systems are examples of *individual scoring systems*, where a score is assigned to each of a set of OA features. The sum of those scores will be the overall OA grade. The main drawback of using atlas methods is that it can be very time consuming, since in order to assign each of the individual scores a radiograph has to be compared with the reference images.

The KL grading system, as described in Table 2.1 is a composite grading method. Radiographs showing the progression of KL grades from 0 to 4 is shown in Figures 2.5, 2.6 and 2.7. That means that the grade is assigned by looking at a set of features simultaneously. Due to the composite nature of the grading, it can be difficult to distinguish what grade a knee should be assigned. Some OA affected knees will have different features developing at different rates, while this grading assumes parallel progression of the structural signs. It is possible to find a knee joint with extremely marked joint space narrowing but almost no signs of osteophytes. According to the KL grading prescriptions, the knee should be considered either grade 1 or 2, despite the lack of joint space indicating a more severe progression of the disease. An example of a quantitative grading method is the Ahlback grading system [3].

Grading systems are assessed by performing experiments on the inter- and intra-rater reliability, finding a measure of how often different graders agree on assigning severity to OA. One of the most relevant studies with this aim was executed by Gossec

Grade	Description
0	Normal
1	Doubtful narrowing of joint space and possible osteophytic lipping
2	Possible narrowing of joint space and definite osteophytes
3	Definite narrowing of the joint space, moderate multiple osteophytes, some sclerosis and possible deformity of bone ends
4	Marked narrowing of joint space, large osteophytes, severe sclerosis and definite deformity of bone ends

Table 2.1: K-L grading scheme

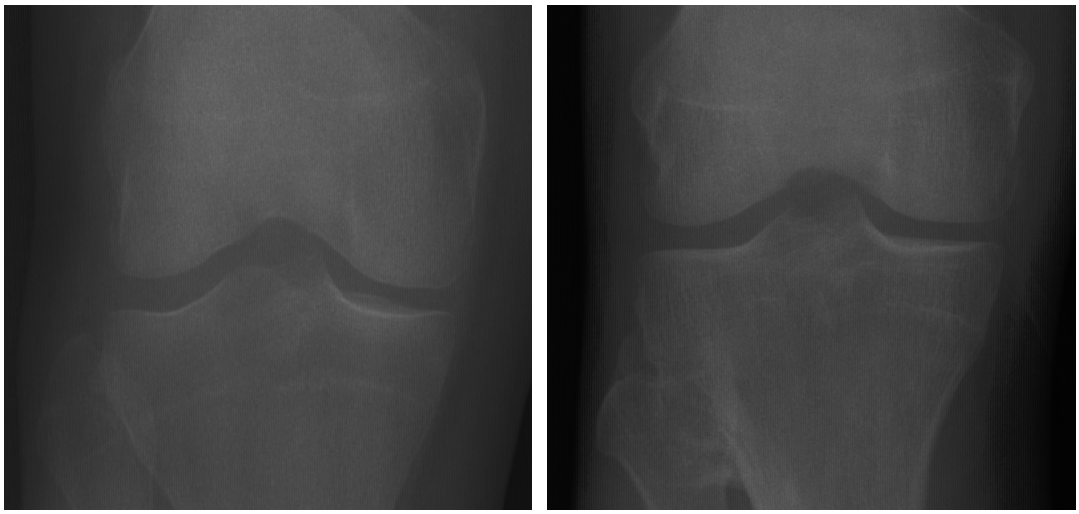


Figure 2.5: KL grade 0 (left) and KL grade 1 (right)

et al. [46], where 1759 x-ray images were graded using KL, OARSI joint space narrowing score and measurement of joint space width (JSW). Results on inter-rater reliability was higher for JSW with value of kappa equal to 0.86 compared to 0.56 and 0.48 of KL and OARSI, respectively. Furthermore, JSW had the highest intra-rater reliability (kappa=0.83) while KL and OARSI resulted with only 0.61 and 0.71. Even though KL grades have been frequently criticised for not being able to give a reliable measure of this complex disease they are still the gold standard.

2.5 The MOST initiative

MOST (Multicenter Osteoarthritis Study) [36] is a longitudinal, prospective study of Knee OA run in the United States and involves 3026 participants, men and women aged

Figure 2.6: KL grade 2 (left) and KL grade 3 (right)

Figure 2.7: KL grade 4

50 to 79 at baseline, with or at high risk of developing knee OA. Eligibility criteria included those who were overweight or obese, with knee pain or with a history of knee injury or operation. Each participant made one visit at baseline and then weight bearing knee radiographs (two lateral and one posterior anterior) and MRI were taken at months 15,30,60 and 85. The study collected a wide variety of data involving physical data such as BMI and height, history of previous injuries, symptoms and medications taken.

The assignment of grades was done in the following way. A panel of three experienced readers was chosen. Radiographs taken at each visit were assessed by a pair of readers and, in case of disagreement, the third reader was required to give an opinion as well. Both KL and OARSI grades were assigned.

Chapter 3

Machine Learning for Osteoarthritis

This chapter is an overview of the techniques used in the project together with the circumstances of their creation. This will include aspects of machine learning, statistical shape analysis and a literature review of various approaches to the automated study of Osteoarthritis, both by looking at medical images and other data sources.

3.1 Object detection

The literature on object detection is very large since locating objects in images is a key step in many applications, one of the main ones being object segmentation, the problem of finding the contours of an object of interest in images. In this section, we will try to give a brief overview of some of the most successful approaches.

The idea behind the earliest approaches for object detection was to take a fixed template for the object shape and attempt to detect similarities between this template and the image [91]. For instance, the template of a square would only need a rigid set of the four corners and connecting edges to be able to accurately detect other squares. This method works well if the object appears with approximately the same orientation as the template, otherwise it will often fail.

3.2 Machine learning algorithms

Significant improvements were obtained with the increasing success and development of machine learning techniques. Nowadays most state of the art techniques in the field of computer vision and its application are machine learning based [70, 49, 120]. We can think of data as a set of vectors, where the number of vectors represents the amount

of examples of data instances coming from a source. Each of those vectors contains measurements of characteristics (features) associated to the specific sample. For instance, if we are describing a person, those measurements can be physical characteristics such as height, weight, blood pressure etc. The length of those vectors represents the number of features used to describe each sample.

The idea underlying machine learning is the exploration of whether it is possible to split samples into groups based on the values of the features. That is done by finding patterns in the way variables change alone and with respect to each other. Another type of exploration is attempting to find relations between two sets of variables. An example of this is predicting the value of the stock market the next day based on news sentiment and information regarding the relationship between companies on the current day (supplier, competitor, etc.).

There are at least two broad categories of learning: supervised and unsupervised, depending of the information available in the data. If data instances also have labels, expressing membership of some class, meaning that for every sample we know the ground truth value of the variable to predict, then we are dealing with *supervised learning*, and the task is to learn the functional relation between the features and the labels. When labels are not available the data is *unsupervised*. One of the most natural things to do is try to find clusters in the data in order to group together instances that share some characteristics.

With regard to both supervised and unsupervised learning, a vast number of models have been designed and optimised. The most popular machine leaning models are: *Random Forest*, an ensemble model made of a sequence of decision trees, each one independently trained on a different bootstrap sample; *Neural Networks*, a model inspired by the way neurons transmit information in the human brain; *Support Vector Machines*, which learn a nonlinear classifier by applying the so called *kernel trick* to find hyperplanes with maximum-margin in separating different classes; *Probabilistic Methods*, a very broad class of models, including Hidden Markov Models and Bayesian Networks; *Boosting Methods*, techniques developed to increase performance of weak classifiers (e.g. AdaBoost).

3.2.1 Cross Validation

Cross Validation(CV) is an evaluation procedure used to show the robustness of machine learning models. The idea is to divide the dataset in n folds. Then the model will be trained and tested n times, each time training on a different combination of $(n-1)$

folds and testing on the remaining fold. This will lead to n different measurements of performance, which can be used to determine what is the average performance and how much variation there is depending on different training datasets.

3.2.2 Machine learning for object detection

The most commonly adopted baseline for object detection is the Viola-Jones object detector [112], that makes use of Haar features and AdaBoost [40]. Haar features involve the creation of integral images to evaluate rectangular features in constant time. AdaBoost is a technique used to make a series of cascade classifiers from a single weak model. The method is quite fast in testing due to the fast computing of the features and flexible enough to be applied in various contexts, but it is not very robust with regard to rotation of the object or significant luminance variation.

Dalal and Triggs [26] designed an object detector by considering the problem as a binary classification task: distinguishing object patches from background patches. Histograms of Oriented Gradients (HOG) are computed for the whole image at multiple resolutions and then score a set of patches extracted from scanning the image. The task in this setting results in an extremely unbalanced binary classification problem because most images contain far more background pixels that object pixels. This method performed very well on the INRIA dataset, but not as well on a more challenging benchmark dataset, the PASCAL [32].

Further improvements were achieved by combining part-based models with the Discriminative Training (D-T) algorithm [37] [38], giving a linear filter per part. The idea was to combine the responses of the individual filters to find the location of the object together with its parts. Part-based models were popular when initial investigations on object detection were carried out by the research community but they were left aside due to lack of computational power.

3.3 Statistical shape and appearance models

Almost everything we can think of is an object, but every object can appear in the real world in many different instances, that can be very different from each other, within the same class of objects. That is why it is useful to have a model that can capture the underlying variations of a class of objects and express them in a compact way.

Statistical shape models exploit the principle that all examples of some types of

objects can be represented as a mean shape, plus some linear combination of modes of variation.

The first thing we need to do is collect a dataset of images containing the object of interest. Then, and this is probably the most crucial step, we have to choose a set of landmark points. These points have to be well defined in every instance and they have to characterize the object class in some sense. When the landmark points have been chosen, we annotate them in each image, ending up with a set of $2n$-dimensional vectors (n is the number of landmark points) giving the coordinates of these points within the corresponding image.

Since objects in images can appear in different sizes and angles, it is necessary to align the shapes to a common reference frame that makes it possible to compare them effectively using Procrustes Analysis [60]. After the alignment of the shapes Principal Component Analysis (PCA) is applied. The principal components we obtain correspond to the most relevant modes of variation in the data. Figure 3.1 shows the first three modes of a shape model built from faces.

Figure 3.1: An example of three modes of a shape model.

In more detail, a shape model is a mathematical object that represents each shape $x = (x_1, y_1, x_2, y_2, \dots)^T$ in the following way

$$x = T(\bar{x} + P_s b_s; t), \tag{3.1}$$

where $\bar{x}$ is a representation of the mean shape in a suitable reference frame, P_s is a matrix containing a set of modes of variation and T applies a global similarity transformation with parameters t.

The shape parameters b_s can be calculated from x using

$$b_s = P_s^T (T^{-1}(x;t) - \bar{x}). \tag{3.2}$$

3.3.1 Shape Model Matching

Building a shape model is useful to study the within-class variance of a set of objects and it is widely used in medical imaging to find correlations between certain shapes and different progressions of the disease of interest.

In this context a question that naturally arises is the following: can we build an algorithm that, given a shape model and a new image, is able to find the outline of the shape? That specifically means locating each one of the points that constitute the outline of the shape. We can think of shape model matching as a case of Landmark Point Detection with a special emphasis on shape features.

There have been various attempts to solve this task. In the following sections we will introduce some of the most successful approaches: Active Shape Models (ASM), Active Appearance Models (AAM) and Constrained Local Models (CLM).

3.3.2 Active Shape Model

The Active Shape Model (ASM) [21] is a well established technique to fit shape models in new images.

First, we assume that we have already built a shape model and we initialise it in the image we want to test. A poor initialization makes it very hard for the algorithm to converge to a good fit. A good initialisation method is to run an object detector over the image in order to obtain a good approximation of the location of the object. We suppose now that the model has been initialised, so we have a first approximation of where the landmark points are within the image. The algorithm then iteratively looks for new locations for each of the points in the shape. When this model was first developed, this was done by extracting one-dimensional profiles of pixel intensities along the normal to the shape and then building a statistical model of the grey-level structure given by them. The statistical model works under the standard assumption that the data is distributed as a Gaussian. Current implementations of ASM use 2D patches instead of one-dimensional profiles.

In the test phase of the search algorithm profiles extracted from the current approximation of the landmark points are compared to a statistical model built from a training

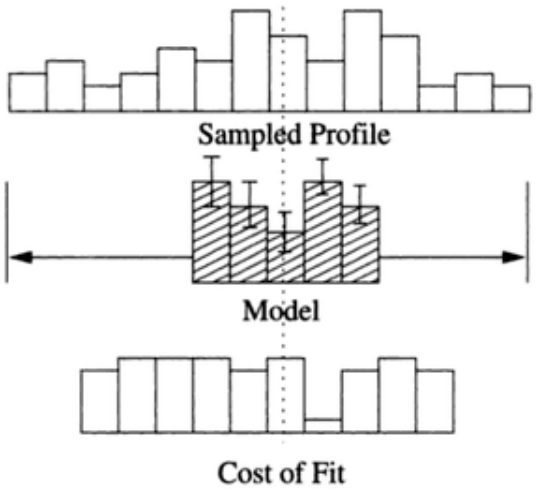


Figure 3.2: Search along sampled profile and evaluation of the fit [20].

set, using the following expression as cost of fit, known as Mahalanobis distance [76] as shown in Figure 3.2.

$$f(g_s) = (g_s - \bar{g})^T {S_g}^{-1} (g_s - \bar{g}), \tag{3.3}$$

where g_s is the vector whose elements are the grey-scale intensities sampled from the profile, $\bar{g}$ is the mean vector of the model and S_g is the covariance matrix. This distance is related to the probability that g_s is drawn from the distribution in the sense that minimising $f(g_s)$ is equivalent to maximising the probability that g_s comes from the distribution.

The ASM model has two major limitations: it treats each model point as independent when searching, and it only makes use of sparse image information around the points. A partial solution to these was given by the Combined Appearance Models [31], that model an object using parameters related to its shape but also the texture intensities that describe it.

3.3.3 Combined Appearance Model

One of the main insights on ASM is that it does not incorporate all grey-level information in its parameters. Combined Appearance Models (CAM) are an attempt to a better use of textural information. They are based on a much more complex statistical model that uses shape as one of its components. In this way we achieve better representation power and this could bring more robustness.

We start by assuming that we have built a shape model of variation, as described in Equation 3.1. Then we warp each of the training images to match the mean shape and we sample the texture information from the resulting objects. To minimise the effect of lighting variation, we normalise the samples using a scale factor α and an offset parameter β. After applying PCA to the set of vectors obtained, we end up with a linear model that follows

$$g = T(\bar{g} + P_g b_g; t), \tag{3.4}$$

where $\bar{g}$ is the mean grey-level vector, P_g is a matrix of eigenvectors, explaining the textural variation and b_g are texture parameters and $T(:,t)$ is an affine transformation (translation, rotation, scaling) with parameter vector t.

An Appearance model can be calculated by concatenating the two models and then a further PCA. This is because shape and texture may be correlated.

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix}, \qquad b = Qc, \tag{3.5}$$

where W_s is a diagonal matrix of weights, the shape eigenvalues, for each shape parameter, b_g are the texture parameters and c are the appearance parameters with corresponding eigenvectors listed in Q.

3.3.4 Random Forest Voting Schemes

Random Forest Voting algorithms, as well as ensemble models in general, work on the principle that a large number of separate independent votes will result in a majority vote on the correct answer.

Algorithms of this kind are now very popular in object detection. Some of them combine the idea of visual codebooks or part-based model to a voting scheme [41, 42].

3.3.5 Random Forest Constrained Local Models

We now focus on one of the implementations of a Random Forest Voting Scheme: the Random Forest Constrained Local Model (RFCLM) [19], which has been shown to produce accurate segmentation consistently with different radiographic datasets in medical imaging [22, 72, 30, 108]. We assume we have a set of annotated images. The idea is to sample patches around each of the landmark points and store the displacement

of the center of the patch with respect to the location of the corresponding annotated point.

This data is used to train a Random Forest Regressor per landmark point. Each forest has to learn the functional relation between the pixel intensities within the patches and the associated displacement. Haar features are used to find the optimal split during training.

During testing, we locally sample a set of patches, around the current approximation of each landmark point. If we feed them into the trained Forest each patch will end in a leaf in each tree of the forest, making a set of predicted displacements. The idea now is to store the locations predicted using the displacements as votes in a response image with the same size as the original image, as shown in Figure 3.3.

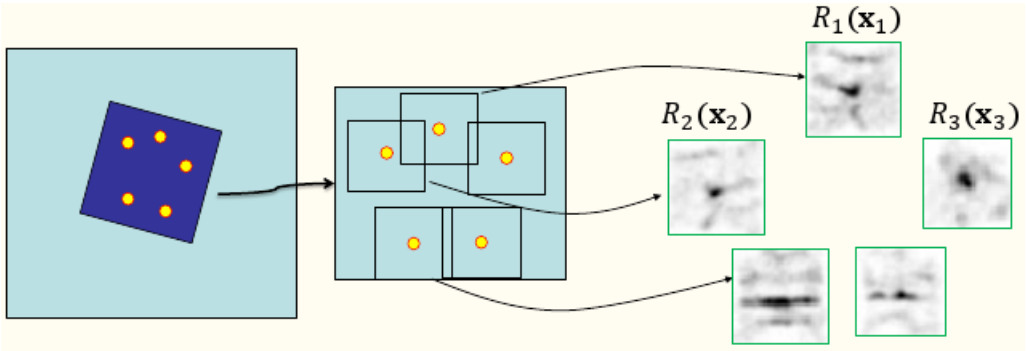


Figure 3.3: CLMs are first initialised on the test image, then a point regressor focuses on a specific region resulting in a response image per point.

We are looking for locations with high number of votes, but we also want the overall outline of the points to agree with the shape model. This is done by setting a constrained optimisation problem. The function to maximise is a function V counting the number of votes of each of the coordinates in the voting image. Then we add a constraint term giving an estimate of the log-likelihood of the shape given the shape parameters associated to the current outline. This term penalises all those shapes that are far from the mean shape, according to the Mahalanobis distance.

The RFCLM is both efficient and robust and has proven to converge even in cases where the initialisation of the model is far from accurate. More details about this method and the way it was used as part of the project are given in the following chapter.

3.4 Automated Methods for studying OA

In this section we explore a variety of computer aided approaches for studying OA. These include methods for automatically evaluating specific radiographic features such as trabecular bone or joint space width; other methods were built trying to diagnose and assess the severity of knee OA by determining the appropriate KL grade. Finally, we will cover examples of techniques used to determine the chance of a person developing the disease in the future depending on the current state of the joint and how well radiographic features can be used to discriminate both people experiencing pain and people who will develop symptoms in the future from their corresponding controls.

It is possible to distinguish two main classes of algorithms: the semi-automated and the fully automated. Any technique of the first kind needs at some point in the process some sort of human intervention, usually by an experienced clinician. On the other hand, fully automated methods do not need any human input.

3.4.1 Fractal Signature based methods

The first kind of approaches are those that study trabecular bone, specifically its variation in thickness and orientation by using Fractal Signature based methods. Knee Osteoarthritis causes the trabecular bone of the tibia to thicken and this can be visualised in radiographs as horizontal striations. Calculating the fractal dimension of texture of those regions can help establishing the degree of thickness due to the disease and consequently its progression.

Kraus et al. [65] performed fractal analysis of the subchondral tibial plateau and combined it with both radiographic measurements such as knee alignment and bone mineral content and covariate features like age, gender and BMI. Using a total of 138 participants from the Prediction of Osteoarthritis Progression (POP) study, the work showed that fractal analysis was able to predict JSN progression on the medial side, while failing to perform as well on the lateral side. The best collections of features found were the combination of radiographical features with age, gender and BMI with the best AUC reported being 0.79.

Podsiadlo et al. [90] introduced a modified Hurst orientation transform (HOT) method able to characterise trabecular bone texture with a better focus on descriptors for roughness and anisotropy. The authors showed the effectiveness of the methodology as well as its robustness with respect to the presence of noise, blurring effects and other types of artifacts.

Wolski et al. [117, 116] aimed to evaluate if there were statistical differences in trabecular bone texture in radiographs between people with OA and those without it. The methodology involves a novel use of the variance orientation transform (VOT), compared to the HOT method. The VOT method was chosen because it allows us to calculate the fractal dimension and signature in all directions efficiently. Results obtained show that the fractal dimension of cases in several locations was lower than the same variable for control patients, while VOT and HOT had the same reproducibility of texture features and the same discriminative power for binary OA diagnosis [1].

Another fully automated method is the one of Stachowiak et al. [102]. In this work knee and hand radiographs are studied. Our review will cover the steps applied to analyse the knee joint. First image enhancement techniques are used and then Active Shape Models retrieve the outline of the tibial plates. ROIs are then placed making sure that there is no overlap with periarticular osteopenia, subchondral bone sclerosis and fibula head. High agreement was shown between automatically found ROIs and regions selected by clinicians (Similarity Index (SI) ≥ 0.81). Fractal signature of these ROIs is then calculated using the VOT method together with the Hurst coefficient and the following features are extracted: the roughest part of the FS, texture aspect ratio signature and texture direction signature. Differences in trabecular bone texture using these measurements were evaluated between different groups (OA vs no-OA, cartilage defects vs no cartilage defects and others) consistently finding significant differences. As an example, texture extracted from OA affected knee radiographs were on average smoother and less anisotropic than textures in healthy joints. These results were consistent with previous findings of similar studies as described above.

3.4.2 Joint Space Width

The first work in automated analysis of the Joint Space Width was by Dacre et al. [25]. The authors used a standard square grid as template and a computer generated refined grid in order to find areas of joint space. Those areas can then be measured just by counting the number of pixels inside them. Similarly, the authors computed joint space distance and showed that healthy knees have higher joint space width than knees with OA.

An early semi-automated study of OA was developed by Duryea et al. [29]. The idea was to design an algorithm to automate the measurement of the minimal Joint

[1] By this, we mean splitting the radiographic set into two groups: the ones with KL grade equal to 0 or 1 and everything else.

