ALGORITHMS FOR AUTOMATIC ANALYSIS OF RADIOGRAPHS OF THE KNEE WITH APPLICATION IN DIAGNOSIS AND MONITORING OF OSTEOARTHRITIS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2016

Jessie Thomson
School of Health Sciences

Contents

Word count 29191

List of Tables

List of Figures

The University of Manchester

Jessie Thomson
Doctor of Philosophy
Algorithms for Automatic Analysis of Radiographs of the Knee with Application in Diagnosis and Monitoring of Osteoarthritis
December 29, 2016

Osteoarthritis (OA) of the knee is a disease that deteriorates the bones and surrounding soft tissue of the affected joint. Categorisation of the disease into grades of severity is subject to errors of measurement and poor observer agreement. There is an urgent need for automated methods to measure radiographic features and remove, as far as possible, the element of subjectivity in assessment. This project creates a fully automated system to analyse all aspects of the knee in radiographs. The methods evaluate explicit and implicit features of: overall shape, trabecular structure, osteophytes, tibial spines and intercondylar notch, and joint space shape. The project develops the first fully automated osteophyte detection algorithms, improved trabeculae features using raw pixel intensities, and a better analysis of joint space using shape models.
This project is the first to combine explicit and implicit features across the whole of the knee, and applies these features to classify radiographs using four main outcomes: current OA, current pain, later onset OA, and later onset pain. The results find a strong current OA classification rate, with an Area Under the ROC Curve (AUC) of 0.904 and weighted kappa of 0.49 (0.48-0.51). The remaining later onset and pain experiments report weaker results; these results suggest that radiographic features in Posterior-Anterior (PA) view radiographs have a weak association with clinical and later onset OA.

Declaration

No portion of the work referred to in the 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 Statement

Acknowledgements

I would like to thank my supervisor Tim Cootes for the enormous amount of help and support across the project, and for all the helpful advice on maths, computing and job related matters that have helped me reach the position I am now. I would like to thank my co-supervisors, David Felson and Terence O'Neill, for all the help and guidance - especially in the medical and clinical aspects of the project. I would also like to add an extra thanks to Matthew Parkes for the invaluable statistical guidance, and Claudia Lindner for help with the direction and motivation of my experiments.

I would also like to thank my partner Stephen Lacy and my mum, Gill Thomson, for the emotional support and life advice that kept me going. My family for their unending support in all my life choices and struggles, and my friends for putting up with my incessant obsession with work and writing.

Abbreviations

AAM Active Appearance Models

ASM Active Shape Models

AUC Area Under ROC Curve

AVOT Augmented Variance Orientation Transform

BMI Body Mass Index

BML Bone Marrow Lesion

CDF Cumulative Distribution Function

CI Confidence Interval

CLM Constrained Local Model

EMD Earth Mover's Distance

FD Fractal Dimension

FSA Fractal Signature Analysis

JS Joint Space

JSA Joint Space Area

JSN Joint Space Narrowing

JSW Joint Space Width

KIDA Knee Images Digital Analysis

KL Kellgren-Lawrence

KOACAD Knee Osteoarthritis Computer-Aided Diagnosis

kw weighted kappa

LBP Local Binary Patterns

LDA Linear Discriminant Analysis

mHOT modified Hurst Orientation Transform

MI Mutual Information

mJSW minimal Joint Space Width

MRI Magnetic Resonance Imaging

OA Osteoarthritis

OAI OsteoArthritis Initiative

OARSI Osteoarthritis Research Society International

OR Odds Ratio

OS osteophytes

PA PosteroAnterior

PCA Principal Component Analysis

PGD Principal Gradient Direction

RF Random Forest

RFCLM Random Forest Constrained Local Model

ROC Reciever Operating Characteristic

ROI Region Of Interest

RPR Raw Pixel Ratios

SDM Signature Dissimilarity Measure

SOM Self Organising Map

SSM Statistical Shape Model

stdev. standard deviation

VOT Variance Orientation Transform

Nomenclature

α Constraint on the DP fitted contours

δ_n Scale of gaussian sampling region in SDM texture analysis

r Vector projected between two shape points

T0 Data at the time point of incidence

T1y Data from 1 year prior to incidence outcome

T2y Data from 2 years prior to incidence outcome

T5y Data from 5 years prior to incidence outcome

C_{xy} Circular pixel sampling region centred at pixels (xy) used in FSA methods

N_{di} Number of pixels along each angle θ_i in region C_{xy}

N_θ The number of angles θ in the region C_{xy}

$R(\theta, d)$ Table of intensity differences along distances N_d per N_θ angles

xJSW Vector of JSW measurements across medial and lateral joint space

About the Author

The author, Jessie Thomson, completed a BSc in Computer Games Programming at Huddersfield University in 2012, and an MSc in Advanced Computer Science at the University of Manchester in 2013. The PhD was started in 2013, and during the three years the author has published and co-authored the folowing papers relevant to the project:

J. Thomson, T. O'Neill, D. Felson and T. F. Cootes, "Automated shape and texture analysis for detection of Osteoarthritis from radiographs of the knee", Proc. MICCAI 2015, Part 2, pp.127-134

J. Thomson, M. Parkes, D. Felson, T. O'Neill and T. F. Cootes, "Automated multi-feature analysis of current and future onset pain in osteoarthritic knees", Int. Workshop on Osteoarthritis Imaging, (abstract).

J. Thomson, T. O'Neill, D. Felson and T. F. Cootes, "Detecting osteophytes in radiographs of the knee to diagnose Osteoarthritis" Proc. MICCAI MLMI 2016 (to appear).

Under review: L. Minciullo, J. Thomson, T. F. Cootes, "Combination of Lateral and PA View radiographs to Study Development of Knee OA and Associated Pain", SPIE 2016.

In preparation: J. Thomson, M. Parkes, D. Felson, T. O'Neill and T. F. Cootes, "Automated Radiographic Multi-feature Analysis of Current and Future Onset OA and Pain", Osteoarthritis and Cartilage

Chapter 1

Introduction

1.1 Motivation

Osteoarthritis (OA) is a degenerative disease in which bones and surrounding soft tissue of the affected joint deteriorate. The knee is one of the most commonly affected sites. Knee OA affects approximately one in five of the UK population (18.2%) with the frequency increasing with age [1]. Due to a demographic shift towards an older population the number of people affected by knee osteoarthritis is set to increase significantly over the next 50 years. Symptoms of the disease include pain, stiffness, occasional swelling, and loss of function. There are currently no treatments which slow progression of the disease.