Space Width in PA images. This work is semi-automated since the radiographs have been manually cropped before training. The techniques used involve the application of the gradient based Gabor filter and a region growing algorithm to find an approximate segmentation of femur and tibia. Then, the minimal distance point between the contours is measured to obtain the JSW estimation. The reproducibility results are perhaps the greatest merit of this work, while the algorithm was noted to not perform in cases where complete surface contact was found between the femur and tibia since the method looks at sub-chondral surfaces.

The KIDA algorithm [78] was proposed by Marijnissen et al. to measure a number of osteoarthritic features from radiographs, JSW being one of them. The algorithm works as follows: first, in order to derive bone density estimation and the magnification value of the radiograph the user has to choose the reference wedge, from a set of 15 options, that best compares with the image texture. The software will automatically compute a reference length based on it. A framework of four lines is then added to manually locate the joint within the image. Those lines will also be used to find the profile of the bone cartilage and the subchondral area. The operator then has to manually place a set of circles in order to more accurately define the interface of the articular cartilage to get an approximation of the joint space width. Further manipulation can be done by the software to obtain other measurements such as osteophyte margin and joint angle.

The work of Grochowski et al. [47] uses lateral knee radiographs (one of the few studies to do so) to build a semi-automated method for measuring the patello-femoral JSW. The method is not fully automated because the user is required to crop the radiograph, obtaining a patch containing the patello-femoral joint. The Canny edge detector [14] was applied to segment the lateral femoral epicondyle. The edge of the patella was found by computing the intensity gradients of the image and then by determining the brightest pixels in each horizontal scan line. The minimal JSW was computed by comparing the x-axis coordinates of the outlines previously found. The repeatability intra and extra observer was evaluated as 0.03 ± 0.09 mm and 0.01 ± 0.12 mm, respectively and reproducibility after the collection of a new set of radiographs two weeks later was measured as 0.09 ± 0.73 mm for the same technician. The size of the dataset used was small, including images taken from only 35 participants.

3.4.3 OA Classification

Oka et al. [87] proposed the KOACAD software to measure various osteoarthritic features. After correcting for magnification and reducing noise with a multiple application of median filtering, the method applies Robert's filter to extract an initial approximation of the outline of both femur and tibia. Then a rectangle containing the joint space is found by applying a vertical neighborhood difference filter. Further application of the same filter plus Canny filter gave the outline of the femoral condyles and the tibial plateau. The middle line between the two profiles was chosen as the lower bound of the JSW. The profiles of the bones found earlier were used as well to compute the joint space area and the tibiofemoral angle. Further processing gives the osteophyte area as well. The paper found that all measurements were highly correlated with osteoarthritic grades (both OARSI and KL grades) and that medial joint space narrowing and varus angulation were risk factors for the onset of knee pain.

The semi-automated work of Hladůvka et al. [50] is based on the manual input of the user who has to select several ROIs from PA high resolution radiographs. The resulting radiographic patches are analysed computing the Hurst coefficient and the Shannon entropy of the texture. These measurements are then used as features in order to perform binary classification of radiographic OA. Results found that a linear classifier trained on only 5 features of the ones extracted leads to an AUC of 0.85, higher than the state of the art at the time, but evaluated on a significantly smaller dataset.

The work proposed by Gornale et al. [45] is a fully automated method that finds the location of the knee joint by applying an Active Contour segmentation technique. The region is analysed by extracting a large number of features (Haralick, mean, area, entropy, Euler number of images, the first four moments and others). A random forest classifier is then trained on the individual sets of features and on the combinations of all descriptors. The best overall performance, with an accuracy of 87.92%, was achieved by the fusion of all available features. Similarly to other cases, the model is evaluated on a rather small dataset containing only a few hundred images.

Another fully automated method is the work of Thomson et al. [108]. The authors worked with 500 posterior anterior knee joint radiographs to automatically classify the presence of Osteoarthritis. To this aim they considered as part of the OA group all those patients with KL grade equal to 2 or more and non-OA all the other patients. The images were manually annotated using 74 points in order to build a shape model of the tibiofemoral joint. A global object detector based on Random Forests was trained

to find the approximate location of two of those points in order to initialise a RFCLM to locate each individual point in the bony outline. Then, shape parameters were extracted together with textural information that quantifies the fractal signatures of tibial texture. Results included an AUC of 0.845 for the fully automated method trained on the whole set of features with the shape features proving to be more discriminative that the textural ones.

Continuing from the previous work, Thomson et al. [109] used the old 74 point model for PA knee radiographs, but extended it to specifically study osteophytes using three different approaches. The first one augmented the model with 44 extra points to capture the shape of osteophytes, where present. The second approach was based on extending the profile defined by the original points by looking for strong edges along the direction normal to the shape surface. This problem was modeled as an optimisation problem and solved using dynamic programming. The third approach was using 4 ROIs based on the 74 point annotation. Those ROIs were located in regions of the joint were osteophytes are most likely to develop. Haar features calculated from the texture of the ROIs were used to train random forest classifiers. Features were associated to the first two approaches by building a statistical shape model of the resulting shapes and extracting shape parameters corresponding to each particular shape instance. The initial problem was to automatically say if an osteophyte was present by splitting the corresponding OARSI grade into two groups (0-1 and 2-3). The best achieved result for this problem was AUC 0.846 ± 0.014, obtained by combining all the available features. Using the same features and methods the authors perform the standard KL grade classification task and Binary OA classification task on a dataset of over 500 radiographs obtaining in both cases the state of the art of $50.2 \pm 0.5\%$ (multi-class accuracy) and AUC 0.931 ± 0.002, respectively.

The work described in [56] was automated detection of OA using Infrared Thermography images, that provide functional information on thermal and vascular conditions of knee joints. The idea was to use a semi-automated feature extractor algorithm based on patella-centering and then feed the found features to a SVM classifier. The size of the dataset does not allow to draw definite conclusions and repeatability of the experiments was not investigated, but this work showed the discriminative potential of features extracted from Infrared Thermography images.

Shamir et al. [97] developed a method for automated detection of Osteoarthritis. The knee joint was found by computing the Euclidean distances between 20 predefined 150×150 pixels knee joint patches and image patches of the same size extracted from

a large number of locations in each image. The idea was then to extract a large set of discriminative features such as: Zernike features, multi-scale histograms, first four moments, Tamura texture features, Haralick features or Chebyshev statistics. Then weights were assigned to each feature using Fisher scores, to give more importance to features with more discriminative potential. A Nearest Neighbor classifier was trained on the weighed features to predict the first 4 KL grades. This method distinguished KL grade 3 and KL grade 2 from KL grade 0 with an accuracy of 91.5% and 80.4%, respectively, on a dataset made of 350 PA radiographs. Feature extraction was very time consuming, especially for Zernike features, making the method not suitable for real time applications.

In a study by Anifah et al. [6], 303 PA knee radiographs collected by the OAI [69] were first preprocessed by applying Contrast Limited Adaptive Histogram Equalization and then segmented using Gabor kernel, template matching, row sum graph and grey level center of mass method. Then the gray tone spatial dependency matrix was built and 4 features, namely the contrast, correlation, energy and homogeneity were extracted and fed to a Self Organizing Map (SOM) [63] for classification. The results were a rate of 93.8% for KL-Grade 70% for KL-Grade 1, 4% for KL-Grade 2, 10% for KL-grade 3 and 88.9% for KL-Grade 4.

CNN based methods

In recent years, we have seen increasing attention by the medical imaging research community dedicated to develop CNN based architectures. Some of them were used for classification purposes, others to look for specific structures in the data. In general, they have shown to be able to solve some of the most relevant tasks in the field, finding ways to get around the use of hand crafted features. CNNs provide features that are learned from the data and consequently have potential to better describe phenomena. Here, we will focus on Osteoarthritis Imaging applications.

The first work exploring the potential of deep learning in imaging for Osteoarthritis, was the work by Antony et al. [8]. A patch containing the knee joint was found by labeling positive patches of the centre of the joint and negative patches outside the center of the joint. Sobel horizontal image gradients were used as features and the binary classification was done using a linear SVM model. In testing phase, a sliding window technique was used to find the centre of the joint in the unseen radiographs. A patch of 300×300 pixels was then extracted based on the smaller centre patch. Two CNN approaches were then used. In the first one features were extracted from a

VGG16 architecture trained on the OAI and a linear SVM was trained on those features to classify radiographs based on their KL grade. The second approach fine-tuned two architectures pre-trained on ImageNet replacing the top fully connected layer in both instances. The best performing model was a linear SVM trained on features extracted from the last fully connected layer of a fine-tuned BVLC NET network. The corresponding accuracy of 57.5% was the state of the art in KL grade classification.

A second work from the same group of authors [7] improved the model further with two main adjustments. The first one was a knee joint detector based on CNNs. The idea is to feed a radiograph and to get as output a binary segmentation mask describing the location the knee joint within the image. Furthermore, the found patches are then used to train Fully Connected Networks(FCN) for KL grade classification. The classification network was trained from scratch to minimise a combined classification and regression loss (categorical cross-entropy and sum of squared differences). The performance of this method achieved a multi-class accuracy of 61.9% further improving on the state of the art.

The last and most recent work that uses CNNs is the one of Tiulpin et al. [110]. Using an automated method for locating the knee joint, the authors developed a collection of three Deep Siamese networks using three different random seeds. All three networks share the same architecture and take as input two smaller patches, one containing only the lateral side of the joint and a second one being the flipped version of the medial side. A softmax layer is responsible for weighting the outputs of the network and providing the final distribution over the KL grades. This work is the first one using the MOST dataset for training and OAI for testing. Multi-class accuracy increases further on previous work reaching 66.71%. Attention maps showing what areas in the images were important to lead to the prediction were also included, showing areas where information about the disease is available.

3.4.4 OA Prediction

Here we focus our attention on studies whose aim was to look at longitudinal patterns, looking at predicting future development based on current measurements.

The first study that is worth mentioning is of Kinds et al. [61]. The authors considered baseline participant of the CHECK cohort [114] with knee pain, and looked at measuring several radiographic osteoarthritic features using the KIDA software. Multivariate regression was used to predict incidence of radiographic OA at a 5-year follow up visit. The study compared whether adding radiographic features to demographic

and clinical characteristics improved performance. The results showed the best sets of features were the combination of osteophyte area and minimum joint space width with demographic and clinical features (AUC of 0.74) but clinical OA development could not be predicted more accurately using any radiographic measurement.

In the study from Bowes et al. [11] over 2000 MRI scans of the knee were segmented using Active Appearance Models [18] allowing for measurements of subchondral bone. Participants with OA showed an increase in bone area over time, while controls remained overall stable. Bone was more responsive than more standard osteoarthritic features such as cartilage thickness and JSW.

In the work of Yoo et al. [119], the authors trained a Logistic Regressor and an Artificial Neural Network to predict future onset of radiographic and symptomatic OA, just using features extracted from clinical data, without any radiographic image. The features used included sex, age, body mass index, educational status, hypertension, moderate physical activity and knee pain. The full dataset contained about 2500 patients from the KNHANES V-1 study. The Logistic regressor and the ANN predicted radiographic OA with AUC of respectively 0.62 and 0.67 and symptomatic OA with AUC of 0.70 and 0.76.

The same technique described in [97] was applied in [96] to detect future onset of Osteoarthritis using a follow-up of approximately 20 years. The data used was collected by the Baltimore Longitudinal Study of Aging [57]. In the experiments they used just a few hundred images to train a classifier to distinguish between patients that will develop OA with KL grade 3 from the ones that will not develop it at all, achieving 72% accuracy. Similar experiments were performed to distinguish whether a knee will change to KL grade 2 from KL grade 0 or it will remain the same. The corresponding accuracy was 62%, also finding that features extracted from the tibial spines had strong predictive signal.

3.4.5 Pain

In this last section we cover studies looking for ways of predicting current and future pain in Osteoarthritis.

Galvan et al. [43] used data from the OAI study to look at current radiographic OA features and build a univariate logistic regression test to find out what features were most associated with future pain. The study found that early osteophytes were the feature with the highest association and, in contrast with previous analyses, joint space reduction was not associated with future joint pain.

Finally the work of Luna-Gomez et al. [75] studied the discriminative ability of MOAKS (MRI Osteoarthritis Knee Score) to predict future knee pain. These scores are manual grades assigned as part of the OAI study. The experiments were divided into three time points all of which achieved significant association with future knee pain. An AUC above 0.60 was reported in each of the experiments.

Chapter 4

Fully Automated Classification and Prediction of Osteoarthritis

4.1 Lateral Knee Radiographs

The project started from the observation that lateral knee radiograph images are often available or, even when they are not, they are easy, cheap and fast to acquire. Furthermore, the lateral view contains very informative features that cannot be read from PA images. Several works in the field have emphasised the need for better use of the available resources and lateral radiographs are often cited as the main addition to explore [79, 35]

4.1.1 Choice of Landmarks

The choice of the landmarks' point locations was done in order to capture anatomical landmarks. Additionally, equally spaced points were inserted between them to sufficiently capture the shape profiles. The annotation was performed by one annotator only, initially placing points by hand. Once the landmark point detector was performing reasonably well the software was used to speed up manual annotation.

4.1.2 Parameter Optimisation

When using the RFCLM algorithm, there are a number of parameters that can be optimised to best fit the data being used to train the model. These parameters relate to various characteristics of the way the algorithm works and the way training and testing images are treated, including the size of the images, and the necessary level of

detail required to analyse the images. To achieve the best accuracy and efficiency of the model, these parameters had to be carefully tuned by experimenting with different value settings.

The most relevant parameters to tune in the RFCLM were:

- *Frame Width*, representing the size that the object of interest will have in the reference frame;
- *Patch Size*, this parameter is highly related to the previous one and it encodes the width of the sub-images that are sampled for training, hence containing all the features used by the CLM;
- *Maximum Displacements*, this parameter defines the maximum pixel displacement that training samples can have with respect to the corresponding landmark point;
- *Search Range Border* defines the size in pixels of an extra border around the mean shape point that the searcher can extract patches from;
- *Optimiser Radius* is the maximum distance allowed for points to be chosen as candidates of landmark points when maximising the response image values constrained to the shape model.

4.2 Papers

The following works were considered state of the art at the time of submission in the tasks that were approached in them. The first paper describes the development of a landmark point detector to find the outline of the knee joint in lateral knee radiographs. The found segmentations are then used to distinguish diseased knees from healthy ones. An investigation of which features are most informative and what is the best way to combine radiographic features was also performed.

The second paper looks at the combination of radiographic features from the two views that are most commonly available. In this work we investigated whether concatenating these features improves performance and we looked at future pain prediction.

Fully Automated Shape Analysis for Detection of Osteoarthritis from Lateral Knee Radiographs

Luca Minciullo, and Timothy F. Cootes

The University of Manchester

This work has been published in the proceedings on the International Conference on Pattern Recognition 2016 (ICPR)

Contribution of the thesis author: literature research, conception and design of the study, classification experiments using the code developed by Tim Cootes and adapted to use for the project, interpretation of results, drafting and revising of paper content, paper submission.

Keywords: Random Forests, Decision Trees, Optimisation

Abstract

Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world. Since no cure has been discovered and considering the financial impact on health systems, any attempt to understand more of this disease could reveal new insights that would help develop new therapies. Lateral knee radiographs are often ignored both by clinicians and the research community when trying to diagnose OA or other diseases that affect the knee joint. Our goal is to show that this view has a considerable potential. We present a fully automated method based on a Random Forest Regression Voting Constrained Local Model (RFCLM) to discriminate radiographs of people that have developed OA from people who have not. The experiments involved models built on different combinations of the four shapes (patella, tibia, medial and lateral femoral condyles) of the knee joint. We show that automated analysis of the lateral view achieves classification performance comparable if not better than similar techniques applied to the frontal view.

4.2.1 Introduction

Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world. It has been reported [36] that by 2030 around 20% of the American population will be above the age of 65, and that half of them (35 million patients) will be at high risk of developing OA.

Since no cure has been discovered and considering the financial impact on health systems [15], any attempt to understand more of this disease could reveal new insights that would help develop new therapies.

OA is currently assessed from radiographs using the Kellgren and Lawrence (KL) [59] grades from 0 to 4, where 0 represents normality and 4 the most severe stage of OA. When a radiograph is taken clinicians assign a discrete KL grade based on features in the image. This is time consuming, subjective and there are shortages of suitably trained radiologists. There is an increasing need for reliable systems that can perform the grading automatically.

We describe the first fully automated system for classifying OA and KL grades from lateral radiographs of the knee. We evaluate its performance and show that analysing the shape of bones in lateral images gives better results than using the shape in PA (frontal) views. The lateral view is often ignored both by clinicians and the research community when trying to diagnose OA or other diseases that affect the knee joint. We show that this view has a considerable potential when trying to assess the state of OA from a radiograph.

We used a Random Forest Regression Voting Constrained Local Model (RFCLM) [19], [22], [72] to locate points in both single bones and combinations of bones. We used an object detector based on Random Forests (RF) to automatically initialise the RFCLM on each image.

The RFCLM returns the found points and the associated shape parameter vector. We used the components of these vectors as features on which to train a Random Forest classifier.

In this work we are interested in the level of classification performance that is achievable using just information coming from the shape of the bones in a lateral knee radiograph. We used features related to the texture only to train our Random Forest landmark point detector.

We performed experiments on both binary (OA vs No OA) and 5 class (the 5 K-L grades) classification, using shape information from the manual annotation and from a fully automated system.

Our approach is similar to that in [108], where the authors studied Posterior-Anterior (PA) knee radiographs to retrieve shape and texture features. They used a RFCLM on a 74 points model and extracted features of tibial texture. Other approaches are [97] and [6], where image processing techniques are applied to PA knee radiographs: in the former the authors extracted image content descriptors and image transforms to use as features in a Nearest Neighbor setting; the latter applied the unsupervised self organizing maps based on Gabor filter to classify the K-L grades. In [56], the authors used medical infrared thermography of the PA view to extract features on which train a SVM classifier.

4.2.2 Method

Our model is made of four different sub-shapes: the patella (21 points), the lateral femoral condyle (24 points), the medial femoral condyle (25 points) and the tibia (32 points). The whole knee model is then made of 102 points (Figure 4.1).

We analysed different combinations of these shapes in order to understand which features are most informative.

4.2.3 Statistical Shape Model

A shape model [20] can be obtained by applying the Principal Component Analysis (PCA) to a set of aligned shapes (vectors). We used a linear model of shape variation, that represents each shape $x = (x_1, y_1, x_2, y_2, \dots)^T$ in the following way

$$x = T(\bar{x} + Pb; t), \tag{4.1}$$

where $\bar{x}$ is a representation of the mean shape in a suitable reference frame, P is a matrix containing a set of modes of variation (Figure 4.2) and T applies a global similarity transformation with parameters t.

The shape parameters b can be calculated from x using

$$b = P^T(T^{-1}(x; t) - \hat{x}). \tag{4.2}$$

These parameters were the features in our classification tasks.

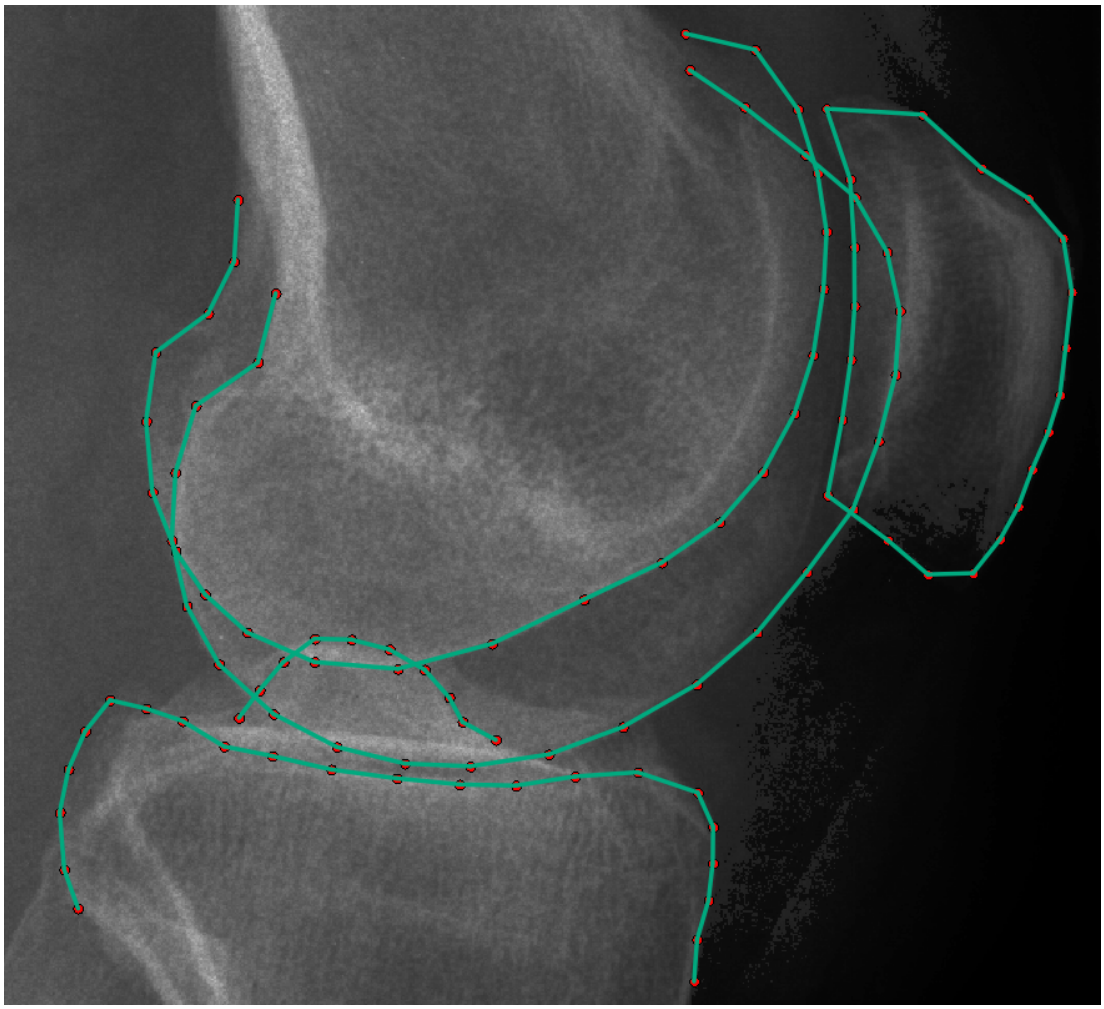


Figure 4.1: An example of the 102 landmark points used to build the shape model.