Characteristic features of Osteoarthritis of the knee on plain radiographs include narrowing of the joint space, thickening of the joint line (bone sclerosis) and new bone formation at the joint margin (osteophytes). Magnetic Resonance Imaging (MRI) provides additional information on knee structures including involvement of the lining of the joint (synovium) and other soft tissue structures. OA may be detected based on these features using semiquantitative approaches, into distinct categories or grades (typically normal, doubtful, minimal, moderate and severe) [2] [3] [4]. The cost and availability of plain radiographs (compared to MRI images) make x-rays the most common imaging modality used in clinical and research settings. Classification criteria based on categorisation into grades is, however, subject to errors of measurement and poor observer agreement. Imprecision when assigning a grade is compounded when

looking at structural change over time, as is the case in both prospective studies and clinical trials. Such errors result in misclassification and make it more difficult to detect change when it really is present. The consequence of this is that for observational studies and clinical trials a larger number of people need to be recruited in order to detect change.

There is an urgent need for automated methods to measure radiographic features and remove, as far as possible, the element of subjectivity in assessment. Osteoarthritis has no cure, but the development of improved methods for detecting and analysing OA will improve understanding of disease development and applying new treatments that may slow or prevent progression of the disease. Further to this a precise system will help analyse the effect of treatments and determine an improvement or worsening of radiographic OA features. Automated methods apply a standardised set of rules to evaluating radiographic features of OA. Current automated systems tend to focus on singular aspects of the disease [5] [6] [7] [8] or implicitly capture OA features [9] [10]. These methods primarily focus on later stages of OA development, with minimal application to early features and taking account of clinical symptoms [6] [11] [12].

In the project we have created a fully automated system to analyse all aspects of the knee in radiographs. The methods evaluated explicit and implicit features of OA and find the optimal combination of features to assess current and later onset disease development. The project was extended to analyse the association of radiographic features with the clinical assessment of pain in the respective knee.

1.2 Aims and Objectives

The main aim of the project was to create a system to automatically analyse the bones of the knee from radiographic images, allowing precise measurements of change across the joint. The development of this was done using a state of the art algorithm for initially detecting the knee, and then the application of various feature extraction methods to analyse implicit and explicit features across the whole joint. The main objectives to meet this aim were:

1. Create an automated system to detect OA or non-OA from radiographs of the knee.

    - This was done by analysing various radiographic OA features, such as: overall bone shape, osteophytes, trabecular structure, tibial spines and intercondylar notch, and Joint Space Narrowing (JSN).

2. Explore the correlation between radiographic features and later onset OA.

    - Radiographic features were extracted from non-OA baseline images and used to predict participants that later develop OA at any point in the follow-up visits.

3. Analyse current and later onset pain reported by the participants

    - Similar to objective 2, the feature analysis was applied to detect current non-painful and painful images, and later onset painful knees from non-painful baseline images.

1.3 Contributions

The project developed a fully automated system to analyse all aspects of knee OA. During the project a number of contributions to knowledge were made, these were:

- A fully automated analysis of radiographic knee images that analysed both implicit and explicit features, and combined the features to strengthen the detection accuracy of both OA and pain assessments.

- A novel technique for analysing trabeculae texture was developed. The method achieves a better detection of current OA than other state-of-the-art trabeculae texture analysis algorithms.

- A novel fully automated osteophyte analysis was developed using combined shape and texture information. The algorithm achieves a higher accuracy than the semi-automated and osteophyte area algorithms reported in the literature.

- An improved measure of joint space change using shape models that achieves better accuracy (in all current and later onset experiments) than the commonly used xJSW features.

1.4 Outline of the Thesis

The next chapter (Chapter 2) presents an overview of the relevant literature. The review describes Osteoarthritis and the effects of knee OA, current manual and automated grading methods, and an introduction to the state-of-the-art algorithm used to detect the knee and localise the feature extraction methods. Chapter 3 describes the feature extraction methods, the data used throughout the experiments (Chapters 4-6), and the techniques used to evaluate the features. The experiments chapters (Chapters 4 and 5) compare and evaluate the feature extraction methods in comparison to other methods from Chapter 3 and features provided in the dataset.

Chapter 6 uses the best features from the previous two experiments chapters and expands the data in each of the four experiments: current OA, current pain, later onset OA, and later onset pain. The results are compared to the manual Kellgren-Lawrence grades provided in the dataset.

We conclude the thesis in Chapter 7 and discuss limitations and future extensions of the project.

Chapter 2

Literature Review

The chapter provides an overview of the structure and function of the knee, occurrence of Osteoarthritis (OA) of the knee, underlying causes and structural features of the disease, and symptoms linked with Osteoarthritis. The focus of the project is quantitative imaging of the knee; the literature relating to imaging of knee OA is summarized including manual grading techniques, past and current automated methods, and concludes with a section exploring the best reported object segmentation methods.

2.1 Structure and Function of the Knee Joint

The knee constitutes one of the largest joints in the body, and comprises an articulation between three bones: tibia, femur and patella (knee cap). The bones are held together by collateral ligaments which attach to the tibia and femur. The anterior and posterior cruciate ligaments which run from the lower femur to the upper tibia provide joint stability (see Fig. 2.1). The bones at the joint are lined by a protective layer of hyaline cartilage. The joint space (JS) between the tibia and femur is lined by a layer of synovium which is responsible for secretion of viscous fluid to reduce friction. Two crescent shape cartilages (menisci) are attached to the medial and lateral tibia and help stabilize the joint. The joint takes the weight-bearing stress during upright movement, and allows flexion, extension and some medial and lateral rotation [13].

The bones in the knee are made up of two types of material: cortical and trabecular

(see Fig. 2.2). Cortical bone, a dense and compact tissue, forms 80% of the bone mass of the human body. Cortical tissue makes up the cortex (outer) layer of bone, and is integral for support and protection of the organs in the body. Trabecular bone is a cancellous (mesh-like) structure situated within the cortical envelope. The trabecular structure has a higher surface area to weight ratio and provides support for the skeleton [14].