Shape Model Matching

The first step was to build a global searcher able to find the individual bones within each image. We used a Hough Forest approach [41]. We defined a bounding box starting from a pair of landmark points and then sampled from each image a set of 23×23 patches with different displacements, angles and scales with respect to the location of the bone of interest. We then trained a Random Forest to learn the functional relation between the pixel intensities in the image patches and the corresponding displacements. This RF is scanned over a new image at multiple scales and orientations, voting for likely knee locations. The output of the global search is a bounding box with two reference points, from which we initialise each model (Figure 4.3).

In the second step we improve the fitting of the model, by applying a sequence of

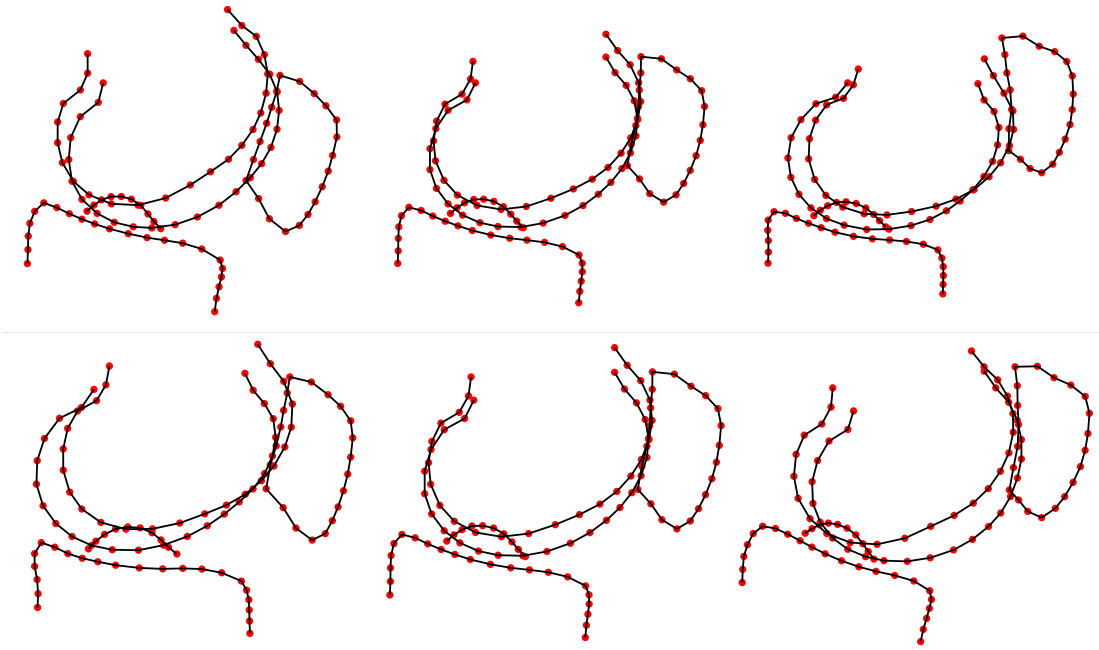


Figure 4.2: The first (above) and the second (below) shape model modes of variation.

increasingly refined Constrained Local Models. The idea is to independently train a point detector per landmark point.

In the search phase we sample a set of patches around the current approximation of the point location. We feed those patches into the Random Forest, receiving a prediction per patch and tree for the location of the landmark point of interest. We combine all the predictions in a voting image $V_i()$ for each point i. The shape model is used to regularise the result, finding the parameters b,t which maximise the total votes $Q(b,t) = \sum_{i=1}^{n} V_i(T(\bar{x}_i + P_i b_i; t))$. Our implementation involved a sequence of three increasingly refined CLM, with frame width equal to 50, 100 and 200 pixels.

Classification

The approach above enables us to fully automatically locate the points of the outlines of the bones in new images. From both manual and automated annotation we can find shape parameters from the statistical shape model. The shape parameters are weights representing which modes of variation were found in the data instance and with what magnitude.

We train Random Forests on combinations of shape parameters, in order to predict: (a) OA vs non-OA, (b) the KL grade. Our implementation involves 100 trees per Forest

Figure 4.3: An example of the bounding boxes found by the Random Forest bone detector.

and we use two stopping criteria when building the trees: the maximum depth that a tree can have and the minimum entropy in the data. As soon as one of these two conditions is met we stop splitting.

4.2.4 Results

Data

Our dataset is made of 300 lateral knee radiographs, 60 images per grade, from the MOST (Multicenter Osteoarthritis Study) dataset [36]. MOST is a longitudinal, prospective study of Knee OA run in the United States and involves 3026 participants, men and women aged 50 to 79. Each participant makes one visit per year, with about 5 visits in total. The dataset also contains various information recorded at each visit, including the KL grades from 0 to 4, indicating respectively: normal, doubtful, minimal, moderate, severe.

For the binary classification task the grades have been split into two groups: non-OA, KL (0,1), and the OA group, KL(2-4).

Table 4.1: Point detection results (mm)

Shape	Mean	Median	90%
Patella	0.24	0.17	0.45
Lateral Femoral Condyle	1.04	0.72	2.21
Medial Femoral Condyle	1.18	0.86	2.35
Tibia	0.98	0.81	2.39

Landmark point detection

Finding Landmark points in a radiograph is challenging due to the way bones overlap in the projection. The lateral view is even more challenging than the frontal view since the two femoral condyles look almost identical, making them difficult to distinguish even for an experienced clinician. Table 4.1 contains the results of experiments evaluating the accuracy of the local search for each of the bony outlines shown in Figure 4.1. The latter results were obtained by training the model on 200 examples and testing on the remaining 100.

Our model performs well on the patella, where the error on 90% of the examples is less than $0.5mm$. This is probably due to the lower spatial variation of the landmark points describing this shape. The accuracy of detection of the other shapes is considerably worse, though we always have a median error of less than a millimeter.

OA Classification

The results shown in this section are obtained performing a 5-fold cross validation. For both binary and 5-class classification we trained a Random Forest made of 100 trees.

Our experiments have been performed first looking at the shape parameters from each shape, both using manual and fully automated annotation. We also concatenated different combinations of the points and then built the shape model on the concatenated points. In this way we can assess what shapes or combination of shapes contain more information and how big is the loss in performance when we move from a manual annotation to a fully automated one.

Table 4.2: AUC for manual and fully automated annotation. Results from concatenating points.

Shape	Manual	Fully Automated
Patella	0.759 ± 0.021	0.651 ± 0.008
Lateral Femoral Condyle	0.666 ± 0.02	0.632 ± 0.007
Medial Femoral Condyle	0.671 ± 0.011	0.711 ± 0.016
Tibia	0.771 ± 0.008	$\mathbf{0.73 \pm 0.017}$
Patella+LCon	0.72 ± 0.005	0.617 ± 0.013
Patella+LCon+MCon	0.754 ± 0.011	0.714 ± 0.014
Pat+LCon+MCon+Tibia	$\mathbf{0.842 \pm 0.017}$	0.711 ± 0.013

We also used the combinations of different shape parameters, independently obtained from different shape models, as features.

Binary Classification. As we can see from Table 4.2, the Tibia and the Patella are the two shapes whose features achieve the best individual classification accuracy. The best overall performance is obtained by the whole knee model in the manual annotation and by the individual tibia model in the fully automated system. The AUC for the Medial Femoral Condyle is higher in the fully automated system than in the model trained on features built on manual annotation.

The best results for the fully automated system were achieved when we concatenated the shape parameters of different shapes, calculated independently. In this way we ignore the relative position of the different shapes. The results corresponding to these experiments are shown in Table 4.3. The overall best binary classification performance of our fully automated system is achieved by the knee model given by the concatenation of the shape parameters of the four sub-shapes. Figure 4.4 shows the results of the concatenation of the different shape parameters with manual annotation. If we add the lateral femoral condyle parameters to the patella we obtain a ROC curve that is consistently lower than the one related to just the patella. Figure 4.5 shows the ROC curves corresponding to the different combination of shapes obtained by concatenating the shape parameters. Concatenating shape parameters leads to improved ROC curves in each case.

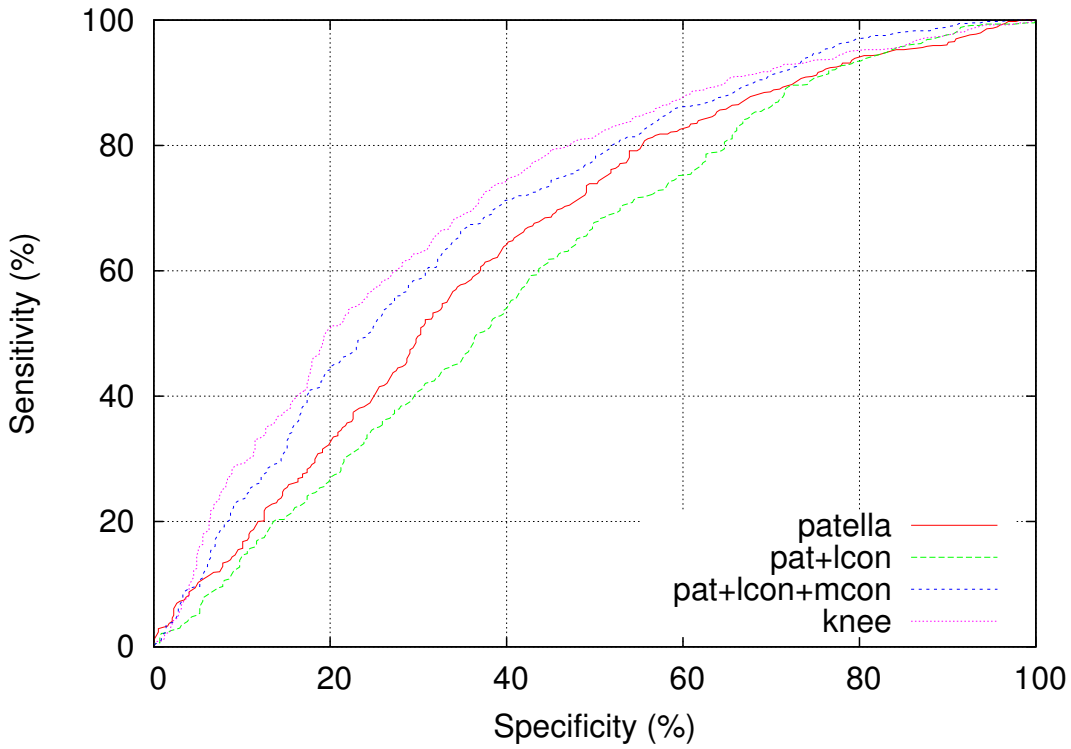


Figure 4.4: The ROC curves corresponding to the different concatenations of the shape parameters based on manual annotation.

We are not aware of any other work investigating the potential of the lateral view. Results using a similar approach on 500 Posterior-Anterior (PA) radiographs from the OAI dataset [69] are given in [108].

The comparison of their best results using shape parameters with our performance is reported in Table 4.4. Although we are dealing with results coming from different views and different datasets this suggests that models trained on the lateral view show great promise.

5-Class Classification. In this section we describe the results corresponding to the 5-class classification task, a considerably more challenging problem. In this case we train a Random Forest classifier to predict the KL grade from the shape parameters. In Table 4.5 we show the results for this new task in terms of the proportion of data correctly classified.

The individual shapes whose shape parameters perform best are the Patella from manual annotation and the Medial Femoral Condyle for the fully automated system. We see again two examples where the fully automated system has better performance

Table 4.3: AUC for concatenation of shape parameters deriving from individually trained shape models.

Shape	Manual	Fully Automated
Patella+LCon	0.755 ± 0.014	0.695 ± 0.007
Patella+LCon+MCon	0.785 ± 0.018	0.719 ± 0.017
Knee	$\mathbf{0.827 \pm 0.006}$	$\mathbf{0.794 \pm 0.015}$

Table 4.4: A comparison between our best results and the ones in [108].

Shape	Our Method	[108]
Manual	$\mathbf{0.842 \pm 0.017}$	0.796
Fully Automated	$\mathbf{0.794 \pm 0.015}$	0.789

Table 4.5: Proportion of the data correctly classified for manual and fully automated annotation. KL-grade classification problem.

Shape	Manual	Fully Automated
Patella	45.3 ± 3.3	29.8 ± 1.3
Lateral Femoral Condyle	32.6 ± 1.2	30.7 ± 1.7
Medial Femoral Condyle	35.6 ± 1.6	36.2 ± 3
Tibia	30.9 ± 2.8	32.1 ± 0.7
Patella+LCon	40.6 ± 0.8	33.7 ± 1.6
Patella+LCon+MCon	45.3 ± 2.8	37.9 ± 1.7
Knee	$\mathbf{47.9 \pm 0.8}$	$\mathbf{43.9 \pm 1}$

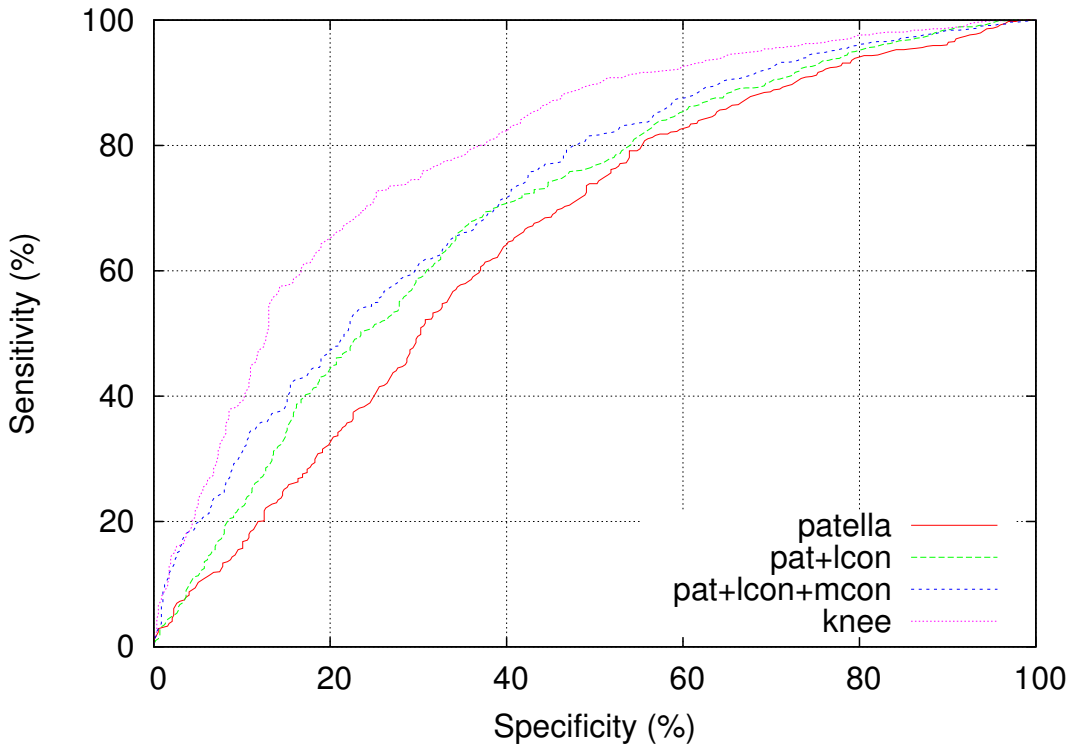


Figure 4.5: The ROC curves corresponding to the different concatenations of the shape parameters based on fully automated annotation.

than the manual system. However, the best overall performances for both annotations are achieved by the full knee model. For completeness we also report the confusion matrices for both knee models (Tables 4.6 and 4.7).

With regard to the manual annotation (Table 4.6), radiographs with grade 0 were easiest to classify. It is encouraging to observe that in the vast majority of instances the mislabeled images were assigned to grades close to the ground truth.

Table 4.7 is the confusion matrix corresponding to the fully automated model. In this case grade 0 turned out to be the hardest to classify, being often mislabeled with grade 1 and 2. Conversely, the fully automated model performs better than the manual one in classifying grade 1 and 2 and they achieve exactly the same accuracy on grade 3.

The proportion of data correctly classified by the model using concatenated shape parameters is reported in Table 4.8. We observe an overall increasing trend when adding more shapes, but unlike what we saw with binary classification the results on automated annotation are consistently worse if compared to the ones obtained by concatenating the different annotations. The overall best results are achieved both by

Table 4.6: Confusion Matrix of the full knee model built on manual annotation (All standard deviations less than 3.2%)

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**55.3**	19.7	12.5	9.2	3.4
Class 1	21.8	**45.7**	20.0	8.6	3.9
Class 2	17.5	9.5	**49.2**	11.1	12.7
Class 3	14.7	9.0	13.3	**38.3**	24.7
Class 4	8.8	8.4	8.8	23.1	**50.9**

Table 4.7: Confusion Matrix of the full knee model built on fully automated annotation (All standard deviations less than 2.5%)

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**33.6**	23.4	20.3	14.9	7.8
Class 1	17.9	**47.5**	16.4	6.4	11.8
Class 2	12.1	13.7	**50.8**	14.6	8.9
Class 3	13.7	14.0	16.7	**38.3**	17.3
Class 4	8.1	12.8	14.4	15.9	**48.8**

Table 4.8: Proportion of the data correctly classified for concatenations of shape parameters deriving from individually trained shape models.

Shape	Manual	Fully Automated
Patella+LCon	45.8 ± 2.8	32.4 ± 1.9
Patella+LCon+MCon	47.1 ± 0.7	35.6 ± 2.3
Knee	$\mathbf{47 \pm 1.6}$	$\mathbf{39.2 \pm 1.4}$

the manual and the automated system by the full knee models built on the concatenated point annotations, with a proportion of data correctly classified of respectively $47.9 \pm 0.8\%$ and $43.9 \pm 1\%$.

4.2.5 Conclusion

We have shown the first attempt at building a fully automated system to classify OA and OA grades using shape information from lateral knee radiographs. The results suggest that the lateral view contains very informative features that can achieve performance of the same level if not better to that of the PA view. One of the reasons for this is that one of the bones that is most affected by knee OA, the patella, is clearly visible in the lateral view, but it is obscured by the femur in PA images.

There is still a great room for improvement in fully automated OA diagnosis. In future work we will use a combination of shape and texture parameters for the lateral view. Informative texture features can be found near the tibial spines and, in general, in all the locations that are most likely to develop osteophytes. When concatenating shape parameters coming from different models it would be interesting to apply some sort of feature selection, in order minimise the noise in the data. Furthermore, it is worth investigating a way of concatenating features from both views.

Finally, we will be studying techniques able to quantify the risk for a patient to develop OA in the near future given the current state of the joint.

4.3 Combining features from both radiographic views

One natural question that arose from the previous study was whether simply combining radiographic features from lateral and PA image could improve performance at classifying if a person had knee OA. Furthermore, several other learning tasks could be attempted such as future radiographic OA prediction and onset of knee pain. In addition, the following work aimed at providing the first direct comparison of the features of the two views, while the previous work was done with PA images of the OAI dataset and lateral images from the MOST dataset.

Combination of Lateral and PA View Radiographs to Study Development of Knee OA and Associated Pain

Luca Minciullo, Jessie Thomson and T.F. Cootes

The University of Manchester

This work has been published in the proceedings on the SPIE Medical Imaging 2017 conference

Contribution of the thesis author: literature research, conception and design of the study, classification experiments using the code developed by Tim Cootes and adapted to use for the project, interpretation of results, drafting and revising of paper content, paper submission.

Keywords: Knee Osteoarthritis, Appearance Models, Combination of Views, Constrained Local Models

4.3.1 Abstract

Knee Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world. The effects of the disease have been studied using the shape and texture features of bones in Posterior-Anterior (PA) and Lateral radiographs separately. In this work we compare the utility of features from each view, and evaluate whether combining features from both is advantageous. We built a fully automated system to independently locate landmark points in both radiographic images using Random Forest Constrained Local Models. We extracted discriminative features from the two bony outlines using Appearance Models. The features were used to train Random Forest classifiers to solve three specific tasks: (i) OA classification, distinguishing patients with structural signs of OA from the others; (ii) predicting future onset of the disease and (iii) predicting which patients with no current pain will have a positive pain score later in a follow-up visit. Using a subset of the MOST dataset we show that the PA view has more discriminative features to classify and predict OA, while the lateral view contains features that achieve better performance in predicting pain, and that combining the features from both views gives a small improvement in accuracy of the classification compared to the individual views.

4.3.2 Introduction

Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world, the chance of developing the disease being particularly high in older people. It has been reported [36] that by 2030 around 20% of the American population will be above the age of 65, and that half of them (35 million of patients) will be at high risk of developing OA, requiring a large amount of public money [15] for treatments and surgery.

The most common signs of OA are: osteophytes, bony spurs that grow on the bones of the spine or around the joints, joint space narrowing (JSN) and calcium deposits. Painkillers and lifestyle changes are the only therapies currently available and eventually most patients have to undertake a total or partial joint arthoplasty.

OA is currently assessed from radiographs using the Kellgren and Lawrence (KL)[59] grades from 0 to 4, where 0 represents normality and 4 the most severe stage of OA. When a radiograph is taken clinicians assign a discrete KL grade based on features in the image. This is time consuming, subjective and there are shortages of suitably trained radiologists. There is an increasing need for reliable systems that can perform the grading automatically. Detecting knee OA and assessing its severity are crucial step for clinical decision making and a reliable prediction of the disease progression.

Current automated systems focus on the PA view, but research [68] indicates the lateral knee view adds information about pain, prediction of disease and other measures. It also allows better analysis of disease by capturing features missed in the PA angle.

We have developed fully automated systems to analyse the shape and texture of bones in both lateral and PA knees. The goal of this work is to compare which view gives the most informative features for studying OA, and to explore whether better results can be achieved by combining information from both views.