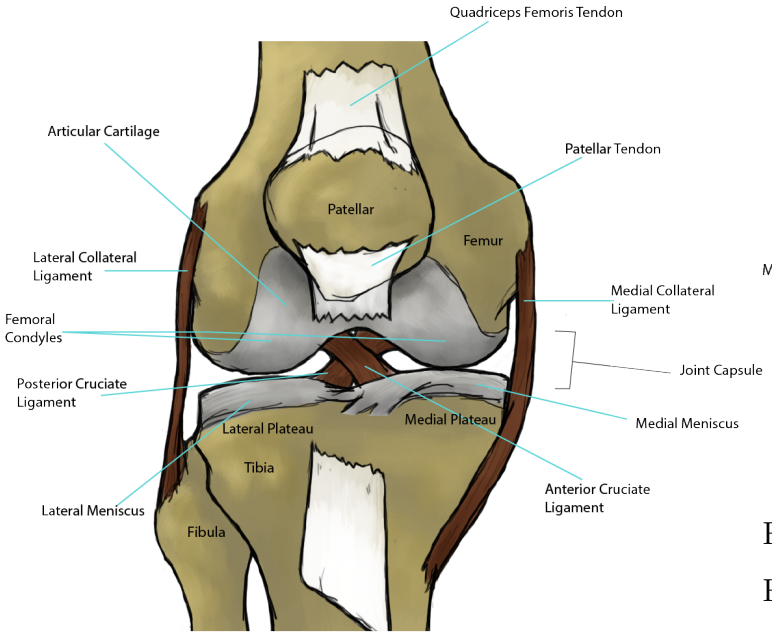

Figure 2.1: Anatomy of the knee joint.

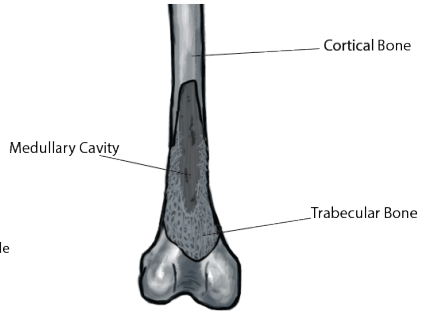

Figure 2.2: Structure of the Femur.

2.1.1 Bone Remodelling

Bone is a living tissue and continuously being renewed, often in response to physiological, environmental and mechanical triggers. It is a normal process to maintain the strength and integrity of bone, whereby mechanosensors within the bone influence remodelling to repair or distribute structure to prevent concentrations of pressure [15]. Bone modelling may be influenced by mechanical stress and bone-related disease pathology (i.e. during Osteoarthritis) [16] [17], creating malformed and enlarged bones around the afflicted areas.

2.2 Osteoarthritis of the Knee

OA of the knee is a disease in which the normal structure of the joint is disrupted with loss of cartilage and bony remodelling resulting in new bone formation (osteophytes) and thickening of bone (sclerosis) at the joint margins. Ultimately this results in joint failure. Originally the disease was considered to be primarily due to a loss of hyaline cartilage though it is now recognised as a disease of the whole joint. Symptoms of the disease include pain, stiffness, occasional swelling, and loss of function. Osteoarthritis of the knee is one of the most frequent causes of disability in the UK affecting just under one in five of the UK population (18.2%) [1]. Due to a demographic shift towards an older population the numbers of people with Osteoarthritis of the knee are set to increase significantly over the next 20 years.

2.2.1 Causes

The causes are split into two groups: primary (caused either by genetic factors or in circumstances without a clear causative mechanism), and secondary (caused as a factor of another disease or mechanical strain).

Environmental and mechanical factors are linked with an increased risk of developing Osteoarthritis, these include: genetics, age, BMI, gender, hormones, race and ethnicity, joint injury, repetitive overloading of joints, joint deformity and periarticular muscle weakness. These risk factors vary between participants, with some knees developing OA with no factors present. This makes diagnosis and prevention difficult to target specific risks or even understand the true cause of any one person's disease development.

Recent findings have found particular subgroups in the disease pathomechanics and the risk factors associated. Whereby pathomechanics are the alterations to the normal function and response of the knee caused by Osteoarthritis. The work of Waarsing et al. [18] found four different subgroups of radiographic OA relating to different risk factors, with the subgroups illustrating different rates of OA feature progression, disease severity and clinical symptoms.

2.2.2 Features of Knee Osteoarthritis

Osteoarthritis is a disease of the whole joint. Structural features of the disease include: Joint Space Narrowing (JSN), osteophytes, denudation and attrition of articular surfaces, cysts, sclerosis, orientation changes and thickening of underlying trabeculae, Bone Marrow Lesions (BML), and inflammation of the synovium. These features are linked to mechanical stress [19]. As OA progresses, mechanical alterations can occur, which further the biomechanical response of the knee. This results in altered mechanical loads and further disease progression. The project focuses on features visible in radiographic images: osteophytes, JSN, bone attrition, joint alignment, and trabeculae changes. The Figure 2.3 below describes the regions (referenced in later sections) typical for these features to occur in PosteroAnterior (PA) radiographs. The terms medial and lateral are used frequently in the text, these describe the areas of the knee (see Figure 2.4), with the medial sides facing the centre (central plane) of the body and the lateral side facing away from the centre. This orientation is consistent throughout the thesis.

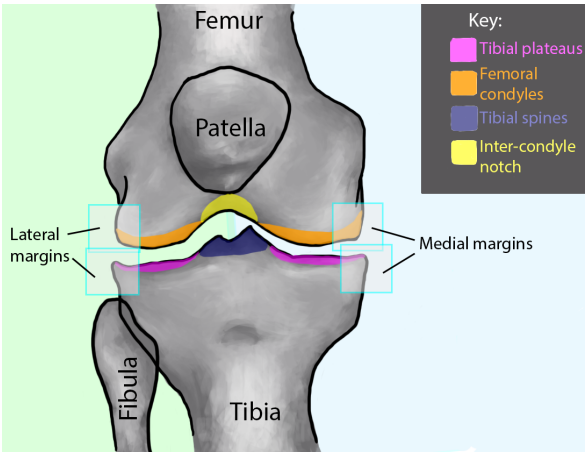

Figure 2.3: Reference regions on an AP right knee.

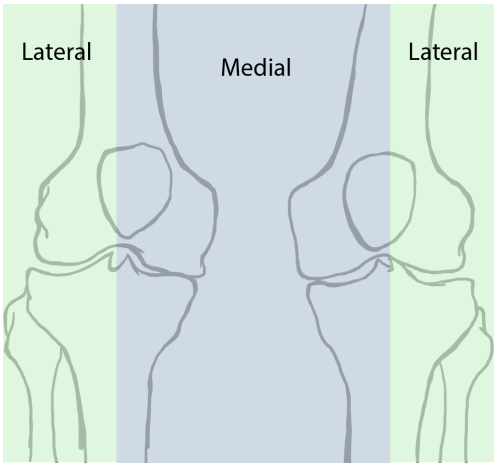

Figure 2.4: Reference medial and lateral regions on left and right knees.