This work follows two different ones [82, 108] in automated OA diagnosis. In the first one the authors built the first automated system to classify OA using lateral knee radiographs. This system was built on the MOST [36] dataset, the only large dataset associated to a longitudinal study where at each visit both lateral and PA radiographs are acquired. This work showed that shape features extracted from lateral radiographs have promising discriminative capabilities, but lacked direct comparison with PA radiograph on the same dataset. The second work studied Posterior-Anterior (PA) knee radiographs from the OAI dataset [69] to retrieve shape and texture features. They used a RFCLM on a 74 points model and extracted features of tibial texture.

Other approaches as the ones of Shamir [97] and Anifah [6], where image processing techniques are applied to PA knee radiographs: in the former the authors extracted image content descriptors and image transforms to use as features in a Nearest Neighbor setting; the latter applied the unsupervised self organizing maps based on Gabor filter to classify the K-L grades. In the work of Jin [56] , the authors used medical infrared thermography of the PA view to extract features on which train a SVM classifier.

Our work aims at combining features from both lateral and PA knee radiographs. For both lateral and PA radiographs the method was the following: we manually annotated a few hundred images with a set of landmark points, obtaining a collection of discrete shapes, from which we built statistical shape and appearance models. We used a Random Forest Regression Voting Constrained Local Model (RFCLM) [19, 22, 72] to locate points in both single bones and combinations of bones. The detection of a ROI containing the joint was done using an object detector based on Random Forests (RF) to automatically initialise the RFCLM on each image.

Once an automated annotation was found, we extracted shape, texture and appearance parameters and combined them to solve three tasks related to OA: (i) OA classification, distinguishing patients with structural signs of OA from the others; (ii) predicting future onset of the disease and (iii) predicting which patients with no current pain will have a positive pain score later in a follow-up visit.

4.3.3 Methods

Our lateral knee model is made of four different sub-shapes: the patella (21 points), the lateral femoral condyle (24 points), the medial femoral condyle (25 points) and the tibia (32 points). We considered the lateral femur as the union of the two femoral condyles (49 points). The whole knee model is then made of 102 points (Figure 4.6). In this work we ignored the points of the patella because there was not a corresponding model for the PA view.

On the other hand, the PA model is made of two shapes: the femur and the tibia (37 points each, for a total of 74 points).

Appearance Model

The way we extracted features was by building an appearance model. Combined Appearance Models (CAM) [20] are an attempt at a better use of textural information

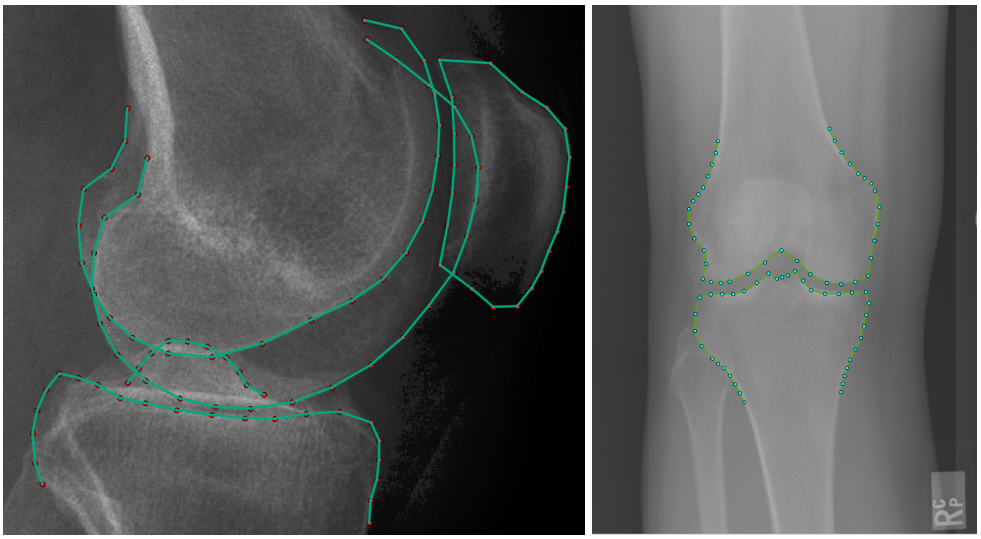


Figure 4.6: An example of the landmark points used to build the two shape models.

and they are based on a statistical model that uses shape as one of its components. In this way we achieve better representation power compared to a shape model and this could bring more robustness. Such a model incorporates non redundant information of the shape and the texture of the object of interest. They are built by first retrieving a statistical shape model of the knee.

A shape model is a mathematical object that represents each shape $x = (x_1, y_1, x_2, y_2, \dots)^T$ in the following way

$$x = T(\bar{x} + P_s b_s; t), \tag{4.3}$$

where $\bar{x}$ is a representation of the mean shape in a suitable reference frame, P_s is a matrix containing a set of modes of variation and T applies a global similarity transformation with parameters t.

The shape parameters b_s can be calculated from x using

$$b_s = P_s^T (T^{-1}(x; t) - \bar{x}). \tag{4.4}$$

In order to build an appearance model we start by assuming that we have built a shape model of variation, as described in Equation 4.3. Then we warp each of the training images to match the mean shape and we sample the texture information from the resulting objects. To minimise the effect of lighting variation, we normalise the samples using a scale factor α and an offset parameter β so that the mean of the pixel values is zero and the sum of their squares is unity. After applying PCA to the set of

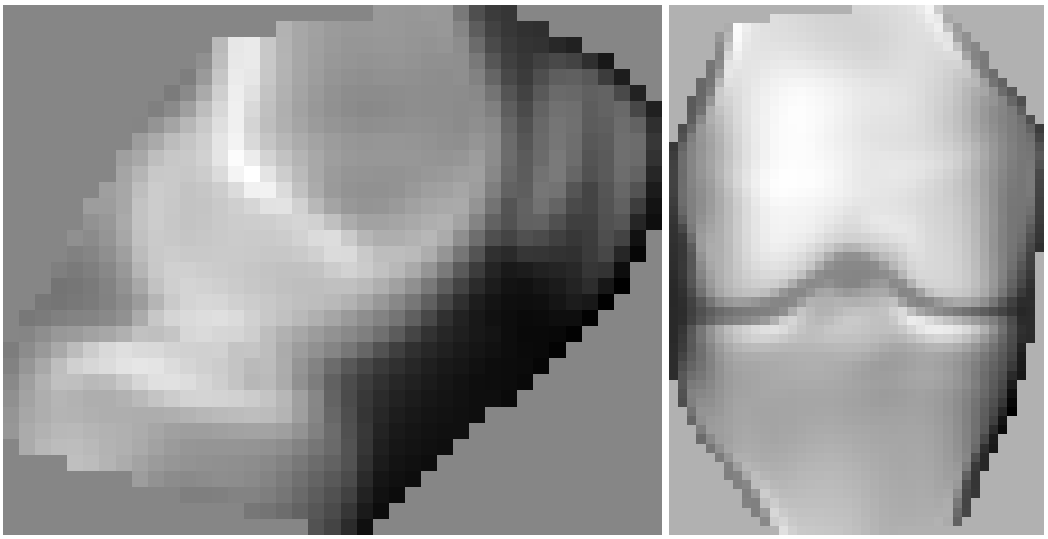


Figure 4.7: The mean appearance of the lateral (left) and Posterior-Anterior (right) models.

vectors obtained we end up with the linear model that follows

$$g = T(\bar{g} + P_g b_g; t), \tag{4.5}$$

where $\bar{g}$ is the mean grey-level vector, P_h is a matrix of eigenvectors, explaining the textural variation and b_g are texture parameters. The number of texture parameters was chosen to be constantly equal to 30.

An Appearance model, as shown is Figure 4.7, can be calculated by concatenating the two models and then via a further PCA. This is because shape and texture are often correlated.

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix}, \qquad b = Qc, \tag{4.6}$$

where W_s is a diagonal matrix of weights, the shape eigenvalues, for each shape parameter, b_g are the texture parameters and c are the appearance parameters with corresponding eigenvectors listed in Q.

Object Detection and Shape Model Matching

The first step in the segmentation of the bones was to build a global searcher able to find the individual bones within each image. Our implementation used a Hough Forest approach [41]. We defined a bounding box starting from a pair of landmark points and

then sampled from each image a set of patches with different displacements, angles and scales with respect to the location of the bone of interest. We then trained a Random Forest to learn the functional relation between the pixel intensities in the image patches and the corresponding displacements. This RF is scanned over a new image at multiple scales and orientations, voting for likely knee locations. The output of the global search is a bounding box with two reference points, from which we initialise each model (Figure 4.8).

Figure 4.8: An example of the bounding boxes found by the Random Forest bone detector.

In the second step we improve the fitting of the model, by applying a sequence of increasingly refined Constrained Local Models. The idea is to independently train a point detector per landmark point. Each model uses regression-voting trees to predict point displacements from patches of image texture and constrains the points using a shape model. The algorithm has been used previously to find hips and knees [73] from radiographs.

In the search phase we sample a set of patches around the current approximation of the point location. We feed those patches into the Random Forest, receiving a prediction per patch and tree for the location of the landmark point of interest. We

combine all the predictions in a voting image $V_i()$ for each point i. The shape model is used to regularise the result, finding the parameters b,t which maximise the total votes $Q(b,t) = \sum_{i=1}^{n} V_i(T(\bar{x}_i + P_i b_i; t))$. With regard to the lateral view, we used a different segmentation model for each of the three shapes involved. Each of those models involved a sequence of three increasingly refined CLM, with frame width equal to 50, 100 and 200 pixels. The frame widths associated to the PA radiographs were 50, 200 and 500, the first two aim to find the whole knee shape at once while the last one improving the fitting of tibia and femur separately.

The RFCLM is trained on manually annotated points placed around the shape outlines. We used 500 images from the MOST dataset to train the lateral RFCLM and 500 images from the OAI dataset for the PA model.

Table 4.9: Number of features per each shape and feature type.

	Lateral view			PA view		
	Shape	Texture	App.	Shape	Texture	App.
Femur	30	30	50	18	30	37
Tibia	20	30	42	19	30	40
Knee	44	30	55	35	30	51

Extraction and combination of the features

The approach above enables us to fully automatically segment the outlines of the bones in new images. We can find shape, texture and appearance parameters given the points from the statistical shape model.

In this work we explore how the parameters from different structures and views can be combined to most effectively classify the disease or pain status of the knee. To combine parameter vectors we simply concatenate them.

4.3.4 Results

The results shown in this section are obtained performing a 5-fold cross validation. For all the classification tasks introduced above we trained a Random Forest made of 25 trees. We show the performance of our technique by giving the areas under the

Receiver Operating Characteristic (ROC) curves (AUC) for each bone and view or combinations of features.

Data

The images were taken from the Multicentre Osteoarthritis Study (MOST) dataset. This is a longitudinal prospective study that collected data from 3026 participants with a 7-year follow-up. The data used in this work only considers data up to 30 months after baseline (second visit). Lateral and PA radiographs have been collected at each time-point for both knees. For the binary classification task the grades have been split into two groups: non-OA, KL (0,1), and the OA group, KL (2-4). KL grades and reported pain within the last 30 days are used as outcomes in the experiments, and different subsets of the data are used to solve the tasks of interest: (i) OA classification, using 4628 OA ($KL \geq 2$) and 6805 non-OA images; (ii) OA prediction, using 3234 baseline images with no OA ($KL \leq 1$) of those 272 develop OA within 30 months and 2962 do not develop OA; (iii) pain prediction, 845 knees with no pain at baseline with 478 later developing pain and 367 not developing pain.

In the following tables, the best results for the lateral view , the PA view and the combined view are highlighted.

OA Classification

The results of the automated diagnosis are shown in Table 4.10. In general, features extracted from the PA view perform better than those of the Lateral view. Furthermore, in various instances combining features of the two views achieved AUCs higher than both the individual models. However, the best overall AUC of the combined model is only as good as the best AUC of the PA model, in both cases corresponding to the appearance features of the full knee model. Finally, texture in all but one instance performes better that shape and appearance parameters in almost everytime perform better that both shape and texture alone.

OA Prediction

The accuracy of predicting future onset of structural OA is shown in the following Table 4.11. Again, in most cases the PA view has better performance than the lateral view. Moreover, features from the combined model are often more discriminative only by a small margin compared to the ones of the individual models. Similarly to the

Table 4.10: Binary OA classification. AUC for the two individual views and their concatenation.

	Lateral view			PA view			Lateral + PA views		
	Shape	Texture	App.	Shape	Texture	App.	Shape	Texture	App.
Femur	74 ± 0.1	80.9 ± 0.01	82.3 ± 0.01	78.7 ± 0.1	87.1 ± 0.1	87.3 ± 0.1	81 ± 0.1	87.9 ± 0.1	87.7 ± 0.1
Tibia	72.3 ± 0.01	79.2 ± 0.01	81 ± 0.2	78.4 ± 0.2	89 ± 0.1	88.9 ± 0.1	80.5 ± 0.1	89.1 ± 0.1	89 ± 0.01
Knee	81.3 ± 0.2	82.7 ± 0.01	$\mathbf{85.3 \pm 0.2}$	89.6 ± 0.1	89.4 ± 0.1	$\mathbf{90.4 \pm 0.1}$	89.7 ± 0.2	89.6 ± 0.1	$\mathbf{90.5 \pm 0.01}$

Table 4.11: Prediction of future onset of OA. AUC for the two individual views and their concatenation.

	Lateral view			PA view			Lateral + PA views		
	Shape	Texture	App.	Shape	Texture	App.	Shape	Texture	App.
Femur	57 ± 0.6	57.4 ± 0.8	$\mathbf{58 \pm 0.8}$	57.5 ± 0.1	62.4 ± 1.2	60.3 ± 0.8	58.9 ± 0.5	63.1 ± 1.1	60.2 ± 1.2
Tibia	54.1 ± 1.3	55.2 ± 1	53.1 ± 0.1	59.6 ± 0.2	$\mathbf{65.1 \pm 0.4}$	64.3 ± 1.2	59.7 ± 0.8	62.6 ± 1	60.4 ± 0.5
Knee	54.8 ± 0.8	56.3 ± 0.5	56.6 ± 1.1	60.3 ± 0.5	64 ± 0.6	63 ± 0.5	60.2 ± 0.8	$\mathbf{63.2 \pm 0.9}$	62.1 ± 0.6

previous table in most cases shape is less discriminative that texture, that in turn is less discriminative than appearance features. The appearance features of the femur have the highest AUC of lateral view features, while the texture features or the tibia have the highest AUC of the PA view and the highest overall.). While texture features perform consistently better than shape features, the performance often decreases when using appearance features.

Pain Prediction

Predicting future onset of knee joint pain is the most challenging of the three tasks. Pain scores are very subjective and finding patterns in the way pain develops with time has proven to be extremely difficult. Results of this task are shown in Table 4.12.

Unlike the previous experiments, features extracted from lateral knee radiographs show consistently better performance than the ones of the PA view. Furthermore, on average combining the two sets of features does not seem to increase the performance. Nevertheless, the best overall AUC 56.9 ± 0.1 is achieved by the combination of textures features of the femur. The appearance features of the femur again have the highest AUC of lateral view features, while the texture features or the femur have the highest AUC of the PA view.

Table 4.12: Prediction of future pain. AUC for the two individual views and their concatenation.

	Lateral view			PA view			Lateral + PA views		
	Shape	Texture	App.	Shape	Texture	App.	Shape	Texture	App.
Femur	53.6 ± 1.1	54.9 ± 1.7	$\mathbf{55.7 \pm 1.3}$	48.3 ± 0.8	$\mathbf{56.8 \pm 1.3}$	53.8 ± 0.1	52.6 ± 1.1	$\mathbf{56.9 \pm 0.1}$	55.1 ± 0.2
Tibia	54.1 ± 0.5	53.5 ± 1.2	55.4 ± 0.2	51.5 ± 1.2	52.9 ± 1.3	53.3 ± 1.2	55.4 ± 0.9	55.6 ± 1.6	54.8 ± 0.7
Knee	53.1 ± 2.6	54.7 ± 0.8	55.3 ± 1.8	48.5 ± 1.1	52.2 ± 0.8	52.4 ± 1.7	55.2 ± 0.2	54 ± 0.3	54.3 ± 1.3

4.3.5 Conclusions and Future Work

In this work we have shown the first attempt at combining shape, texture and appearance parameters of radiographs of the knee joint acquired from different views. Our experiments show that such concatenation leads to improved accuracy in various tasks, though often by a small margin. As far as we are aware this work is the first large scale direct comparison of the two views when studying OA and its future development.

The results show that the combination of the two views contais more discriminative features, but the magnitude of the improvement in performance is not large. Future work will involve the development of alternative ways of combining features from the two views. For example, we will be investigating ways of building a combined model for the appearance and apply redundancy reduction techniques. We also aim to design a deep learning architecture to merge the information coming from the radiographs.

It would also be interesting to use combined features to solve other OA related classification problems in the computer aided study of the disease such as: automated assessment of the severity of OA, detection of osteophytes severity and joint space narrowing.

ACKNOWLEDGMENTS

The research leading to this results has received funding from EPSRC Centre for Doctoral Training grant 1512584.

Table 4.13: Comparison between our method and previous approaches.

	Binary OA prediction	Binary Pain prediction
Ours	58.8	54
Shamir et al. [96]	55.7	54.6

4.4 Appendix

4.4.1 Hyper-parameter tuning

The number of shape and appearance parameters have been chosen in order to explain 95% of variance when applying PCA. The number of texture parameters have been chosen by looking at previous works, considering what was done for similar shapes in terms of number of points. This is due to the inherent redundancy of textural features. A more comprehensive tuning of the texture parameters was done prior to the indecisive tree experiments. This also explain the slight boost in performance between this chapter and the next one.

4.4.2 Comparisons with previous works

In this section we want to compare the performance of our lateral knee model with the method proposed by Shamir et al. [96]. We used the same features extracted in the paper and trained gradient boosted decision forests for classification. The lateral knee crop was obtained using our own knee joint detector.

We can see in Table 4.13 that our method outperforms the previous method. The distance is more significant on the disease prediction task. This is confirming the idea that hand crafted features perform worse than those learned from the data.

4.4.3 Summary of Deep learning results

Table 4.14 contains the results of the most well known approaches for knee OA classification using deep learning. Both of them use exclusively the PA view. The method from Tiulpin et al. [110] uses a siamese architecture to simultaneusly analyse two PA patches of the same knee. The network was trained on the MOST dataset and tested on OAI dataset. On the other hand, Anthony et al. [8] use a finetuned BLVC network

Table 4.14: Results on deep learning approaches in knee OA related learning tasks.

	Binary OA classification(AUC)	KL-grade classification(Accuracy %)
Tiulpin et al [110]	93	66.71
Anthony et al. [8]		57.9

to extract features from the penultimate fully connected layer. An SVM classifier then is trained to predict the appropriate KL grade. The results reported in Table 4.14 are reflecting evaluations on the OAI dataset.

Chapter 5

Improving the classification algorithm: Indecisive Forests

It is commonly believed that improved classification performance is achievable via either improving the quality of the data used or by refining the machine learning model. While the work that we presented earlier was mainly focused on the former, this part of the project was trying to address the latter issue.

This chapter describes the development of a novel classification technique based on random forests. The model can be potentially applied to any classification task, but we will show its effectiveness at solving two of the tasks associated to our project, namely binary structural knee OA classification and future onset of structural OA. The idea was to start from the Random Forest implementation used previously and look at ways to improving its generalization potential.

Indecisive Trees for Classification and Prediction of Knee Osteoarthritis

Luca Minciullo, Paul. A. Bromiley, David T. Felson and T.F. Cootes

The University of Manchester
ARUK Epidemiology Unit, The University of Manchester

This work has been published on the proceedings on the International Workshop of Machine Learning in Medical Imaging 2017 (MLMI)

Contribution of the thesis author: literature research, conception and design of the study, classification experiments using the code developed by Tim Cootes and adapted to use it for the project, interpretation of results, drafting and revising of paper content, paper submission.

Keywords: Random Forests, Decision Trees, Optimisation

Abstract

Random forests are widely used for classification and regression tasks in medical image analysis. Each tree in the forest contains decision nodes which choose whether a sample should be passed to one of two child nodes - a binary decision. We demonstrate that replacing this binary choice with something less decisive (some samples may go to both child nodes) can lead to improvements in performance for both individual trees and whole forests. Introducing a soft decision at each node means that a sample may end up at multiple leaves. The output of a tree should thus be a weighted sum of the individual leaf values - we show how the leaves can be optimised to give the best results. We also show how backpropagation can be used to optimise the parameters of the decision functions at each node. We show that the new method outperforms an equivalent random forest on a disease classification and prediction task.

5.1 Introduction

Decision trees, particularly in the form of Random Forests [12], are widely used in medical image analysis for tasks such as landmark location [19], segmentation [92] and classification [82, 83]. In most cases each tree uses hard decision nodes (a threshold on a feature response derived from the input), in which a sample is channelled to either the left or right child node. Thus one input sample ends up at exactly one leaf, which holds the output for the tree.

A natural extension is to replace this binary decision with something softer, so that a sample can go down both branches, but with different weights or probabilities depending on the feature response at the node. An early example of this approach were "Fuzzy Decision Trees" [105] in which a sigmoidal function is used to assign a weight to be passed down each child branch. The approach was extended in [64], where a forest of such trees was integrated into a deep network allowing end-to-end training.

However, one problem with using a sigmoidal transfer function is that every input effectively ends up at every leaf of the tree with a non-zero weight - though at most leaves the weight may be very close to zero. This is potentially very inefficient for deep trees.

In this paper we introduce trees in which only samples near the decision boundary are propagated to both children - most samples only go to one child. This is equivalent to using a simple sloped step function to compute the weights. Each input is then propagated to a relatively small number of leaves. This allows us to use deep trees and retain most of the efficiency of binary decision trees. In the following we describe the approach in detail, including a greedy method for training a tree. Like the Fuzzy trees, we can optimise both the values stored at the leaf nodes and the parameters of the transfer functions using either closed form or gradient descent approaches, leading to better performance than that from the greedy training. We demonstrate that replacing random forests with these more indecisive trees leads to improvements in overall performance on a classification task. We show the improvement in performance of our methodology on Osteoarthritis (OA) classification and prediction tasks. OA is the most common form of arthritis, affecting millions of people around the world, the chance of developing the disease being particularly high in older people. The most common signs of OA are: osteophytes, bony spurs that grow on the bones of the spine or around the joints, joint space narrowing (JSN) and calcium deposits. We train our new trees to use features which measure the shape and appearance of the knee in radiographs to classify OA status and predict who is at risk of developing the disease.

5.2 Background

Random Forests [12] are a very successful machine learning ensemble model, where each of the sub-models is a binary decision tree. The randomness comes from two main features: first, each of the decision trees is trained on a different sample of the original dataset obtaining by generating multiple bootstrap samples; second, the optimal split is found by considering only a random subset of the features appearing in the data.

Ren et al. [92] showed how the leaves of a forest could be mutually optimised to give better performance than that of a forest with independent trees. Fuzzy decision trees, which can be optimised by a backpropagation-like algorithm, were introduced in [105]. They proposed training a tree in the normal way, then replacing the binary decision threshold with a sigmoidal function to indicate branch membership, the parameters of which can then be optimised.

Kontschieder et al. [64] extended this idea to full decision forests, using a sigmoidal decision function. They too used a stochastic gradient descent approach to optimise the parameters of the decision nodes and the leaves. The decisions at each node are based on the output of one node of a deep convolutional network, making the entire system amenable to end-to-end training.

When using a sigmoidal function for branch membership, every sample ends up being propagated to every leaf of the tree, even though at some leaves the membership value may be very small. This may lead to inefficiencies for deep trees. To overcome this we use a ramp function for the membership propagation:

$$\pi(\mathbf{x};t_0,t_1) = \begin{cases} 0 & \text{if} \quad f(\mathbf{x}) \leq t_0 \\ \frac{(f(\mathbf{x})-t_0)}{(t_1-t_0)} & \text{if} \quad t_0 < f(\mathbf{x}) < t_1 \\ 1 & \text{if} \quad f(\mathbf{x}) \geq t_1 \end{cases} \tag{5.1}$$

where $f(\mathbf{x})$ is a feature derived from the input $\mathbf{x}$ and $t_0 < t_1$ are two thresholds defining the ramp function. Thus if the membership for either branch is zero, we do not need to propagate down that branch. During training we choose the thresholds so that a given proportion of the training samples are in the ambiguous region (see below).

5.2.1 Evaluating the result from a tree

A tree is a collection of decision nodes and leaf nodes. Each decision node has two child nodes (left and right), a function which computes a scalar feature value from the

input $f(\mathbf{x})$ and two threshold values defining the transfer function, t_0, t_1. Each leaf node contains an output value.

When an input, $\mathbf{x}$, is evaluated with the tree, the output is a set of leaf values and associated weights, $\mathcal{S} = \{(\mathbf{v}_i, w_i)\}$. Starting at the root node, we propagate an input through the nodes, exploring only the branches with non-zero weights. Each node either adds its value (if it is a leaf) to a set of outputs, or it propagates the input and weight to one or both of its child nodes. This can be computed with a recursive function, starting at the root node with a unit weight: $\mathcal{S}$ =EVALUATE(root,$(\mathbf{x}, 1.0), \{\}$). The function is defined as follows:

1: **function** EVALUATE(node,$(\mathbf{x}, w), \mathcal{S}$)
2: **if** $node.isLeaf$ **then**
3: $\mathcal{S} \leftarrow \{\mathcal{S}, (node.value, w)\}$
4: **else**
5: $\mu = \pi(node.f(\mathbf{x}), node.t_0, node.t_1)$
6: $w_L = (1-\mu)w$
7: $w_R = \mu w$
8: **if** $w_L < w_R$ **then**
9: if $(w_L < w_t)$ then $w_L \leftarrow 0, w_R \leftarrow w$
10: **else**
11: if $(w_R < w_t)$ then $w_R \leftarrow 0, w_L \leftarrow w$
12: **end if**
13: if $(w_L > 0)$ $\mathcal{S}$ ←EVALUATE(node.leftChild,$(\mathbf{x}, w_L), \mathcal{S}$)
14: if $(w_R > 0)$ $\mathcal{S}$ ←EVALUATE(node.rightChild,$(\mathbf{x}, w_R), \mathcal{S}$)
15: **end if**
16: **return** $\mathcal{S}$
17: **end function**

The tests in lines 8-12 allow a threshold (w_t) on the smallest allowable weight to be enforced. If a split would cause the weight propagated to one child node to fall below the threshold, then that child node is ignored and all the weight is passed to the other child. Setting $w_t > 0$ ensures that no leaf is reached with a weight lower than w_t, and focuses processing on the branches with higher weights. It thus also limits the maximum number of leaves that can be returned to w_t^{-1}. The output of the tree can then be computed from $\mathcal{S}$ as the weighted sum of the leaf outputs, $\mathbf{v} = \sum_i w_i \mathbf{v}_i$.

5.3 Training and Optimising Indecisive Trees

In a similar way to training a normal decision tree, an indecisive tree is trained using a greedy recursive algorithm in which each node finds a feature and threshold to split the data arriving at it so as to minimise a cost function. During training a sample consists of a triplet, $(\mathbf{x},\mathbf{y},w)$, containing the input vector, the target output and a weight. To train a node, we consider the set of n samples $\mathcal{D}$ arriving from the parent node. To evaluate a particular choice of feature,$f(\mathbf{x})$, and thresholds t_0,t_1, we compute the sets of data $\mathcal{D}_L$ and $\mathcal{D}_R$ that would be propagated to the child nodes, and the cost function

$$C(f,t_0,t_1) = C(\mathcal{D}_L)+C(\mathcal{D}_R) \tag{5.2}$$

The cost $C(\mathcal{D})$ depends on the task (classification or regression). For instance, for regression, it can be the sum of square differences. When used as part of a random forest, a random selection of features and possible thresholds is evaluated, and those giving the lowest cost retained.

Since finding the optimal pair of thresholds can be computationally expensive, we use the following approach. For each input $(\mathbf{x}_i,\mathbf{y}_i,w_i)$ we compute the feature value $f_i = f(\mathbf{x}_i)$, then rank the samples using this value. Let $(\mathbf{x}_j,\mathbf{y}_j,w_j)$ be the j^{th} sample in this ranked list. By computing running sums through this ranked data we can efficiently locate the index, k, for the hard split leading to the lowest total cost (all samples $j \leq k$ are sent to one child, all $j > k$ to the other). We then introduce an ambiguous region to include a proportion of approximately $r \in [0,1]$ of the samples by setting $j0 = \max(1,k-0.5rn)$, $j1 = \min(n,k+0.5rn)$, and selecting $t_0 = f_{j0}$, $t_1 = f_{j1}$.

Since those samples in the ambiguous region will go to both children, the total number of samples propagated from nodes at depth d will be approximately $n_0.(1+r)^d$, where n_0 is the original number of training examples, though it should be remembered that the total weight for each of the original samples will always sum to unity. In order to avoid propagating large numbers of samples with small weights, we use the same technique as described above (Sec. 5.2.1) - if a sample weight would fall below w_t when propagated to one child node, we ignore that child and propagate all the sample weight to the other child. Decision nodes are added in a recursive manner until a suitable stopping condition (a maximum depth, minimum number of samples or measure of spread) is reached. The values at the leaf nodes can then be set to the weighted mean

of the samples reaching that node, for instance for regression, the value

$$\mathbf{t} = \left(\sum w_i \mathbf{y}_i\right) / \left(\sum w_i\right) \tag{5.3}$$

5.3.1 Optimising the leaf values

A tree with vector output can be expressed as a function of input $\mathbf{x}$

$$\mathbf{y} = \mathbf{V}\mathbf{w}(\mathbf{x}) \tag{5.4}$$

where $\mathbf{V}$ is a matrix whose columns are all the leaf vectors, and $\mathbf{w}(\mathbf{x})$ is the sparse vector of weights returned by the tree, which selects the leaves to which $\mathbf{x}$ is propagated. Thus the outputs corresponding to the training inputs can be expressed as

$$\mathbf{Y} = \mathbf{V}\mathbf{W} \tag{5.5}$$

where $\mathbf{Y} = (\mathbf{y}_1|...|\mathbf{y}_n)$ and $\mathbf{W} = (\mathbf{w}(\mathbf{x}_1)|...|\mathbf{w}(\mathbf{x}_n))$ is a sparse matrix.

For regression, as in [92], the leaf values can be found by minimising

$$Q(\mathbf{V}) = ||\mathbf{V}\mathbf{W} - \mathbf{Y}||_2 + \alpha||\mathbf{V}||_2 \tag{5.6}$$

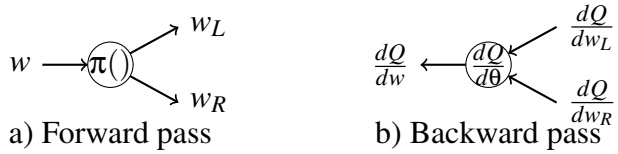
where α is an optional ridge regression regularisation function. Since $\mathbf{W}$ is sparse the solution can be found efficiently with conjugate gradient descent.

5.3.2 Optimising the decision nodes

If the leaf values are fixed, each decision node only affects the final output through the way it changes the weights on the samples passing through it. As in [105, 64] we can use a gradient descent-based backpropagation algorithm to optimise the parameters. However, in our case, since each sample only passes through a small subset of nodes, this can be significantly more efficient - we only have to compute values at the nodes visited.

The cost function to be minimised is of the form

$$Q_T(\theta; \{(\mathbf{x}_i, \mathbf{y}_i)\}) = \sum_i Q(\mathbf{V}\mathbf{w}(\mathbf{x}_i, \theta), \mathbf{y}_i) \tag{5.7}$$



a) Forward pass b) Backward pass

Figure 5.1: During the forward pass (root to leaves), weights are calculated. During the backward pass (leaves to root), gradients are calculated.

where θ are the parameters affecting the weights and $Q(\mathbf{t},\mathbf{y})$ is the cost function comparing the output of the tree, $\mathbf{t} = \mathbf{V}\mathbf{w}(\mathbf{x},\theta)$ with the target output $\mathbf{y}$.

Gradient at leaf nodes: The contribution to the output from a single leaf node is given by $w\mathbf{v}$, where w is the weight of the sample arriving at the leaf. For one leaf,

$$\frac{dQ}{dw} = \frac{dQ}{d\mathbf{t}}\frac{d\mathbf{t}}{dw} = \mathbf{v}^T\frac{dQ}{d\mathbf{t}} \tag{5.8}$$

Gradient at decision nodes: At a decision node, the weights passed to the output nodes are given by

$$\begin{pmatrix} w_L \\ w_R \end{pmatrix} = w\begin{pmatrix} 1-\pi(f,t_0,t_1) \\ \pi(f,t_0,t_1) \end{pmatrix} \tag{5.9}$$

If the parameters at the decision nodes are θ, then

$$\begin{aligned} \frac{dQ}{d\theta} &= \frac{dw_R}{d\theta}\frac{dQ}{dw_R} + \frac{dw_L}{d\theta}\frac{dQ}{dw_L} \\ &= w\frac{d\pi}{d\theta}\frac{dQ}{dw_R} - w\frac{d\pi}{d\theta}\frac{dQ}{dw_L} \\ &= w\frac{d\pi}{d\theta}\left(\frac{dQ}{dw_R} - \frac{dQ}{dw_L}\right) \end{aligned} \tag{5.10}$$

Similarly

$$\frac{dQ}{dw} = \pi(\theta)\frac{dQ}{dw_R} + (1-\pi(\theta))\frac{dQ}{dw_L} \tag{5.11}$$

During the backward pass we use (5.10) to compute the gradient w.r.t. the thresholds t_0 and t_1. In the experiments below we keep the features fixed, but it would also be possible to compute gradients of any parameters of the features.

We use the following algorithm to update the parameters of each node (t_0,t_1):

1: **function** UPDATENODES($\mathcal{X} = \{(\mathbf{x}_i,\mathbf{y}_i)\}$)
2: **for all** $\mathbf{x}$ in $\mathcal{X}$ **do**
3: Feed $\mathbf{x}$ forward through tree to calculate weights
4: Visit each node in reverse depth order - compute gradients
5: Update estimate of mean gradient over batch

6: **end for**
7: Update parameters using mean gradient
8: **end function**

In the following the parameter update is made using a momentum term, but something more sophisticated could be used.

5.4 Experiments

Here we focus on two classification tasks related to knee osteoarthritis. The features we used were shape, texture and appearance parameters extracted from lateral knee radiographic images (Figure 4.1). Those features were obtained by first building a statistical appearance model [20] of the knee. This model is a PCA based combination of statistical shape and texture models and it was built on fully automated annotation found using a 3-stage Constrained Local Model [22].

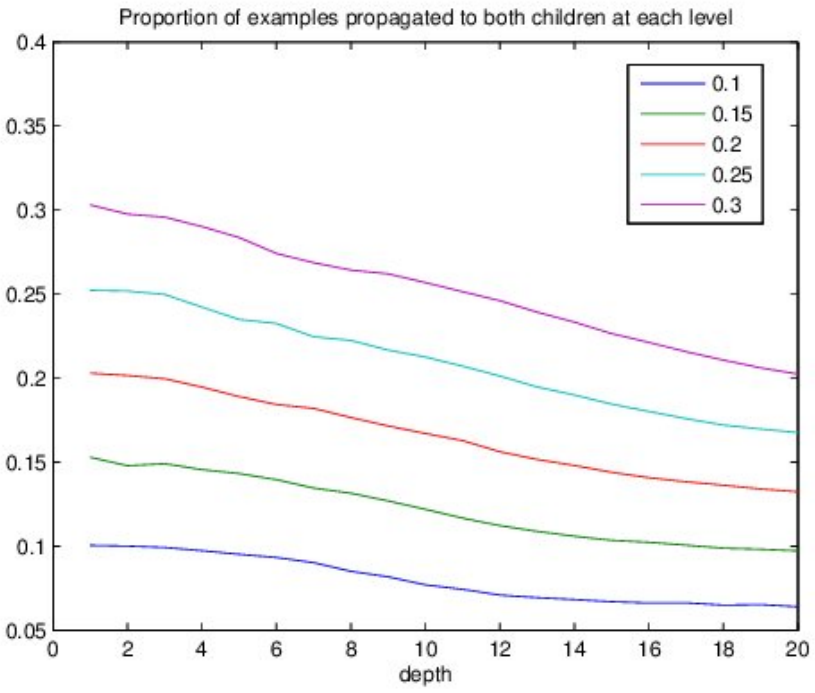


Figure 5.2: Proportion of examples within the indecisive window at each level for different choices for window width.

Data. The images were taken from the Multicentre Osteoarthritis Study (MOST) dataset [36]. MOST is a longitudinal prospective study that collected data from 3026

	OA Classification	OA Prediction	Timings (OA Prediction)
Baseline Forest	86.35 ± 0.99	59.03 ± 1.20	9.3s
IF	87.61 ± 0.94	$\mathbf{61.11 \pm 1.79}$	94.9s
OIF	$\mathbf{88.15 \pm 0.91}$	59.11 ± 2.01	between 3s and 2 minutes

Table 5.1: AUC for the two knee OA tasks: comparing a standard Random Forest with both an Indecisive Forest(IF) and an Optimised Indecisive Forest(OIF).

participants with a 7-year follow-up. Lateral radiographs have been collected at each time-point for both knees and a KL(Kellgren-Lawrence) grade assessing the severity of the disease was assigned to each knee. For our binary classification tasks the grades have been split into two groups: non-OA, KL (0,1), and the OA group, KL(2-4). The first task is an OA classification task, where the goal is to distinguish patients from the two groups and uses 8606 OA ($KL \geq 2$) and 10604 non-OA images. In the second task we consider 3478 baseline images with no OA ($KL \leq 1$) and aim to discriminate those that will develop OA within 84 months from those who will not.

Knee OA classification tasks. We compare the performance of our Indecisive Forest (IF) with a standard Random Forest (RF) in 5-fold CV experiments. A parameter sweep suggested that a good choice for the parameter responsible for the width of indecision window $r = 0.3$. In addition we applied the tuning algorithm described above to optimise the IF, and evaluated the performance. We report the area under the ROC curve to evaluate each of the models in Table 5.1. This shows that for both classification tasks the IF performs better than a standard Random Forest, with an improvement of at least 2% for both Classification and Prediction. The optimisation improves the results for the OA classification task, while the OA prediction performance does not change significantly. Our results on both tasks achieve the state of the art on the MOST dataset using only lateral knee radiographs (compared to [83]).

Figure 5.2 (Right) shows that the proportion of examples within the indecisive region increases when the window width increases and decreases linearly as examples go deeper in the trees.

Timings. The average time to train a standard tree on the prediction dataset was 9.3s, compared to 94.9s for each indecisive tree. The average tree optimisation time depended on the dataset and the parameter choice, it ranges from 3s to 2 minutes. There is only a small difference in performance when applying the trees.

5.5 Discussion and Conclusions

We have presented an improvement on the standard random forest that uses a ramp function with an ambiguous region to train and test random forests. We showed improved performance, compared to a standard Random Forest, on two OA related classification tasks. The combined leaf and node optimisation further improved the results on one of the tasks. The indecisive forests take longer to train and optimise. Pilot experiments on regression tasks have shown small but encouraging improvements, something that we will explore in future work.

5.6 Acknowledgments

The research leading to this results has received funding from EPSRC Centre for Doctoral Training grant 1512584. This publication presents independent research supported by the Health Innovation Challenge Fund (grant no. HICF-R7-414/WT100936), a parallel funding partnership between the Department of Health and Wellcome Trust, and by the NIHR Invention for Innovation (i4i) programme (grant no. II-LB_0216-20009). The views expressed are those of the authors and not necessarily those of the NHS, NIHR, the Department of Health or Wellcome Trust.

5.7 Appendix

5.7.1 Details of the hardware used

The experiments reported in this chapter were obtained on a 3.10GHz Intel core x4 CPU machine with 16 GB RAM.

5.7.2 Comparison with alternative methodologies

In this section we want to compare the indecisive tree approach and its optimisation with the alternative methods cited in the paper. The code implementing the Deep Neural Decision Forest [64] is publicly available and after writing a parsing for our dataset we trained and tested it with the same 5 fold CV setup. Table 5.2 report the results of this comparison. Our indecisive forest performs better on the prediction task. On the OA Classification task the DNDF performs better than than the Indecisive Forest but there does not seem to be significant difference between the two methods after the

	OA Classification	OA Prediction
Deep Neural Decision Forest [64]	$\mathbf{88.28 \pm 1.20}$	56.22 ± 5.04
IF	87.61 ± 0.94	$\mathbf{61.11 \pm 1.79}$
OIF	88.15 ± 0.91	59.11 ± 2.01
XGBoost	86.43 ± 0.18	53.16 ± 1.54

Table 5.2: Direct comparison between the indecisive forest and the Deep Neural Decision Forest [64].

optimisation step. As an added comparison we compared with the Gradient Boosted trees [58], using the open source python library XGboost. This method performs worse than both the Indecisive Forest and its optimisation.

Chapter 6

Correlating Symptomatic and Radiographic Osteoarthritis

After mainly focusing on ways to automatically diagnose the disease and improving the performance of machine learning classifiers for radiographic OA, the last part of the project was around the relationship between radiographic features and symptoms experienced by the participants of an OA related study. Being able to show evidence of this relationship would allow for further understanding of what is causing those symptoms and consequently investigate ways of reducing them.

The paper in this chapter deals with this problem. We attempt to answer the following questions:

- What are the main individual sources of pain?
- Measure the discriminative ability of manual radiographic measurements at predicting pain
- Does adding demographic information improve performance?
- Does removing people with widespread pain improve performance?
- Is consistent pain more correlated with radiographic features?
- Can we extract features automatically to achieve the same performance?

Our work is the first we are aware of using manual grades of lateral view radiographs and reports the highest AUC of a machine learning classifier when predicting pain from radiographic measurements.

Comparing Image Analysis Approaches vs Expert Readers: the Relation of Knee Radiograph Features to Knee Pain

Luca Minciullo, Matthew Parkes, David T. Felson and T.F. Cootes

The University of Manchester
ARUK Epidemiology Unit, The University of Manchester

A shortened version of this paper has been submitted to the Annals of Rheumatic Diseases

Contribution of the thesis author: literature research, conception and design of the study, classification experiments using the code developed by Tim Cootes and adapted to use it for the project, interpretation of results, drafting and revising of paper content, paper submission.

Keywords: Symptomatic Osteoarthritis, Machine Learning, Radiographic features, Frequent Knee Pain

Abstract

Objectives: The relationship between radiographic evidence of osteoarthritis and knee pain has been weak, but this may be because features that best discriminate knees with pain have not been included in analyses. We tested the correlation between knee pain and radiographic features of osteoarthritis taking into account both features automatically extracted from radiographs and manual scores. Methods: Using the baseline visit of the Multicenter Osteoarthiritis Study, we tested how well x-ray features discriminated those with frequent knee pain (one question at one time) or consistent frequent knee pain (3 questions at 3 times) from those without it. We used posteroanterior and lateral radiographs and examined grades assigned by readers as well as imaging features such as shape and texture. Random forest classifiers were used to predict whether participants had knee pain or not. We used the area under the ROC curve (AUC) to quantify how well radiographs classified those with and without pain. Results: X-rays were better at classifying those with pain using 3 questions compared to one. When we used all radiographic features scored by readers, the AUC was 70.4. Using the best model from automated image analyses or a combination of these and manual grades, no significant improvement in performance over manual grading alone was found. Conclusions: X-ray changes of OA are more strongly associated with repeated reports of knee pain than pain reported once. In addition, a fully automated image analysis technique that assessed features not scored on x-ray performed no better than manual grading of features.