Osteophytes

Osteophytes are bony spurs that form from the articular cortical surface of the bone, typically around the joint margins with some formation in the inter-condylar notch. The osteophytes are reported to be a pro-inflammatory reaction to damage of the

cartilage and tendons [20]. They occur throughout OA development (Figs. 2.5-2.8), and often in later stages of the disease multiple large osteophytes (Fig. 2.9) will be found.

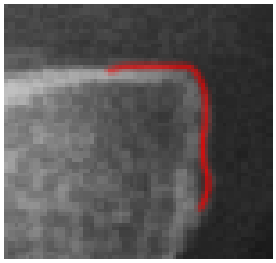

Figure 2.5: Normal tibial margin

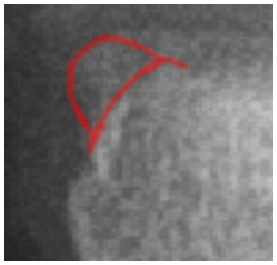

Figure 2.6: Mild osteophyte

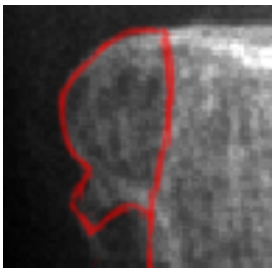

Figure 2.7: Moderate osteophyte

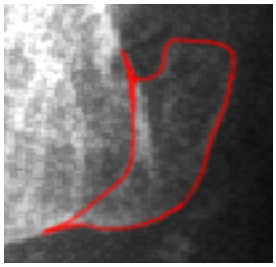

Figure 2.8: Severe osteophyte

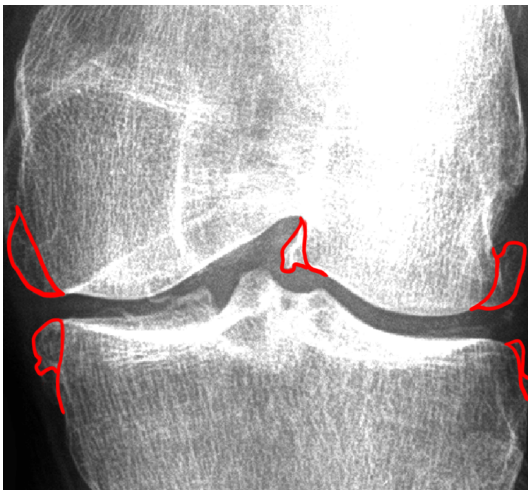

Figure 2.9: Multiple severe and moderate osteophytes

Bone Sclerosis

In the early stages OA bone resorption increases and there is thinning of the subchondral bone. In later disease stages there is lower resorption and increased bone formation, resulting in thickened low mineral bone (sclerosis) at the bony articular surface (Fig 2.10), and some widening and flattening of the plateaus (attrition) (Fig. 2.11).

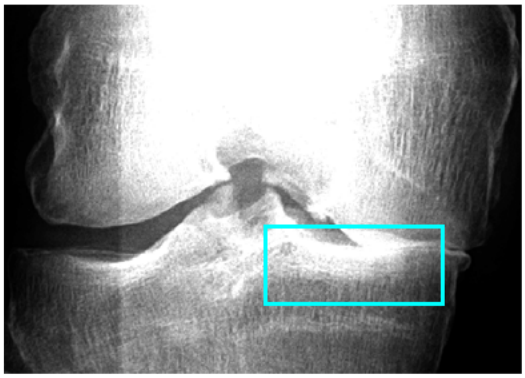

Figure 2.10: Sclerosis seen beneath blue region

Figure 2.11: Red arrows show the movement of attrition and the pocket formed in the medial plateau

Joint Space Narrowing

Joint Space Narrowing (JSN) occurs as a result of cartilage degradation [21] and develops during more advanced stages of the disease. Tears in the menisci or meniscal extrusion may also be linked with joint space narrowing [22]. The medial compartment is typically more likely to be affected by cartilage loss and JSN than the lateral compartment (Figures 2.12-2.15). As the disease progresses, cartilage loss and JSN increase (see Fig. 2.15).

Figure 2.12: Normal joint space

Figure 2.13: Mild narrowing

Figure 2.14: Moderate narrowing (lateral)

Figure 2.15: Severe narrowing

Joint Mal-alignment Another factor of OA which is linked to JSN, is joint mal-alignment. This occurs when the tibia and femur begin to angle outside the typical amounts, away or towards the central plane (illustrative vertical line dissecting the body lengthways) from the inflection point (knee joint) [23]. This is often caused by compartment JSN, with the tibia and femur bones shifting from the acute angle formed in the joint space. This mal-alignment is termed either varus (angulation away from the central plane and caused by medial JSN) or valgus (angulation towards the central plane and caused by lateral JSN) [24]. Examples of both can be seen in the Figures 2.16 and 2.17 below.

Figure 2.16: Varus alignment of the right knee.

Figure 2.17: Valgus alignment of the right knee.

Trabeculae

In OA affected knees, trabeculae under the weight-bearing subchondral surfaces of the tibia will thicken and alter direction in conformity with Wolff's Law [25], acting to alleviate the strain caused by focal pressure. The trabeculae are strong to compression along the typical direction (moving down the leg on a normal aligned joint) but are weak to tension forces. Through the shift in weight and pressure, the horizontal trabeculae start to thicken to handle the extra tension forces during early OA [26] [27] [28] and more vertically orientated trabeculae appear in the later stages to handle the increased pressure from complete joint loss (see Figure 2.15 above).

Tibial Spines

The tibial spines become rough, spiked and irregular in shape (see Figs. 2.18-2.19). These changes are caused as a result of the formation of new bone across the knee.

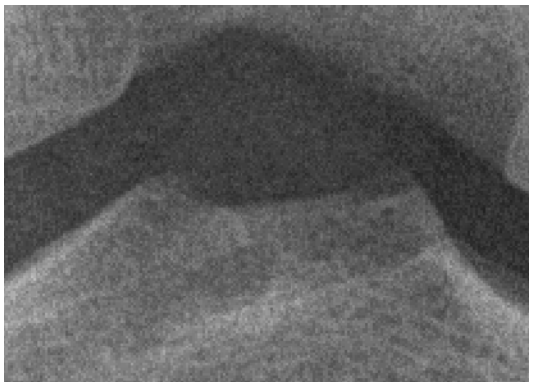

Figure 2.18: Normal tibial spines.