6.1 Introduction

One of the main points of interest in research in osteoarthiritis(OA) is the investigation of pain and its relation with structural changes from radiographic images. Despite considerable effort the existence of pain has not found to be strongly correlated with radiographic OA [43, 48, 98, 107, 67]. In general, only about half the people with knee pain in population studies have radiographic OA and likewise, only around half of the knees with radiographic OA in such studies are afflicted with knee pain [10, 48]. Firstly, this poor agreement between radiographs and pain may be because the global measures of radiographic disease that are used in these studies, such as Kellgren and Lawrence grades, are insensitive to subtle or specific features that are better correlated with pain than global scores. Secondly, these studies have generally been limited to uniplanar radiographs and therefore may miss features that are correlated with the presence of pain. Thirdly, some individuals may have knee pain as part of a syndrome of widespread pain and do not have OA. Lastly, knee pain is often transient and radiographic disease may be more likely in persons in whom it is more consistently reported.

Previous studies involve the investigation of correlation between individual structural features such as osteophytes and joint space narrowing (JSN) [10, 48, 67] and pain. Even those explorations have not found a strong correlation of pain with radiographic features. Felson and colleagues [35] gave an alternative definition of OA based on a combination of structural features and showed a modestly improved correlation with pain. Minciullo et al. [83] used Constrained Local Models (CLM) to find landmark points for the knee joint in both Lateral and PA radiographs and extracted features related to the shape, texture and their combination to predict future onset of knee pain, showing a weak correlation with structural features and suggesting that the lateral view contains features that are significantly more discriminative at predicting future knee pain compared to the PA view. Galvan-Tejada et al. [43] used radiographs from the OAI to prove that osteophytes are early predictors of joint pain, while joint space reduction is not clearly associated with future joint pain.

The objective of our work was to determine the correlation between knee pain and various sets of radiographic features of OA obtained at the time of the pain report, using both features automatically extracted from knee radiographs and manual grades assigned by clinicians. To do so we built random forest classifiers using a large collection of features, both extracted from radiographs using state of the art landmark point

detectors and manual grades. Unlike most previous works we used both posteroanterior and lateral radiographs. We also tried combining structural features with image independent features such as age and BMI, which are known to increase risk of developing OA [107]. Furthermore, we tried to exclude from the study people who were experiencing widespread pain, under the assumption that such pain may not be due to OA.

6.2 Methods

Images were taken from the Multicentre Osteoarthritis Study (MOST) dataset [36]. Bilateral PA standing flexed and unilateral weight bearing, flexed lateral radiographs were obtained at baseline for both knees. At baseline subjects were asked three times whether they had knee pain, aching or stiffness on most of the last 30 days. Firstly, a telephone screening (TScreen) done roughly 2 weeks before the clinic visit was performed to check eligibility criteria. Secondly, before the visit, participants filled a Self Assessed Questionnaire (SAQ) at home. Lastly, an interview was done as part of the clinical visit (Clinic). We used the telephone screening, the Clinic and SAQ variable together to create a measure we called 'Consistent Pain'. By consistent pain we meant selecting participants that gave the same binary score at all three time points. In our experiments we only considered data from the baseline visit and only the right knee in order to remove the effect of considering structurally non independent information (multiple visits of the same participants or the two knees).

The radiographic grades used in our work were assigned by central readers as part of the MOST study protocol. Two main types of features have been used in our experimental setup. The first ones were manual grades for features of OA assigned by readers during the MOST study. We used scores for all the features that were read on both the PA and lateral views. These readers also provided a Kellgren and Lawrence grade for each knee.

The second set of features was automatically extracted using Constrained Local Models (CLM) to find landmark points in radiographs. This latter model has been successfully used in medical imaging in numerous occasions and with a large variety of radiographic images [73, 82, 83]. Our knee model for the lateral radiograph was made of four different sub-shapes: the patella (21 points), the lateral femoral condyle (24 points), the medial femoral condyle (25 points) and the tibia (32 points). We considered the femur as the union of the two femoral condyles (49 points) (Figure 6.1,

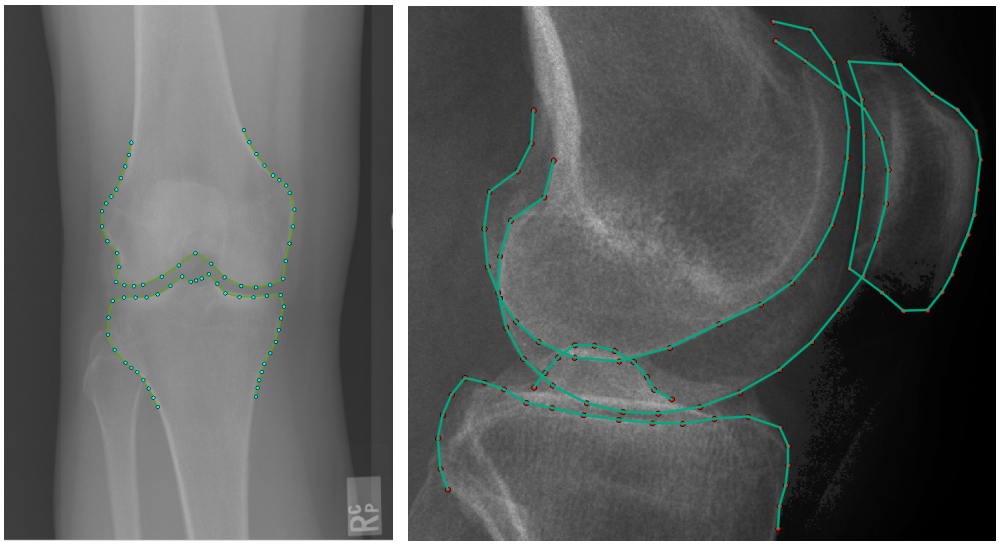


Figure 6.1: The PA (left) and Lateral (right) knee models.

right). The PA model was made of two shapes: the femur and the tibia (37 points each, a total of 74 points) (Figure 6.1, left).

6.2.1 Appearance Model

We extracted features by building an appearance model. Combined Appearance Models (CAM) [20] are an attempt to better use textural information and are based on a statistical model that uses shape as one of its components. Such a model incorporates non redundant information of the shape and the texture of the object of interest. In this way we achieved better representation power compared to a shape model and this could bring more robustness.

A shape model is a mathematical object that represents each shape $x = (x1, y1, x2, y2, \dots)^T$ in the following way

$$z = T(\bar{x} + P_s b_s; t) \tag{6.1}$$

where $\bar{x}$ is a representation of the mean shape in a suitable reference frame, P_s is a matrix containing a set of modes of variation and $T(\cdot;t)$ applies a global similarity transformation with parameters t.

The shape parameters b_s can be calculated from x using

$$b_s = {P_s}^T (T^{-1}(x,t) - \bar{x}), \tag{6.2}$$

In order to build an appearance model we started by assuming that we have built a shape model of variation, as described in Equation 6.1. Then we warped each of the training images to match the mean shape and we sampled the texture information from the resulting objects. To minimise the effect of lighting variation, we normalised the samples using a scale factor α and an offset parameter β so that the mean of the pixel values was zero and the sum of their squares was unity. After applying Principal Component Analysis (PCA) to the set of vectors obtained, we ended up with the following linear model:

$$g = T(\bar{g} + P_g b_g; t) \tag{6.3}$$

where $\bar{g}$ is the mean grey-level vector, P_g is a matrix of eigenvectors, explaining the textural variation and b_g are texture parameters. The number of texture parameters was chosen to retain 90% of the variance obtaining 30 PA features and 63 lateral features.

An appearance model can be calculated by concatenating the two models and then via a further PCA

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix}, \qquad b = Qc, \tag{6.4}$$

where W_s is a diagonal matrix of weights (the shape eigenvalues for each shape parameter), b is the weighted concatenation of shape and texture parameters and c are the appearance parameters with corresponding eigenvectors listed in Q.

6.2.2 Object Detection and Shape Model Matching

We developed an automatic system to locate the outlines of the bones in both radiographic views. It first finds the position of a bounding box around the joint, then refines this with a shape model matching algorithm – for full details see [41, 82, 19]

6.2.3 Analysis Approach

First, we tested the relation of individual features and KL grades with the presence of pain. All the experiments were performed training and testing a random forest classifier with 40 trees, running a 5-fold cross validation with 5 repeats and we used the area under this ROC curve to determine the relation of knee pain with radiographc features. We report the standard deviation of the performance evaluated using the AUC over 5 repetitions.

We compared a single question for frequent knee pain (obtained at the clinic visit) vs. the same question administered three times in relation to the baseline MOST visit. For the latter approach, we compared persons who consistently reported knee pain to those who did not report knee pain at any of the 3 time points.

The subsequent analyses tested whether automated image analysis generated a higher area than did a combination of manually scored features. In addition, we tested whether a combination of information provided by image analysis and manual grading improved upon the ROC curve area compared with manual grading alone.

χ^2 tests were used to assess the difference in AUC between the manual scoring (as the gold standard), adding BMI and sex and the best fully automated model. The p-value of 0.05 or below was selected to indicate that the ROC curve differed from the gold standard significantly.

In additional analyses, we limited our sample to knees without radiographic OA signs to see if there were imaging features that might help identify knees with pain. By without radiographic OA signs we mean participants with both radiographs graded with all OARSI grades equal to 0.

6.3 Results

6.3.1 Testing individual radiographic features

There are 36 individual radiographic features scored from the PA and lateral radiographs (listed in Table 6.1). We tested how well each grade could classify the pain score collected once at the clinic visit. For each we measured the AUC when using the grade as a feature in a classifier. We observe that KL grade, osteophytes, joint space narrowing and sclerosis were the most discriminative with the KL grade achieving the best result. On the other hand chondrocalcinosis, cyst, attrition and ossification of the patella-tendon performed no better than chance. While some of these results were expected, bone attrition (as MRI feature) was previously found to be associated with OA pain [67]. Furthermore, grades associated with the medial compartment were consistently better at classifying frequent pain.

6.3.2 Using shape, texture and appearance parameters

We compared the previous method of using features manually assigned by experts with features automatically extracted by the CLM Model using PA and lateral views. We

Table 6.1: Testing each radiographic feature individually using the pain score reported during the visit (Clinic).

Variable	AUC (%)
Chondrocalcinosis (OARSI grades 0-1) PF joint on LA view	50 ± 0.3
Osteophytes(OARSI grades 0-3) femur anterior PF joint on LA view	58.3 ± 0.2
Osteophytes(OARSI grades 0-3) femur posterior PF joint on LA view	60 ± 0.5
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on LA view	55.2 ± 0.2
Joint space narrowing (OARSI grades 0-3) medial TF compartment on LAT view	59.1 ± 0.3
Effusion (OARSI grades 0-1) PF joint on LA view	56 ± 0.3
Kellgren & Lawrence (grades 0-4) on PA view	$\mathbf{64.8 \pm 0.1}$
Chondrocalcinosis (OARSI grades 0-1) lateral TF compartment on PA view	50.4 ± 0.3
Cyst (OARSI grades 0-3) femur lateral TF compartment on PA view	50.6 ± 0.3
Osteophytes (OARSI grades 0-3) femur lateral TF compartment on PA view	60.2 ± 0.2
Sclerosis (OARSI grades 0-3) femur lateral TF compartment on PA view	54.3 ± 0.3
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on PA view	54.9 ± 0.3
Attrition (OARSI grades 0-1) lateral TF compartment on PA view	50.6 ± 0.2
Cyst (OARSI grades 0-3) tibia lateral TF compartment on PA view	50.6 ± 0.2
Osteophytes (OARSI grades 0-3) tibia lateral TF compartment on PA view	60 ± 0.2
Sclerosis (OARSI grades 0-3) tibia lateral TF compartment on PA view	54.2 ± 0.3
Chondrocalcinosis (OARSI grades 0-1) medial TF compartment on PA view	50.7 ± 0.1
Cyst (OARSI grades 0-3) femur medial TF compartment on PA view	50.8 ± 0.3
Osteophytes (OARSI grades 0-3) femur medial TF compartment on PA view	61 ± 0.3
Sclerosis (OARSI grades 0-3) femur medial TF compartment on PA view	57.7 ± 0.2
Joint space narrowing (OARSI grades 0-3) medial TF compartment on PA view	57.7 ± 0.2
Attrition (OARSI grades 0-1) medial TF compartment on PA view	52.1 ± 0.2
Cyst (OARSI grades 0-3) tibia medial TF compartment on PA view	51.5 ± 0.3
Osteophytes (OARSI grades 0-3) tibia medial TF compartment on PA view	59.7 ± 0.2
Sclerosis (OARSI grades 0-3) tibia medial TF compartment on PA view	58.3 ± 0.4
Ossification (OARSI grades 0-3) patella tendon lower PF joint on LA view	49.5 ± 0.1
Ossification (OARSI grades 0-3) patella tendon upper PF joint on LA view	50 ± 0.6
Ossified loose body (OARSI grades 0-1) femur posterior PF joint on LA view	52.2 ± 0.3
Ossification of QF insertion (OARSI grades 0-3) PF joint on LA view	51 ± 0.3
Cyst (OARSI grades 0-3) PF joint on LA view	51 ± 0.2
Joint space narrowing (OARSI grades 0-3) PF joint on LA view	53.2 ± 0.3
Sclerosis (OARSI grades 0-3) PF joint on LA view	53.1 ± 0.4
Osteophytes (OARSI grades 0-3) patella inferior PF joint on LA view	59.5 ± 0.2
Osteophytes (OARSI grades 0-3) patella superior PF joint on LA view	60.1 ± 0.3
Osteophytes (OARSI grades 0-3) tibia anterior PF joint on LA view	55.5 ± 0.1
Osteophytes (OARSI grades 0-3) tibia posterior PF joint on LA view	59.4 ± 0.3

report the classification performance (AUC, %) for the features derived from the two views separately, and from combining the features from both views (Table 6.2). We notice that, similarly to previous works [83], the PA view performs better than the lateral view at discriminating people experiencing pain regardless of the pain score used. Furthermore, the combination of both view almost in all cases leads to improved results. The best overall results are achieved by the combined PA+lateral models trained on the whole set of appearance features. Consistent pain was the easiest to classify, followed by SAQ, Clinic and TScreen.

6.3.3 Testing combinations of radiographic features

Next we combined all the available manually graded features and repeated the same experiments, considering as well the SAQ, TScreen and consistent pain scores. The input to the random forest classifiers was thus a vector containing all grades. Each node in each tree could thus branch by examining any one of the features (grades). (see Table 6.4). Removing participants with widespread pain made little difference, though it dramatically reduced the size of the dataset. Adding BMI and gender significantly improved the ROC for both Clinic and SAQ pain, while it is not different in the TScreen and Consistent dataset. While we could not perform statistical tests evaluating whether consistent frequent knee pain AUCs were different from one time point pain reports, the AUCs for consistent pain were higher especially for manual grades (e.g. 73.9 vs. 62.8 – 66.7) and the standard deviations around these estimates were narrow.

When working with each pain score individually, the results show that, although there was variation depending on what features and images were used, using the best performing automated model gives results that were not significantly different from those of manual grades. The best model was chosen as the one achieving the highest AUC overall. The consistent pain score was the only one to have all fully automated models performing worse than manual grades [1].

Finally we combined manual grades with the appearance features extracted from the model and found that this combination was not more discriminative than using manual grades alone. Following the usual order, TScreen, Clinic, SAQ and Consistent pain the results were respectively 65.6, 63, 68 and 75.6.

In our dataset there were a number of participants that had no signs of radiographic OA in either PA or lateral view and still experienced frequent pain. We wished to

[1] Comparison eliminating participants with widespread pain was performed using manual grades+Gender+BMI features.

	Lateral view				PA view				Lateral + PA views			
	TScreen	Clinic	SAQ	Consist	TScreen	Clinic	SAQ	Consist	TScreen	Clinic	SAQ	Consist
Shape	58.7 ± 0.4	60.1 ± 0.6	62.1 ± 0.7	62.8 ± 0.3	59.5 ± 0.5	62.9 ± 0.4	62.7 ± 0.5	67.5 ± 0.9	61 ± 1	63.7 ± 0.7	65.1 ± 0.5	68.7 ± 1.6
Texture	62.5 ± 1	62 ± 0.6	65.3 ± 0.7	67 ± 0.4	63 ± 0.5	65.3 ± 0.4	66.4 ± 0.3	72.8 ± 0.5	63.8 ± 0.4	65.2 ± 0.1	67.6 ± 0.3	72.9 ± 0.5
Appearance	62.1 ± 0.5	62.1 ± 0.8	63.9 ± 0.7	66.1 ± 0.8	62.1 ± 0.5	65.1 ± 0.6	66.6 ± 0.2	72.3 ± 0.4	$\mathbf{63.8 \pm 0.2}$	$\mathbf{65.6 \pm 0.9}$	$\mathbf{67.7 \pm 0.3}$	$\mathbf{73.1 \pm 0.7}$

Table 6.2: AUC using shape, texture and appearance parameters extracted from fully automatically found points. Appearance concatenates shape and texture measurements.

	Lateral view				PA view				Lateral + PA views			
	TScreen	Clinic	SAQ	Consist	TScreen	Clinic	SAQ	Consist	TScreen	Clinic	SAQ	Consist
Shape	54 ± 1	54.3 ± 0.9	58.8 ± 1	47.4 ± 0.6	56.2 ± 0.7	58 ± 0.5	60.5 ± 0.5	55.8 ± 0.4	57.3 ± 0.9	57.8 ± 0.4	61.9 ± 0.9	54.2 ± 1.4
Texture	57.6 ± 0.3	56.5 ± 1.2	63.4 ± 0.5	54.5 ± 1.3	59 ± 1	$\mathbf{59.4 \pm 0.5}$	64.5 ± 0.7	$\mathbf{62.9 \pm 1}$	$\mathbf{59.7 \pm 0.7}$	58.3 ± 1.2	64.4 ± 0.7	58.6 ± 2.4
Appearance	58 ± 1	56.3 ± 1.1	62.3 ± 0.9	52.7 ± 1.9	59.4 ± 0.1	57.9 ± 0.7	63.7 ± 0.8	58.7 ± 1.2	58.7 ± 0.7	58.1 ± 0.6	$\mathbf{65.8 \pm 0.8}$	56.6 ± 3.1

Table 6.3: Considering only knees with no sign of osteoarthritis. Can we distinguish who is experiencing pain using the fully automated model?

	Features	# Samples	$AUC \pm SD$	P-Value vs. Referent
Telephonic Screening Interview	Manual Grades	2756	62.8 ± 0.4	Referent
	Manual Grades+Gender+BMI		$\mathbf{66 \pm 0.5}$	0.001
	Best Automated		63.8 ± 0.2	0.15
	Manual+Automated		65.6 ± 0.3	<0.001
	Removing WS Pain	1374	61 ± 0.2	0.51
Clinic	Manual Grades	2756	66.4 ± 0.2	Referent
	Manual Grades+Gender+BMI		$\mathbf{68.8 \pm 0.2}$	<0.001
	Best Automated		65.6 ± 0.9	0.29
	Manual+Automated		63 ± 0.3	0.01
	Removing WS Pain	1374	61 ± 0.2	0.02
Self-Assessed Questionnaire(HOME)	Manual Grades	2756	66.7 ± 0.3	Referent
	Manual Grades+Gender+BMI		68.9 ± 0.4	<0.001
	Best Automated		67.7 ± 0.3	0.30
	Manual+Automated		68 ± 0.2	0.05
	Removing WS Pain	1374	$\mathbf{69 \pm 0.2}$	0.10
Consistent Pain	Manual Grades	1066	73.9 ± 0.5	Referent
(answered yes to pain at all time points)	Manual Grades+Gender+BMI		76.1 ± 0.2	0.01
	Best Automated		73.1 ± 0.7	0.97
	Manual+Automated		75.6 ± 0.6	0.14
	Removing WS Pain	565	$\mathbf{78 \pm 1}$	0.04

Table 6.4: Performance of RF classifiers when using all the available clinician grades as features. The p-values depicted compare the AUCs with the referent in that pain group. For example, for telephone screening, compared with manual grades, none of the other approaches was significant.

explore whether shape and appearance features could be used to discriminate those experiencing pain from those who did not, among people with no clinician graded radiographic signs. We carried out separate analyses for the three types of pain and the consistent pain score. (see Table 6.3). The highlighted entries correspond to the highest AUCs for each pain score. Texture parameters give the best performance for three out of the four pain score and the SAQ is the one with the highest AUC overall. Consistent knee pain is the hardest to correctly classify using these features. This is, considering the results in Table 6.4, shows the very close link between structural features and consistent pain. It has to be said, that for many of the experiments in Table 6.3 is was only possible to find slight correlation between radiographic features and the pain score, in several instances the classifier performed no better than random.

6.4 Discussion

We built a number of binary classifiers to try and distinguish participants with frequent knee pain from those without pain, using various sets of features and selecting subsets of participants satisfying certain conditions. We found that identifying persons with consistent knee pain from manually read radiographic features gave the highest AUC. The best model using features computed automatically from the images could be used

to discriminate pain from non-pain, without significant loss in AUC compared to using grades assigned by experienced clinicians. Furthermore, removing participants with widespread pain does not make the classification easier for either of the pain scores considered.

This fully automated approach has similarly been applied to diagnose osteoporosis vertebrae and wrist fractures [13, 30] among other applications. We do not believe that previous studies have formally compared manual approaches to automated ones.

The main strengths of this work are in the size of the dataset used, one order of magnitude larger that most similar studies. In addition, we presented the most comprehensive corpus of experiments looking at correlations between radiographs and symptomatic OA, using both PA and lateral view images, therefore including PF joint [10] and posterior compartments. Finally, we explored for the first time OARSI grades of the lateral view of the MOST study and their combination with other radiographic features.

Limitations are the absence of skyline view radiographs, which could provide further discriminative information, but were not acquired during the MOST study. MRI volumes, that have shown to be more correlated with symptoms were also excluded from this work, and they are the most promising addition to improve performance.

Future work will involve the study of MR images and the grades that are commonly assigned to them as well as the training of a convolutional neural network architecture to study the images. Further areas of interest will be the research for patterns in functional MR images of the brain related to pain perception in participants with OA or at risk of developing it.

Contributors LM conceived the project, contributed to the study design, analysed and interpreted data, drafted the article and approved the final version for submission; DF conceived and oversaw the project, contributed to the study design, analysed and interpreted data, drafted the article and approved the final version for submission; MP contributed to the statistical analysis and drafting the paper TFC conceived and oversaw the project, contributed to the study design, analysed and interpreted data, drafted the article and approved the final version for submission

Funding The research leading to these results has received funding from EPSRC Centre for Doctoral Training grant 1512584.