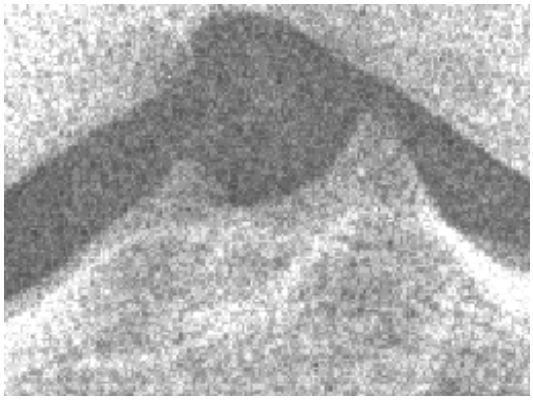

Figure 2.19: Tibial spines during OA.

Pain and Structural Changes

One of the main factors clinicians base treatment of Osteoarthritis on is the symptoms, particularly the pain associated with the disease. There is significant discordance between symptoms and radiographic change in knee OA with some relatively unaffected joints experiencing severe pain, and some with severe OA having relatively little or any pain. These discrepancies have been attributed to the subjective nature of pain and disability, as well as other environmental and mechanical influences outside the disease.

There is evidence of an association between radiographic features of the disease and pain scores. The work in [6] found a positive association between Joint Space Width (JSW) and pain in the respective joint, this is supported by [22] who found a correlation between extended pain severity and an increase in JSN. Similarly [29] found an association between the angle of mal-alignment in the knee and average pain reported by the participant.

These findings are contradicted by studies that looked at a broader range of features, comparing JSN, osteophytes and overall disease grade to the level of pain in the respective joint. Shin and Lee [20] and Cicuttini et al. [30] found osteophytes detected and predicted pain better than JSN. The study on tibial spines [31] found some weak association with spike angulation and pain. Whereas the study by [32] shows that combining multiple radiographic features creates a stronger prediction for current and later onset chronic pain.

2.2.3 Management of Knee Osteoarthritis

Current management of the disease focuses on reducing pain and functional impairment. There are a range of pharmacological and non-pharmalogical interventions (including nutriceutical and orthotics) which may be used. Lifestyle advice, such as losing weight and taking regular exercise, are also important. There are currently no treatments that have been shown to slow down progression of the disease.

2.2.4 Grading of Knee Osteoarthritis

Different grading systems have been developed to classify features and severity of OA. The main purpose of these criteria is to facilitate comparison of results between studies and enhance understanding of the causes and treatment of the disease. The different methods typically comprise a set of numerical stages of increasing severity; some focus on individual structural features while others focus on composite features. Methods that use radiographs focus on measures of the overall shape and texture change across the bone, whereas MRI methods tend towards analysing soft tissue and 3D bone properties. MRI gives much more information than radiographs, however the process is costly and labour intensive. Radiographs are a faster and cheaper alternative and

the typical imaging modality of choice used by clinicians and large sample research trials.

2.3 Manual Grading

Manual grading methods can be split into two groups: quantitative, where the grading makes use of specific measurements of the Osteoarthritic features; and semi-quantitative, that makes use of comparing radiographs against typical representations of the different grades. The commonly used methods for each are: Kellgren-Lawrence [2] grading and atlas grading methods [3] [33] for semi-quantitative methods; and Ahlback grading [34] for quantitative methods.

2.3.1 Semi-Quantitative

Semi-quantitative grades are split into composite scorings, where the stages of the disease are detailed in a combination of OA features; and individual scorings, which have separate scales for each feature.

Composite Scoring

The most widely used composite OA grading is the Kellgren-Lawrence (KL) method [2], which splits disease development into five classes: normal (KL0), doubtful (KL1), minimal (KL2), moderate (KL3) and severe (KL4). Onset of the disease is usually taken to be KL2 and above (see Table 2.1 for details of each grade). KL grading is performed through visual inspection, comparing the signs in the radiographs to the documented features. The reliance on experience and training can make the grading susceptible to subjective views of the observer. This is shown through only moderate inter-observer variability, especially when distinguishing between the central grades (KL 1-3). Inter-observer variability is shown in papers reporting weighted kappa (kw) in the range 0.36 - 0.8 [35], and discrepancies between observers reaching 0.41 kw [36], where 0 means agreement equivalent to chance, and 1.0 perfect agreement.

Some issues of variability in the KL grading have been improved by altering the detail of the stages to make the classifications more distinct [37] [38]. Felson et al. [37] uses an

Grade	Description
0 - No disease	No evidence of disease (see Figure 2.20)
1 - Doubtful	Possible JSN and osteophytes at the subchondral edges (see Figure 2.20)
2 - Mild	Visible osteophytes and possible JSN (see Figure 2.21)
3 - Moderate	Multiple osteophytes, definite JSN with possible deformity of the bone (see Figure 2.21)
4 - Severe	Large osteophytes, JSN and definite bone deformity around the joint (see Figure 2.22)

Table 2.1: Table of the Kellgren-Lawrence grades and descriptions

altered Kellgren-Lawrence scale that defines OA incidence at grade 2 as having both Joint Space Narrowing and osteophyte development, but then also splits the grade to classify joints that only show osteophyte development. Whereas, Brandt et al. [38] removed the occurrence of osteophytes, using only Joint Space Narrowing to grade the radiographs.