Patient consent Obtained.

Ethics Approval This study was conducted in accordance with the declaration of Helsinki, Good Clinical practices and applicable regulatory requirements.

Competing Interest The authors have no competing interests.

6.4.1 Aknowledgements

The authors would like to thank our Raja Ebsim, Luke Chaplin and Manuele Reani for their useful comments.

6.5 Supplementary Results

This section reports supplementary material on the paper described in this chapter.

6.5.1 Descriptive statistic analysis of the presence of frequent knee pain for different KL grades

Figure 6.2 shows the number of people from our study experiencing pain for all possible values of the KL grade. There were no participants with pain among the KL0 group. The KL1 group participants were almost uniformly distributed between painful and non-painful knees. We notice that as the KL grade increases the proportion of painful knees increases reaching around 75% for KL4. This is to support the claim that radiographic information is related with pain symptoms.

6.5.2 Results of using individual structural features to predict frequent knee pain when we use the consistent knee pain score

The corresponding results are shown in Table 6.5. Similarly to what was seen previously the individual features reporting the highest AUCs are a combination of both PA and lateral view grades, further underlining the need for using multiple views. The KL grade score is consistently the individual feature reporting the highest AUC value. The TScreen score (see Table 6.6) which is the one collected the earliest with respect to image acquisition is also the one showing the weakest correlation with radiographic features, followed by the Clinic score (reported in the previous paper), the SAQ score (see Table 6.7) and finally the consistent knee pain score. This result is somewhat surprising. One would expect the Clinic score to have the highest correlation with imaging

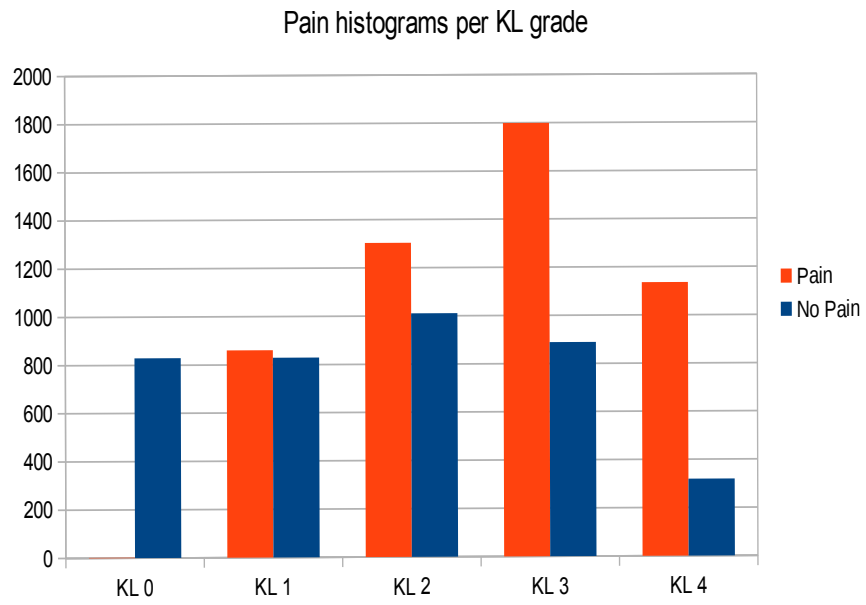


Figure 6.2: The proportion of painful knees increases as the KL grade increases. Data from baseline knees of the MOST study.

features, being acquired in proximity of image acquisition. A possible explanation for this is that the SAQ score is acquired by the study participants in a familiar environment and without any time pressure, allowing them to better consider their responses. Bone attrition also has does not have any current pain signal. This result should be further investigated to understand whether there is any strong inconsistency between the way this variable is graded in radiographs as opposed to MRIs. Finally, overall there is small difference between the pain predictive power of the KL grade alone and the whole set of 36 features, suggesting that the KL grade is a good proxy variable, but if we want to understand pain source we should be looking at different quantitative assessments.

6.5.3 Summary of results on the MOST dataset

This section summarises the majority of results achieved on the MOST dataset from the works described in this thesis. Table 6.8 reports the experimental results for the OA related tasks investigated in this work. When reporting binary classification performance we use AUC (%), while for multi-class we use the mean per class accuracy (%). The manual annotation results correspond to using features extracted from human

annotated bone landmarks. All results are the best results in each scenario. For example, the manual annotation results were achieved using only a few hundred images (see Chapter 4 for details), while to achieve the automated results we could deploy the entire MOST dataset. All in all, when a direct comparison is possible and if we consider the best results only, the PA view performs better than the Lateral view at the Binary OA tasks, while the lateral view performs better on pain related tasks. The combination of the two views only marginally improved performance.

Table 6.5: Testing each radiographic feature individually using the consistent pain score.

Variable	AUC (%)
Chondrocalcinosis (OARSI grades 0-1) PF joint on LA view	51.2 ± 0.2
Osteophytes(OARSI grades 0-3) femur anterior PF joint on LA view	62.2 ± 0.4
Osteophytes(OARSI grades 0-3) femur posterior PF joint on LA view	64.7 ± 0.6
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on LA view	56.3 ± 0.6
Joint space narrowing (OARSI grades 0-3) medial TF compartment on LAT view	63.7 ± 0.7
Effusion (OARSI grades 0-1) PF joint on LA view	58 ± 0.7
Kellgren & Lawrence (grades 0-4) on PA view	$\mathbf{71.7 \pm 0.2}$
Chondrocalcinosis (OARSI grades 0-1) lateral TF compartment on PA view	50.8 ± 0.2
Cyst (OARSI grades 0-3) femur lateral TF compartment on PA view	50.6 ± 0.2
Osteophytes (OARSI grades 0-3) femur lateral TF compartment on PA view	64.3 ± 0.6
Sclerosis (OARSI grades 0-3) femur lateral TF compartment on PA view	55.6 ± 0.2
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on PA view	57 ± 0.5
Attrition (OARSI grades 0-1) lateral TF compartment on PA view	50.9 ± 0.3
Cyst (OARSI grades 0-3) tibia lateral TF compartment on PA view	50.7 ± 0.1
Osteophytes (OARSI grades 0-3) tibia lateral TF compartment on PA view	64.7 ± 0.3
Sclerosis (OARSI grades 0-3) tibia lateral TF compartment on PA view	55.3 ± 0.5
Chondrocalcinosis (OARSI grades 0-1) medial TF compartment on PA view	51.4 ± 0.5
Cyst (OARSI grades 0-3) femur medial TF compartment on PA view	51.1 ± 0.5
Osteophytes (OARSI grades 0-3) femur medial TF compartment on PA view	66.2 ± 0.4
Sclerosis (OARSI grades 0-3) femur medial TF compartment on PA view	62.7 ± 0.4
Joint space narrowing (OARSI grades 0-3) medial TF compartment on PA view	63.5 ± 0.2
Attrition (OARSI grades 0-1) medial TF compartment on PA view	52.7 ± 0.5
Cyst (OARSI grades 0-3) tibia medial TF compartment on PA view	52.1 ± 0.5
Osteophytes (OARSI grades 0-3) tibia medial TF compartment on PA view	64.5 ± 0.6
Sclerosis (OARSI grades 0-3) tibia medial TF compartment on PA view	63.5 ± 0.5
Ossification (OARSI grades 0-3) patella tendon lower PF joint on LA view	49.1 ± 1.3
Ossification (OARSI grades 0-3) patella tendon upper PF joint on LA view	50 ± 0.6
Ossified loose body (OARSI grades 0-1) femur posterior PF joint on LA view	53.5 ± 0.4
Ossification of QF insertion (OARSI grades 0-3) PF joint on LA view	52.4 ± 0.6
Cyst (OARSI grades 0-3) PF joint on LA view	50.1 ± 0.8
Joint space narrowing (OARSI grades 0-3) PF joint on LA view	54 ± 0.4
Sclerosis (OARSI grades 0-3) PF joint on LA view	52.7 ± 0.5
Osteophytes (OARSI grades 0-3) patella inferior PF joint on LA view	64.1 ± 0.4
Osteophytes (OARSI grades 0-3) patella superior PF joint on LA view	64.7 ± 0.5
Osteophytes (OARSI grades 0-3) tibia anterior PF joint on LA view	58.1 ± 0.4
Osteophytes (OARSI grades 0-3) tibia posterior PF joint on LA view	64.5 ± 0.7

Table 6.6: Testing each radiographic feature individually using the Telephone Screening interview pain score.

Variable	AUC (%)
Chondrocalcinosis (OARSI grades 0-1) PF joint on LA view	49.8 ± 0.3
Osteophytes(OARSI grades 0-3) femur anterior PF joint on LA view	56.4 ± 0.3
Osteophytes(OARSI grades 0-3) femur posterior PF joint on LA view	58.1 ± 0.2
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on LA view	54 ± 0.2
Joint space narrowing (OARSI grades 0-3) medial TF compartment on LAT view	58.7 ± 0.4
Effusion (OARSI grades 0-1) PF joint on LA view	55.9 ± 0.2
Kellgren & Lawrence (grades 0-4) on PA view	$\mathbf{62.7 \pm 0.3}$
Chondrocalcinosis (OARSI grades 0-1) lateral TF compartment on PA view	49.6 ± 0.6
Cyst (OARSI grades 0-3) femur lateral TF compartment on PA view	50.3 ± 0.1
Osteophytes (OARSI grades 0-3) femur lateral TF compartment on PA view	58.6 ± 0.1
Sclerosis (OARSI grades 0-3) femur lateral TF compartment on PA view	53.5 ± 0.2
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on PA view	53.8 ± 0.4
Attrition (OARSI grades 0-1) lateral TF compartment on PA view	50 ± 0.4
Cyst (OARSI grades 0-3) tibia lateral TF compartment on PA view	50.6 ± 0.3
Osteophytes (OARSI grades 0-3) tibia lateral TF compartment on PA view	58 ± 0.3
Sclerosis (OARSI grades 0-3) tibia lateral TF compartment on PA view	53.7 ± 0.5
Chondrocalcinosis (OARSI grades 0-1) medial TF compartment on PA view	49 ± 0.5
Cyst (OARSI grades 0-3) femur medial TF compartment on PA view	50.6 ± 0.1
Osteophytes (OARSI grades 0-3) femur medial TF compartment on PA view	60.3 ± 0.2
Sclerosis (OARSI grades 0-3) femur medial TF compartment on PA view	58.3 ± 0.3
Joint space narrowing (OARSI grades 0-3) medial TF compartment on PA view	58.4 ± 0.3
Attrition (OARSI grades 0-1) medial TF compartment on PA view	52.2 ± 0.2
Cyst (OARSI grades 0-3) tibia medial TF compartment on PA view	50.6 ± 0.3
Osteophytes (OARSI grades 0-3) tibia medial TF compartment on PA view	59.6 ± 0.3
Sclerosis (OARSI grades 0-3) tibia medial TF compartment on PA view	58.6 ± 0.2
Ossification (OARSI grades 0-3) patella tendon lower PF joint on LA view	50.4 ± 0.1
Ossification (OARSI grades 0-3) patella tendon upper PF joint on LA view	49.8 ± 0.6
Ossified loose body (OARSI grades 0-1) femur posterior PF joint on LA view	51.7 ± 0.3
Ossification of QF insertion (OARSI grades 0-3) PF joint on LA view	50.7 ± 0.2
Cyst (OARSI grades 0-3) PF joint on LA view	50.7 ± 0.4
Joint space narrowing (OARSI grades 0-3) PF joint on LA view	52.9 ± 0.3
Sclerosis (OARSI grades 0-3) PF joint on LA view	52 ± 0.4
Osteophytes (OARSI grades 0-3) patella inferior PF joint on LA view	58 ± 0.1
Osteophytes (OARSI grades 0-3) patella superior PF joint on LA view	58 ± 0.2
Osteophytes (OARSI grades 0-3) tibia anterior PF joint on LA view	55.4 ± 0.3
Osteophytes (OARSI grades 0-3) tibia posterior PF joint on LA view	58.6 ± 0.1

Table 6.7: Testing each radiographic feature individually using the Self Assessed Questionnaire(Home) pain score.

Variable	AUC (%)
Chondrocalcinosis (OARSI grades 0-1) PF joint on LA view	49.6 ± 0.4
Osteophytes(OARSI grades 0-3) femur anterior PF joint on LA view	58.2 ± 0.2
Osteophytes(OARSI grades 0-3) femur posterior PF joint on LA view	60.3 ± 0.5
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on LA view	53.5 ± 0.5
Joint space narrowing (OARSI grades 0-3) medial TF compartment on LAT view	60.2 ± 0.4
Effusion (OARSI grades 0-1) PF joint on LA view	56 ± 0.1
Kellgren & Lawrence (grades 0-4) on PA view	$\mathbf{65.9 \pm 0.2}$
Chondrocalcinosis (OARSI grades 0-1) lateral TF compartment on PA view	49.5 ± 0.5
cyst (OARSI grades 0-3) femur lateral TF compartment on PA view	50.4 ± 0.1
osteophytes (OARSI grades 0-3) femur lateral TF compartment on PA view	59.6 ± 0.3
Sclerosis (OARSI grades 0-3) femur lateral TF compartment on PA view	53.1 ± 0.1
Joint space narrowing (OARSI grades 0-3) lateral TF compartment on PA view	53.8 ± 0.2
Attrition (OARSI grades 0-1) lateral TF compartment on PA view	50.7 ± 0.1
Cyst (OARSI grades 0-3) tibia lateral TF compartment on PA view	50.2 ± 0.3
Osteophytes (OARSI grades 0-3) tibia lateral TF compartment on PA view	59.5 ± 0.3
Sclerosis (OARSI grades 0-3) tibia lateral TF compartment on PA view	53.1 ± 0.3
Chondrocalcinosis (OARSI grades 0-1) medial TF compartment on PA view	49.8 ± 0.5
Cyst (OARSI grades 0-3) femur medial TF compartment on PA view	50.7 ± 0.2
Osteophytes (OARSI grades 0-3) femur medial TF compartment on PA view	62.2 ± 0.2
Sclerosis (OARSI grades 0-3) femur medial TF compartment on PA view	59.9 ± 0.3
Joint space narrowing (OARSI grades 0-3) medial TF compartment on PA view	60.9 ± 0.1
Attrition (OARSI grades 0-1) medial TF compartment on PA view	51.6 ± 0.1
Cyst (OARSI grades 0-3) tibia medial TF compartment on PA view	51.1 ± 0.1
Osteophytes (OARSI grades 0-3) tibia medial TF compartment on PA view	61.1 ± 0.2
Sclerosis (OARSI grades 0-3) tibia medial TF compartment on PA view	60 ± 0.3
Ossification (OARSI grades 0-3) patella tendon lower PF joint on LA view	50.4 ± 0.5
Ossification (OARSI grades 0-3) patella tendon upper PF joint on LA view	51 ± 0.2
Ossified loose body (OARSI grades 0-1) femur posterior PF joint on LA view	51.9 ± 0.2
Ossification of QF insertion (OARSI grades 0-3) PF joint on LA view	52.4 ± 0.2
Cyst (OARSI grades 0-3) PF joint on LA view	51.1 ± 0.2
Joint space narrowing (OARSI grades 0-3) PF joint on LA view	54.3 ± 0.2
Sclerosis (OARSI grades 0-3) PF joint on LA view	53.6 ± 0.2
Osteophytes (OARSI grades 0-3) patella inferior PF joint on LA view	59.5 ± 0.4
Osteophytes (OARSI grades 0-3) patella superior PF joint on LA view	61 ± 0.2
Osteophytes (OARSI grades 0-3) tibia anterior PF joint on LA view	55 ± 0.2
Osteophytes (OARSI grades 0-3) tibia posterior PF joint on LA view	60.2 ± 0.5

Table 6.8: A table summarising our results on the MOST dataset. All results are reported using the AUC(%) with the exception of KL grade classification for which we used the mean per class accuracy (%). N/A means that the corresponding experiments were not performed as part of our work. 'Auto' stands for fully automated landmark annotation, while 'Manual' stands for manually annotated points.

	Landmarks	Method	Binary OA	KL Grade	Binary Future OA	Pain	Future Pain
PA	Auto	RF-CLM	90.4 ± 0.1	N/A	$\mathbf{65.1 \pm 0.4}$	65.3 ± 0.4	56.8 ± 1.3
Lat	Manual	RF-CLM	84 ± 2	$\mathbf{47.9 \pm 0.8}$	N/A	N/A	N/A
Lat	Auto	RF-CLM	85.3 ± 0.2	43.9 ± 1	58 ± 0.8	62.1 ± 0.8	55.7 ± 1.3
Lat	Auto	IF	87.6 ± 0.9	N/A	61.1 ± 1.8	N/A	N/A
Both	Auto	RF-CLM	$\mathbf{90.5 \pm 0.01}$	N/A	63.2 ± 0.9	$\mathbf{65.6 \pm 0.9}$	$\mathbf{56.9 \pm 0.1}$

Chapter 7

Discussion

We have described the first fully automated system to segment the knee joint from lateral view radiographs and compared its discriminative potential for automatically studying osteoarthritis against standard PA radiographs. We looked into the combination of radiographic features from the two views, developed a new machine learning classifier and showed that it gave better performance on our task.

Finally, we studied the relationship between pain and radiographic features. We determined which radiographic features are most correlated with current frequent pain and built a fully automated model that performs as well as manually graded features.

7.1 Conclusions

Our work contributed to improve the classification accuracy of different knee osteoarthritis related tasks: binary OA classification, prediction of both future frequent pain and structural OA and current pain classification. These improvements were achieved by: extending the data to both PA and lateral images; including a number of subjects and images two orders of magnitude larger in size than previous studies and by refining the machine learning classifier.

Our work leads us to conclude the following:

- Lateral knee radiographs contain very informative disease information. These images are significantly more challenging to segment due to the presence of bone profiles that look almost identical and can intersect or even completely overlap. The lateral view includes features that have lower discriminative ability when diagnosing knee osteoarthritis, but are extremely useful to study the patellofemoral joint.

- Combining features from different views only slightly improves performance. Predicting future frequent pain is the hardest task and the only one where features extracted from lateral view radiographs perform on average better than those from the PA view. Our work on the combination was the first to use a large scale dataset (20k images).

- Unlike other works in the field, all the experiments performed as part of this project did not include any image selection based on data quality. This means that our results are more realistic and our models have higher capability of dealing with real data.

- Introducing indecisive windows in a random forest can help improve performance on classification tasks. Our indecisive forest achieved the state of the art at binary OA classification and prediction of future onset of OA using lateral images. It is as yet unclear whether the proposed back-propagation based optimisation of a pre-trained indecisive forest is in general beneficial or not.

- The current pain classification experiments showed that the proposed consistent pain measure is the pain score with the highest correlation with manual radiographic features. Furthermore, it is possible to build a fully automated feature extractor whose features perform as well as grades assigned by experienced clinicians.

- Osteophytes, medial JSN and both tibial and femoral medial sclerosis were found to be most strongly linked with frequent knee pain in our study.

- When focusing on radiographs with no signs of osteoarthritis it was not possible to distinguish painful knees from non-painful ones.

7.2 Future Work

Potential future research directions include:

Grading Scheme. There is need for more reliable grading methods to use as ground truth. Current grading schemes carry too much subjectivity and inter-rater variability. While these are reduced when dealing with a machine learning classifier, having a more robust and reproducible assessment would be highly beneficial for the AI system.

Include more radiographic views. Adding lateral view images helps when studying structural features that could not be studied with PA radiographs only. Skyline radiographs are often available and provide further information on the patellofemoral joint and femoral osteophytes. These images are rarely included in studies and including them could help achieve a more comprehensive disease evaluation just using radiographic images.

Different ways of combining the two views. In our work, we described the most natural way of combining features, which is by concatenating them. While this can work and give insight to the level of synergy between two sets of features, other ways can be explored. Among them it is worth mentioning deep learning approaches, such as Deep Siamese Networks [62] and 2D-3D reconstruction approaches [113].

Indecisive forest for regression problems. The results we obtained with the Indecisive forest showed its effectiveness when dealing with classification tasks. Preliminary experiments have also been performed on regression tasks, resulting in somewhat inconclusive outcomes. Further investigation is needed to find out whether this novel methodology is beneficial for tasks such as landmark point detection.

MRI images. Knee osteoarthritis is a disease that usually initially affects articular cartilage only. As a result, the best way to assess early knee OA is by looking at MRI volumes. Similar techniques to what we have described in this work can be applied to 3D images. A landmark point detector can be applied to a set of 2D slices obtaining a bone segmentation together with a shape model. Alternative ways of studying MRI images use CNN based segmentation architectures similar to U-Net [93], such as [81, 16].

Radiographic image augmentation. There is need for methodologies for understanding radiographic texture structure to the extent of being able to generate realistically looking radiographic images. Research in this field can help design novel augmentation techniques for radiographic datasets and improve performance on abnormality detection tasks. Examples of techniques that could be used for this aim are Image Auto-encoders [77] and Generative Adversarial Networks (GANs) [44]. GANs could be used by using a U-Net like generator architecture to reconstruct radiographic images and a VGG16 or ResNet architecture to discriminate real from fake images.

Once trained, the middle layer of the U-Net could be replaced by noise sampled from the distribution of real radiographs to generate realistic, synthetic images. Apart from the standard GAN loss terms, a good base for training the architecture would be L1 loss and perceptual loss for reconstruction and sigmoid cross entropy for the discriminator as in [55].

Use of CNNs depending on the available data. Medical imaging applications suffer from data scarcity. That is why designing strategies for the exploitation of small amounts of annotated data is particularly important.

- *Very little data*: A possible strategy for this would be few-shot learning approaches, for examples deep learning models based on meta-learning [39, 100], where one of the objectives of the training phase is learning the learning algorithms from a small set of annotated images.

- *Few Hundred images*: In medical images there is also need for a recognised pre-training dataset, such as ImageNet for Image Recognition and MSCoco [71] or Open Images [66] for Object Detection. With a dataset of this size it would be possible to successfully pre-train deep feature extractor blocks such as ResNet50 or ResNet152 and then fine-tune application specific architectures using the feature extractor as backbone. In 2017, Stanford University announced that it was going to release a pre-training dataset for medical imaging [1], but the data has not been made public yet.