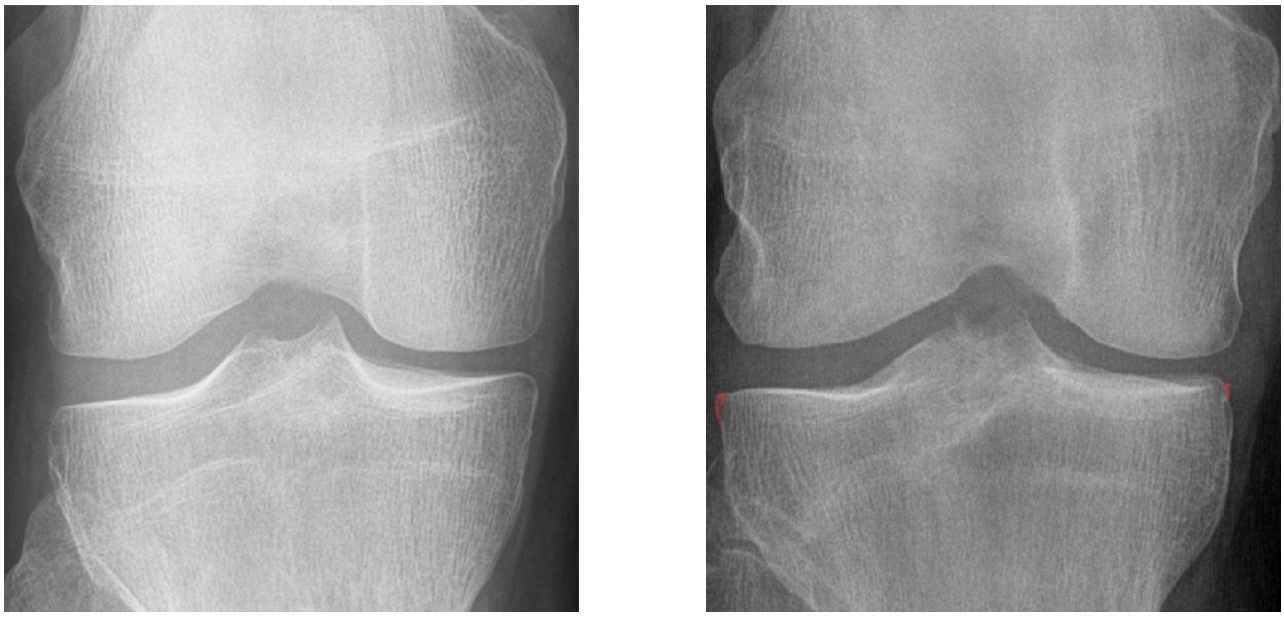

Figure 2.20: KL0, no signs of OA (left image) and KL1, possible osteophytes on tibia (right image).

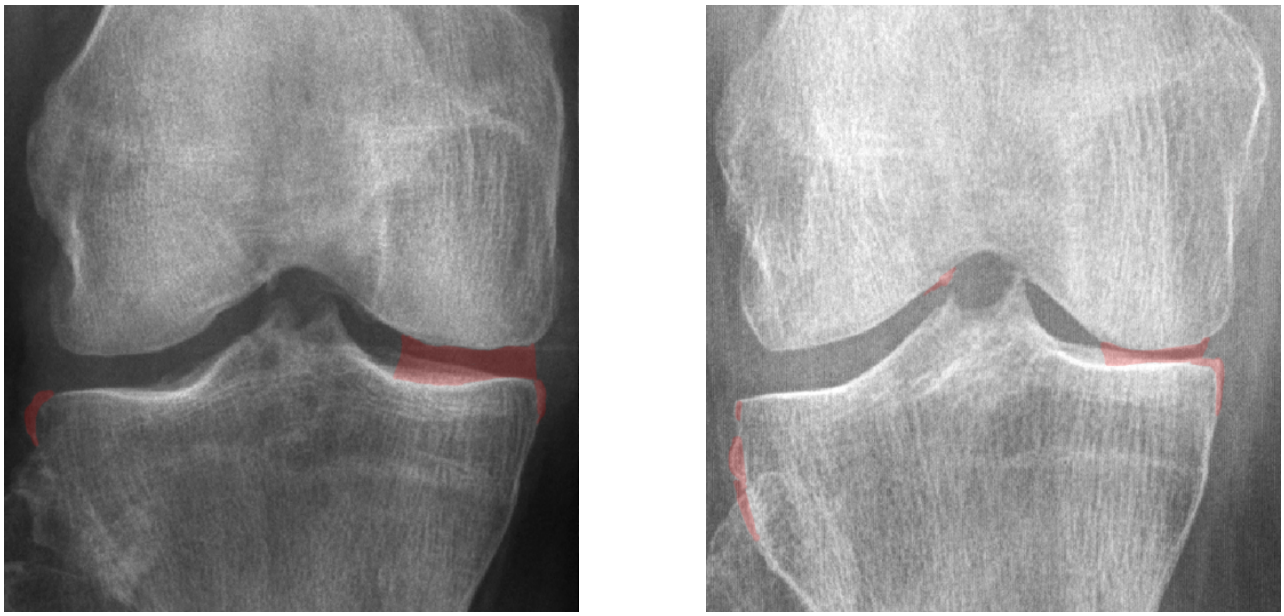

Figure 2.21: KL2 osteophyte on lateral tibial margin, possible medial JSN (left image) KL3 definite medial JSN, multiple small osteophytes on lateral tibia and inter-condyle notch (right image).

Figure 2.22: KL4 large osteophytes on lateral (femur and tibia) and medial margins (tibia only), medial JSN and definite bone deformity on lateral side of tibia.

Individual Scoring

The individually scored methods include the work of Altman et al. [33] and Nagaosa et al. [3], both of which make use of an atlas describing the grading of each feature

on a series of images. The OARSI [33] atlas uses a series of radiographic images of the different features of Osteoarthritis. Whereas, the Line Drawing Atlas [3] uses a simplified outline of the OA features to help categorize the development of OA. The atlas maps the bone changes across the knee, focusing on Joint Space Narrowing (JSN) and osteophyte growth in each individual compartment of the knee. The development is split into four stages: normal (0), mild (1), moderate (2) and severe (3). The reliability across a comparison of extended line atlas grades achieved a high weighted kappa and Confidence Interval (CI) of 0.86 (CI 0.85-0.87) for JSN, and 0.78 (CI 0.77-0.79) in osteophyte development [39]. This is similar to the osteophyte inter-observer reliability kw 0.72 (CI 0.64-0.8) in [40]. The OARSI grades can vary between observers, however, with some studies reporting OARSI atlas inter-observer kw as 0.48 (JSN) and 0.64 (osteophytes) [41]. The use of atlas methods often takes longer in distinguishing grades because the images need to be compared to each representative image in the atlas' set of features.

2.3.2 Quantitative Methods

Quantitative methods focus on the exact measurements of OA features in radiographs. The work of Ahlbäck [34] uses the specific measurement of JSN and attrition, where the latter grading scores are influenced by the amount of growth (in millimetres) the tibial plateaus advance over the course of OA. Despite being a relatively simple measurement with few confounding factors, the reliability of the method was found to be very low, with a weighted kappa of 0.23 [42].

Older methods, such as the work of [43] and [44] measured features of osteophytes, sclerosis and JSN by hand using a pair of dividers and a magnifying glass. The work of Lane et al. [43] measured individual features of the disease for hands, hips and spines. Focusing on specific disease features occurring in each with inter-observer reliability ranging from 0.42 for sclerosis and 0.66 for osteophytes, to 0.8-0.93 for minimal JSW (mJSW). The method split the quantitative measures into stages according to the severity of the disease per feature. A similar method was conducted in [44] with the individual features then being combined into one overall assessment of severity based on all features. The inter-observer reliability (kw) improved to 0.74 knees, 0.73 hips and 0.74 hands. However, these methods relied heavily on the exact magnification,

quality, flexion and rotation of the knees in the radiographs, and did not account for discrepancies between mJSW sites the observers measured [45].

2.3.3 Summary

Manual methods are useful in giving a visual representation of the grades and Osteoarthritic features to be compared and allow flexibility of feature development during progression.

Despite this, manual grading methods are time consuming and have problems consistently producing reproducible results. Reproducibility problems arise from generalisations of features (semi-quantitative measurements); or combinations of features to define grades, removing precise definitions relative to the varied OA features. These problems often lead to differing standards applied to radiographs between studies and clinical sites. This proves problematic when coallating data and transferring participants between clinics.

Quantitative measurements solve these problems by finding distinctions in severity based on precise distance and size measurements, however, these methods are often time consuming and require clear, un-rotated x-rays to avoid erroneous measures and miss-classifications of the disease [45].

The application of automated mathematical measurements is a technique for avoiding these problems. The early work of Buckland et al. [46] used magnified radiographs to accurately measure the progression of an arthritic disease, calculating joint space and the margin of bony erosions using an automated system and a series of projected lines. This quantitative method cultivated various semi-automated projects [45] [47] [48]. The system by Dacre et al. [48] sought to semi-automate the process via overlaying graph lines over the knee radiographs, measuring the Joint Space Area (JSA) and Joint Space Width (JSW) in the tibial-femoral space in the knee. These methods have been noted to be cumbersome and inefficient [45], due to the old technology. However, the general principle of automating the analysis and measurement of Osteoarthritis opens up the concept of systematically detecting and predicting OA development. Utilising a system that precisely and automatically measures the knee and the Osteoarthritic features, allows for further study into the pathogenesis of the disease and its progression [45] [47] [48] [6][49].

2.4 Automated Methods

Since the early automated methods [45] [47] [48], there have been considerable advancements in the area of automated disease analysis. Complex algorithms can be applied to analysing object features in radiographic images. These methods typically focus on specific features found correlated to disease progression from manual grading methods [2] [33]. Due to the nature of radiographs, these features are found through analysing shape and texture properties in the 2D images.

This section is split into the three core testing outcomes for knee OA in radiographs: methods that detect current OA (cross-sectional OA analysis), methods that analyse the clinical symptoms of pain in OA (pain detection), and methods that predict later onset OA (longitudinal OA prediction).

2.4.1 Cross-Sectional Osteoarthritis Analysis

Cross-sectional analysis focuses on the current level of Osteoarthritis in the knee. The methods focus on specific OA features, combinations of features and in a few examples, overall shape and texture containing implicit signs of OA.
The following section is split into the various OA feature models and combinations of them which have been developed.

Trabeculae

The trabeculae are parts of the underlying cancellous bone structure (see Section 2.2.2) and can be seen in the subchondral areas of the joint (beneath the tibial plateau and above the femoral condyles) as a series of thin uniform lines. The earliest method to find a correlation between trabeculae changes and OA [50] measured intensity change across the trabeculae by scanning individual horizontal and vertical lines across regions of interest (ROI) in the radiographs.
From this various other semi-automated methods analysed the variation in the trabeculae, finding an increase in volume of trabeculae in the OA affected compartment [51] and a decrease in horizontal trabeculae during early OA that progresses to thicker more compact horizontal trabeculae with increased vertical trabeculae [27] [26] in later

OA development. Eventually, the method of Fractal Signature Analysis (FSA) was linked to the variation in thickness and orientation in the underlying trabeculae.

Fractal Signature Analysis The fractal signature of an image is a measure of the underlying 3D structure, or Fractal Dimension (FD), which in trabecular structure measures the number, spacing and cross-connectivity of the trabeculae [52]. The work of Pentland [53] and Peleg et al. [54] brought about applying FD to analysing image texture. These methods were based off Mandelbrot's calculations of fractal dimensions in nature, whereby the fractal dimension of an object, or 'roughness' of a surface, could be measured via taking the pixel intensity differences at varying scales of the image. The main feature for calculating fractal dimension, is that the structures must have the properties of a Brownian fractal [26]. These properties are: 1) the pixel differences across varying scales must be normally distributed, and 2) a line can be fitted to a log-log plot of the intensities across the scales. These properties were found to be true for trabeculae seen in radiographs, under the cortical surface (see Fig. 2.23).

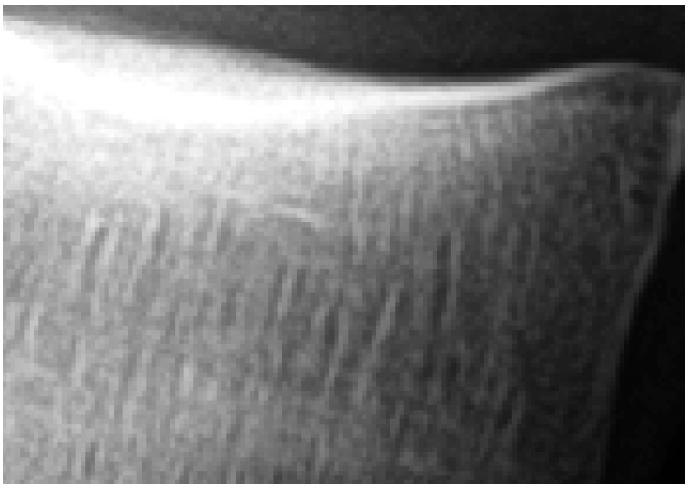

Figure 2.23: Trabeculae seen under the cortical surface of the tibial plateau.