Refining the shape model. Our model can be extended to include both tibial condyles and the femoral crista and articulation of both medial and lateral patellofemoral joints. This would require time and support from experienced radiologists but would allow the model to capture more information on the spatial relation between bones in the knee joint.

Objectivising the study of pain patterns. When we want to study pain patterns we usually need to rely on patients' self reports. This makes our data prone to subjectivity. Pain symptoms can be severely affected by mood and individual perceptions. That is why there is need for ways to assess it that rely on quantifiable measurements. One way of doing this is by looking at patterns of brain signals. Electroencephalography (EEG) and functional MRI (fMRI) have been recently used for this kind of problem

[27, 111, 104]. Such an approach has the chance of making the measure more reliable and robust, though requiring more time and money.

Unbalanced datasets. One common issue with medical applications of machine learning is class imbalance. One often has to deal with datasets where one of the classes is significantly under-represented. This makes the classification problem harder for the model, risking building something unable to correctly detect the minority class. For this reason, there has been interest in these kinds of problems [106], but there is need to further investigate ways to both transform the datasets to reduce the unbalance at its source (under- and over-sampling are the simplest approaches to this) and to develop models that are able to cope with heavy imbalances.

Bibliography

[1] Medical image net, a petabyte-scale, cloud-based, multi-institutional, searchable, open repository of diagnostic imaging studies for developing intelligent image analysis systems. http://langlotzlab.stanford.edu/projects/medical-image-net/.

[2] New neural algorithm can paint photos in style of any artist from van gogh to picasso. https://www.boredpanda.com/computer-deep-learning-algorithm-painting-masters/.

[3] S. Ahlbäck. Osteoarthrosis of the knee. a radiographic investigation. *Acta Radiologica: diagnosis*, pages Suppl–277, 1968.

[4] R. D. Altman, M. Hochberg, J. W. Murphy, F. Wolfe, and M. Lequesne. Atlas of individual radiographic features in Osteoarthritis. *Osteoarthritis and Cartilage*, 3:3–70, 1995.

[5] L. Ameye, D. Aria, K. Jepsen, A. Oldberg, T. XU, and M. F. YOUNG. Abnormal collagen fibrils in tendons of biglycan/fibromodulin-deficient mice lead to gait impairment, ectopic ossification, and Osteoarthritis. *The FASEB Journal*, 16(7):673–680, 2002.

[6] L. Anifah, I. K. E. Purnama, M. Hariadi, and M. H. Purnomo. Osteoarthritis classification using self organizing map based on gabor kernel and contrast-limited adaptive histogram equalization. *The open Biomedical Engineering Journal*, 7:18, 2013.

[7] J. Antony, K. McGuinness, K. Moran, and N. E. O'Connor. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 376–390. Springer, 2017.

[8] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran. Quantifying radiographic knee Osteoarthritis severity using deep convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1195–1200. IEEE, 2016.

[9] R. Aspden. Osteoarthritis: a problem of growth not decay? *Rheumatology*, 47(10):1452–1460, 2008.

[10] J. Bedson and P. R. Croft. The discordance between clinical and radiographic knee Osteoarthritis: a systematic search and summary of the literature. *BMC musculoskeletal disorders*, 9(1):116, 2008.

[11] M. A. Bowes, G. R. Vincent, C. B. Wolstenholme, and P. G. Conaghan. A novel method for bone area measurement provides new insights into Osteoarthritis and its progression. *Annals of the Rheumatic Diseases*, 74(3):519–525, 2015.

[12] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes. Classification of osteoporotic vertebral fractures using shape and appearance modelling. In *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pages 133–147. Springer, 2017.

[14] J. Canny. A computational approach to edge detection. In *Readings in Computer Vision*, pages 184–203. Elsevier, 1987.

[15] A. Chen, C. Gupte, K. Akhtar, P. Smith, and J. Cobb. The global economic cost of Osteoarthritis: how the uk compares. *Arthritis*, 2012.

[16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[17] P. Conaghan, D. Felson, G. Gold, S. Lohmander, S. Totterman, and R. Altman. MRI and non-cartilaginous structures in knee Osteoarthritis. *Osteoarthritis and Cartilage*, 14:87–94, 2006.

[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[19] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291. Springer, 2012.

[20] T. F. Cootes and C. J. Taylor. Statistical models of appearance for medical image analysis and computer vision. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 236–249. International Society for Optics and Photonics, 2001.

[21] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[22] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 3, pages 929–938, 2006.

[23] M. Cross, E. Smith, D. Hoy, L. Carmona, F. Wolfe, T. Vos, B. Williams, S. Gabriel, M. Lassere, N. Johns, et al. The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. *Annals of the Rheumatic Diseases*, 2014.

[24] M. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. L. Hill, et al. The global burden of hip and knee Osteoarthritis: estimates from the global burden of disease 2010 study. *Annals of the Rheumatic Diseases*, pages annrheumdis–2013, 2014.

[25] J. Dacre, J. Coppock, K. Herbert, D. Perrett, and E. Huskisson. Development of a new radiographic scoring system using digital image analysis. *Annals of the Rheumatic Diseases*, 48(3):194, 1989.

[26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[27] G. Deshpande, S. LaConte, G. A. James, S. Peltier, and X. Hu. Multivariate granger causality analysis of fmri data. *Human brain mapping*, 30(4):1361–1373, 2009.

[28] P. Dieppe, H. Basler, J. Chard, P. Croft, J. Dixon, M. Hurley, S. Lohmander, and H. Raspe. Knee replacement surgery for osteoarthritis: effectiveness, practice

variations, indications and possible determinants of utilization. *Rheumatology (Oxford, England)*, 38(1):73–83, 1999.

[29] J. Duryea, J. Li, C. Peterfy, C. Gordon, and H. Genant. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. *Medical physics*, 27(3):580–591, 2000.

[30] R. Ebsim, J. Naqvi, and T. Cootes. Detection of wrist fractures in X-ray images. In *Workshop on Clinical Image-Based Procedures*, pages 1–8. Springer, 2016.

[31] G. J. Edwards, A. Lanitis, C. J. Taylor, and T. F. Cootes. Statistical models of face images—improving specificity. *Image and Vision Computing*, 16(3):203–211, 1998.

[32] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[33] T. Fairbank. Knee joint changes after meniscectomy. *Bone & Joint Journal*, 30(4):664–670, 1948.

[34] D. T. Felson, J. J. Anderson, A. Naimark, A. M. Walker, and R. F. Meenan. Obesity and knee Osteoarthritis: the Framingham study. *Annals of Internal Medicine*, 109(1):18–24, 1988.

[35] D. T. Felson, T. E. McAlindon, J. J. Anderson, B. W. Weissman, P. Aliabadi, S. Evans, D. Levy, and M. P. LaValley. Defining radiographic Osteoarthritis for the whole knee. *Osteoarthritis and Cartilage*, 5(4):241–250, 1997.

[36] D. T. Felson, J. Niu, T. Neogi, J. Goggins, M. C. Nevitt, F. Roemer, J. Torner, C. E. Lewis, and A. Guermazi. Synovitis and the risk of knee Osteoarthritis: the most study. *Osteoarthritis and Cartilage*, 24(3):458–464, 2016.

[37] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[39] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[40] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[41] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *Decision forests for computer vision and medical image analysis*, pages 143–157. Springer, 2013.

[42] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, 2011.

[43] J. I. Galvan-Tejada, J. M. Celaya-Padilla, V. Treviño, and J. G. Tamez-Pena. Knee osteoarthritis pain prediction from X-ray imaging: Data from osteoarthritis initiative. In *International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 194–199. IEEE, 2014.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[45] S. S. Gornale, P. U. Patravali, and R. R. Manza. Detection of Osteoarthritis using knee X-ray image analyses: a machine vision based approach. *International Journal of Computer Applications*, 145(1), 2016.

[46] L. Gossec, J. Jordan, S. Mazzuca, M.-A. Lam, M. Suarez-Almazor, J. Renner, M. Lopez-Olivo, G. Hawker, M. Dougados, and J. Maillefert. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee Osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the oarsi–omeract task force. *Osteoarthritis and Cartilage*, 16(7):742–748, 2008.

[47] S. Grochowski, K. Amrami, and K. Kaufman. Semi-automated digital image analysis of patellofemoral joint space width from lateral knee radiographs. *Skeletal radiology*, 34(10):644–648, 2005.

[48] M. T. Hannan, D. T. Felson, and T. Pincus. Analysis of the discordance between radiographic changes and knee pain in Osteoarthritis of the knee. *The Journal of rheumatology*, 27(6):1513–1517, 2000.

[49] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[50] J. Hladůvka, B. T. M. Phuong, R. Ljuhar, D. Ljuhar, A. M. Rodrigues, J. C. Branco, and H. Canhão. Femoral ROIs and entropy for texture-based detection of osteoarthritis from high-resolution knee radiographs. *arXiv preprint arXiv:1703.09296*, 2017.

[51] M. Hoffman. Picture of the knee joint. `https://www.webmd.com/pain-management/knee-pain/picture-of-the-knee#1`.

[52] D. Hoy, L. March, P. Brooks, F. Blyth, A. Woolf, C. Bain, G. Williams, E. Smith, T. Vos, J. Barendregt, et al. The global burden of low back pain: estimates from the global burden of disease 2010 study. *Annals of the Rheumatic Diseases*, 73(6):968–974, 2014.

[53] D. Hoy, L. March, P. Brooks, F. Blyth, A. Woolf, C. Bain, G. Williams, E. Smith, T. Vos, J. Barendregt, et al. The global burden of low back pain: estimates from the global burden of disease 2010 study. *Annals of the Rheumatic Diseases*, 73(6):968–974, 2014.

[54] D. J. Hunter, J. Niu, Y. Zhang, S. Totterman, J. Tamez, C. Dabrowski, R. Davies, M. H. Le Graverand, M. Luchi, Y. Tymofyeyev, et al. Change in cartilage morphometry: a sample of the progression cohort of the osteoarthritis initiative. *Annals of the Rheumatic Diseases*, 68(3):349–356, 2009.

[55] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[56] C. Jin, Y. Yang, Z.-J. Xue, K.-M. Liu, and J. Liu. Automated analysis method for screening knee Osteoarthritis using medical infrared thermography. *Journal of Medical and Biological Engineering*, 33(5):471–477, 2013.

[57] C. Kawas, S. Resnick, A. Morrison, R. Brookmeyer, M. Corrada, A. Zonderman, C. Bacal, D. D. Lingle, and E. Metter. A prospective study of estrogen replacement therapy and the risk of developing Alzheimer's disease the baltimore longitudinal study of aging. *Neurology*, 48(6):1517–1521, 1997.

[58] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[59] J. Kellgren and J. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, 16(4):494, 1957.

[60] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989.

[61] M. Kinds, A. Marijnissen, K. Vincken, M. Viergever, K. Drossaers-Bakker, J. Bijlsma, S. Bierma-Zeinstra, P. Welsing, and F. Lafeber. Evaluation of separate quantitative radiographic features adds to the prediction of incident radiographic Osteoarthritis in individuals with recent onset of knee pain: 5-year follow-up in the check cohort. *Osteoarthritis and Cartilage*, 20(6):548–556, 2012.

[62] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[63] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[64] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bulo. Deep neural decision forests. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1467–1475. IEEE, 2015.

[65] V. B. Kraus, S. Feng, S. Wang, S. White, M. Ainslie, A. Brett, A. Holmes, and H. C. Charles. Trabecular morphometry by fractal signature analysis is a novel marker of Osteoarthritis progression. *Arthritis & Rheumatology*, 60(12):3711–3722, 2009.

[66] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, et al. The open images dataset v4:

Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[67] P. Lanyon, S. O'Reilly, A. Jones, and M. Doherty. Radiographic assessment of symptomatic knee Osteoarthritis in the community: definitions and normal joint space. *Annals of the Rheumatic Diseases*, 57(10):595–601, 1998.

[68] M. P. LaValley, S. McLaughlin, J. Goggins, D. Gale, M. C. Nevitt, and D. T. Felson. The lateral view radiograph for assessment of the tibiofemoral joint space in knee Osteoarthritis: its reliability, sensitivity to change, and longitudinal validity. *Arthritis & Rheumatology*, 52(11):3542–3547, 2005.

[69] G. Lester. Clinical research in OA - the NIH Osteoarthritis Initiative. *J Musculoskelet Neuronal Interact*, 8(4):313–314, 2008.

[70] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[71] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[72] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE Transactions on Medical Imaging*, 32(8):1462–1472, 2013.

[73] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, T. F. Cootes, arcOGEN Consortium, et al. Accurate bone segmentation in 2D radiographs using fully automatic shape model matching based on regression-voting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 181–189. Springer, 2013.

[74] S. Linn, B. Murtaugh, and E. Casey. Role of sex hormones in the development of Osteoarthritis. *PM&R*, 4(5):S169–S173, 2012.

[75] C. D. Luna-Gómez, L. A. Zanella-Calzada, M. A. Acosta-García, J. I. Galván-Tejada, C. E. Galván-Tejada, and J. M. Celaya-Padilla. Can multivariate models based on MOAKS predict OA knee pain? data from the osteoarthritis initiative.

In *Medical Imaging: Computer-Aided Diagnosis*, volume 10134, page 1013445. International Society for Optics and Photonics, 2017.

[76] P. C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Science of India*. National Institute of Science of India, 1936.

[77] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems*, pages 2802–2810, 2016.

[78] A. C. Marijnissen, K. L. Vincken, P. A. Vos, D. Saris, M. Viergever, J. Bijlsma, L. Bartels, and F. Lafeber. Knee images digital analysis (kida): a novel method to quantify individual radiographic features of knee Osteoarthritis in detail. *Osteoarthritis and Cartilage*, 16(2):234–243, 2008.

[79] T. McAlindon, S. Snow, C. Cooper, and P. Dieppe. Radiographic patterns of Osteoarthritis of the knee joint in the community: the importance of the patellofemoral joint. *Annals of the Rheumatic Diseases*, 51(7):844, 1992.

[80] S. P. Messier, R. F. Loeser, G. D. Miller, T. M. Morgan, W. J. Rejeski, M. A. Sevick, W. H. Ettinger, M. Pahor, and J. D. Williamson. Exercise and dietary weight loss in overweight and obese older adults with knee Osteoarthritis: the arthritis, diet, and activity promotion trial. *Arthritis & Rheumatology*, 50(5):1501–1510, 2004.

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

[82] L. Minciullo and T. Cootes. Fully automated shape analysis for detection of Osteoarthritis from lateral knee radiographs. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3787–3791. IEEE, 2016.

[83] L. Minciullo, J. Thomson, and T. F. Cootes. Combination of lateral and pa view radiographs to study development of knee oa and associated pain. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013411. International Society for Optics and Photonics, 2017.

[84] M. A. Minor, R. R. Webel, D. R. Kay, J. E. Hewett, and S. K. Anderson. Efficacy of physical conditioning exercise in patients with rheumatoid arthritis and Osteoarthritis. *Arthritis & Rheumatology*, 32(11):1396–1405, 1989.

[85] Y. Nagaosa, M. Mateus, B. Hassan, P. Lanyon, and M. Doherty. Development of a logically devised line drawing atlas for grading of knee Osteoarthritis. *Annals of the Rheumatic Diseases*, 59(8):587–595, 2000.

[86] M. C. Nevitt and D. T. Felson. Sex hormones and the risk of Osteoarthritis in women: epidemiological evidence. *Annals of the Rheumatic Diseases*, 55(9):673, 1996.

[87] H. Oka, S. Muraki, T. Akune, A. Mabuchi, T. Suzuki, H. Yoshida, S. Yamamoto, K. Nakamura, N. Yoshimura, and H. Kawaguchi. Fully automatic quantification of knee Osteoarthritis severity on plain radiographs. *Osteoarthritis and Cartilage*, 16(11):1300–1306, 2008.

[88] B. Oliestad, L. Engebretsen, K. Storheim, and M. Risberg. Knee Osteoarthritis after anterior cruciate ligament injury. *Am J Sports Med*, 37(7):1434–1443, 2009.

[89] G. Peat, R. McCarney, and P. Croft. Knee pain and Osteoarthritis in older adults: a review of community burden and current use of primary health care. *Annals of the Rheumatic Diseases*, 60(2):91–97, 2001.

[90] P. Podsiadlo and G. Stachowiak. Analysis of trabecular bone texture by modified hurst orientation transform method. *Medical physics*, 29(4):460–474, 2002.

[91] A. L. Ratan, W. E. L. Grimson, and W. M. Wells. Object detection and localization by dynamic template warping. *International Journal of Computer Vision*, 36(2):131–147, 2000.

[92] S. Ren, X. Cao, Y. Wei, and J. Sun. Global refinement of random forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 723–730, 2015.

[93] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[94] G. M. Salzmann, S. Preiss, M. Zenobi-Wong, L. P. Harder, D. Maier, and J. Dvorák. Osteoarthritis in football: With a special focus on knee joint degeneration. *Cartilage*, 8(2):162–172, 2017.

[95] N. A. Segal and N. A. Glass. Is quadriceps muscle weakness a risk factor for incident or progressive knee Osteoarthritis? *The Physician and sportsmedicine*, 39(4):44–50, 2011.

[96] L. Shamir, S. M. Ling, W. Scott, M. Hochberg, L. Ferrucci, and I. G. Goldberg. Early detection of radiographic knee Osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage*, 17(10):1307–1312, 2009.

[97] L. Shamir, S. M. Ling, W. W. Scott Jr, A. Bos, N. Orlov, T. J. Macura, D. M. Eckley, L. Ferrucci, and I. G. Goldberg. Knee X-ray Image Analysis Method for Automated Detection of Osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415, 2009.

[98] V. Silverwood, M. Blagojevic-Bucknall, C. Jinks, J. Jordan, J. Protheroe, and K. Jordan. Current evidence on risk factors for knee Osteoarthritis in older adults: a systematic review and meta-analysis. *Osteoarthritis and Cartilage*, 23(4):507–515, 2015.

[99] E. Smith, D. Hoy, M. Cross, T. R. Merriman, T. Vos, R. Buchbinder, A. Woolf, and L. March. The global burden of gout: estimates from the global burden of disease 2010 study. *Annals of the Rheumatic Diseases*, 73(8):1470–1476, 2014.

[100] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[101] T. D. Spector and A. J. MacGregor. Risk factors for Osteoarthritis: genetics. *Osteoarthritis and Cartilage*, 12:39–44, 2004.

[102] G. W. Stachowiak, M. Wolski, T. Woloszynski, and P. Podsiadlo. Detection and prediction of Osteoarthritis in knee and hand joints based on the X-ray image analysis. *Biosurface and Biotribology*, 2(4):162–172, 2016.

[103] I. Starodubtseva. Prevalence of secondary Osteoarthritis in patients with rheumatoid arthritis and risk factors for its progression. *Advances in Gerontology= Uspekhi gerontologii*, 27(4):693–698, 2014.

[104] J. Stern, D. Jeanmonod, and J. Sarnthein. Persistent EEG overactivation in the cortical pain matrix of neurogenic pain patients. *Neuroimage*, 31(2):721–731, 2006.

[105] A. Suárez and J. F. Lutsko. Globally optimal fuzzy decision trees for classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1297–1311, 1999.

[106] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017.

[107] B. Szebenyi, A. P. Hollander, P. Dieppe, B. Quilty, J. Duddy, S. Clarke, and J. R. Kirwan. Associations between pain, function, and radiographic features in Osteoarthritis of the knee. *Arthritis & Rheumatology*, 54(1):230–235, 2006.

[108] J. Thomson, T. O'Neill, D. Felson, and T. Cootes. Automated shape and texture analysis for detection of Osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 127–134. Springer, 2015.

[109] J. Thomson, T. O'Neill, D. Felson, and T. Cootes. Detecting osteophytes in radiographs of the knee to diagnose osteoarthritis. In *International Workshop on Machine Learning in Medical Imaging*, pages 45–52. Springer, 2016.

[110] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific reports*, 8(1):1727, 2018.

[111] I. Tracey. Can neuroimaging studies identify pain endophenotypes in humans? *Nature Reviews Neurology*, 7(3):173, 2011.

[112] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2001.

[113] J. Weese, T. M. Buzug, C. Lorenz, and C. Fassnacht. An approach to 2D/3D registration of a vertebra in 2D X-ray fluoroscopies with 3D CT images. In *CVRMed-MRCAS'97*, pages 119–128. Springer, 1997.

[114] J. Wesseling, M. Boers, M. A. Viergever, W. K. Hilberdink, F. P. Lafeber, J. Dekker, and J. W. Bijlsma. Cohort profile: cohort hip and cohort knee (check) study. *International journal of epidemiology*, 45(1):36–44, 2014.

[115] C. Williams. Efficient mapping of the training of convolutional neural networks to a cuda-based cluster. `http://www.apostherapy.co.uk/en/blog/anatomy-of-the-knee`.

[116] M. Wolski, P. Podsiadlo, and G. Stachowiak. Directional fractal signature analysis of trabecular bone: evaluation of different methods to detect early Osteoarthritis in knee radiographs. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 223(2):211–236, 2009.

[117] M. Wolski, P. Podsiadlo, G. Stachowiak, L. Lohmander, and M. Englund. Differences in trabecular bone texture between knees with and without radiographic Osteoarthritis detected by directional fractal signature method. *Osteoarthritis and Cartilage*, 18(5):684–690, 2010.

[118] R. W. Wright, J. R. Ross, A. K. Haas, L. J. Huston, E. A. Garofoli, D. Harris, K. Patel, D. Pearson, J. Schutzman, M. Tarabichi, et al. Osteoarthritis classification scales: interobserver reliability and arthroscopic correlation. *The Journal of bone and joint surgery. American volume*, 96(14):1145, 2014.

[119] T. K. Yoo, D. W. Kim, S. B. Choi, and J. S. Park. Simple scoring system and artificial neural network for knee Osteoarthritis risk prediction: a cross-sectional study. *PloS one*, 11(2):e0148724, 2016.

[120] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.