The simplest FD calculation is using a box-counting method. This calculates the FD placing a grid over the object and counting the number of boxes that contain part of the structure (n = number of boxes that contain the structure at the specific scale). The grid squares are then scaled down so there is double the number within the same area and the process is repeated. Once finished the n for each iteration is plotted in a log-log plot and a line of best fit projected through the points. The FD is equal

to $1-\beta$ where β is the slope of the line fitted to the data. The work in [55] uses a 2-D box counting method (measuring horizontal and vertical trabeculae within the knee) to predict progression of OA. The paper achieves a fairly good Area Under the ROC Curve (AUC) with 0.75 (CI 0.65 - 0.84). The single FD limits the accuracy of the algorithm as the direction of the trabeculae has been found to be an important factor in OA development [27] [26], calculating fixed angles will potentially miss slight changes throughout the progression of OA. FD calculations progressed to analyse all directions of trabeculae in the modified Hurst Orientation Transform (mHOT) method [56].

The mHOT method finds the greatest intensity difference (between pairs of pixels) in every direction across a ROI. The method samples a circle region C_{xy} across the patch of image and taking the absolute difference between the central pixel intensity $I(x,y)$ and all other pixels within the region $I(x_{ij},y_{ij})$, such that $(x_{ij},y_{ij}) \neq (x,y)$, $(x_{ij},y_{ij}) \in \mathbf{C}_{xy}, i = 1,2,...N_\theta, j = 1,2,...N_{di}$. Where N_θ is the maximum number of angles being measured, this is determined by selecting the angles with ≥ 4 pixels, and N_{di} is the number of distances along angle θ_i (see Fig. 2.24). The absolute differences are stored in a table $R(\theta,d)$ splitting the values by angle θ_i, taken as the angle between a vector connecting both pixels and the image horizontal axis, and the distance d_{ij} between the pixels. The region $\mathbf{C}_{xy}$ is then shifted to a new centre and any absolute differences that are larger than the corresponding values stored in $\mathbf{R}(\theta,d)$ are replaced. Once the image has been scanned, a log-log plot of each row of the table is plotted and the Hurst coefficient, equal to $H = \beta/2$ where β is the slope of the line fitted to the plot points, is then related to the Fractal Dimension is by the equation FD$= 3 - H$.

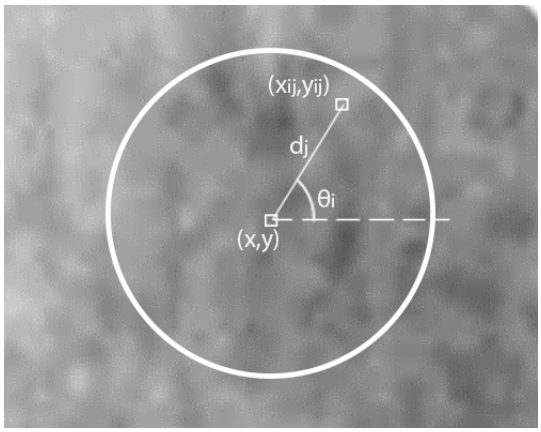

Figure 2.24: Circular region $\mathbf{C}_{xy}$ and sampling intensity difference $I(x, y) - I(x_{ij}, y_{ij})$ at angle θ_i and distance d_j.

To remove errors prone to digitization an inner radius is also set. In most work the best reported inner and outer radii are 4 and 16 pixels respectively [56] [49] [57], so that each distance is $4 \geq d_{ij} \leq 16$ pixels. Each direction with ≤ 4 pixels are removed from the table as there is not enough information to construct a useful Hurst coefficient.

Data: Patch of trabeculae from radiograph
Result: fractal dimensions for each direction.
foreach *pixel* (x, y) *in Image* **do**
  Set region $\mathbf{C}_{x,y}$ at new centre (x, y).;
  foreach *pixel* (x_{ij}, y_{ij}) *in* $\boldsymbol{C}_{xy}$, *such that* $(x_{ij}, y_{ij}) \neq (x, y)$ *and* $4 \geq d_{ij} \leq 16$ **do**
    Calculate angle θ_i and distance d_{ij} to pixel (x_{ij}, y_{ij});
    Take the absolute difference between the pair $diff = I(x, y) - I(x_{ij}, y_{ij})$;
    if $diff > \boldsymbol{R}(\theta_i, d_{ij})$ **then** $\mathbf{R}(\theta_i, d_{ij}) = diff$;
  end
end
Plot each row of maximum differences verses the distances in a log-log plot.;
Fit a line of best fit to the data. Take the slope β;
Calculate FD using $FD = 3 - (2/\beta)$;

Algorithm 1: mHOT algorithm to calculate FDs over an image patch

The mHOT method has been used to measure roughness on surface images [56] and trabecular structures [57]. The method was found to achieve the same if not better results in estimating FD and handling noise and blur in the images [57]. The main issues with this method are: the calculations are based on pixel maximal intensity difference, which makes the method susceptible to noise, and the method has problems generalising over different image magnifications. Two main improvements were made from the mHOT method, these were the Variance Orientation Transform (VOT) [49] and the Augmented Variance Orientation Transform (AVOT) [58].

The VOT method uses the mHOT method, but instead of using a variable number of differences per angle, the method selects the missing values along each angle which have not been filled. For each angle with fewer values than the major axes of the circle ($d_{ij} < 13$). The algorithm searches for pixels to fill the empty distances in the table by sampling a 3×3 region along a line projected down the angle, making sure the pixels were not included in any other direction. The next major change is the maximum absolute intensity difference is changed to be the variance of the absolute intensity differences, to reduce the effect of noisy data. Finally, to handle the different resolutions of the image, the log-log plot points are split into overlapping subsets, shifted by one point. Each of the subsets must be an odd number and have at least 3 pixels (the central distance of each set represents the scale). This returns multiple lines, fitted to each subset per angle, to get Hurst coefficients at multiple scales. The VOT method achieved better results than its predecessors mHOT and other similar algorithms (Fractal Signature HOT and Blanket Rotating Grid algorithms) when analysing trabeculae. The changes make the algorithm less susceptible to magnification and exposure [49]. In the paper by Wolski et al. [59] they found the VOT method returned more information to quantify trabecular differences between OA and non-OA knees than the mHOT method. The results indicated changes in trabecular angle (potentially from the abnormal stresses and loading of the joint), change in anisotropy at different sizes and a change in thickness along the large trabeculae.

The AVOT [58] improved upon the VOT algorithm to handle variable image sizes,

changing the region $\mathbf{C}_{xy}$ radii with the equations $floor(\sqrt{min(\mathbf{R}_w, \mathbf{R}_h)})$ and $floor(r_2/4)$ to handle the outer (r_2) and inner (r_1) radii (respectively) and $\mathbf{R}_w$ and $\mathbf{R}_h$ indicate the width and height of the texture region being analysed. The splitting of log-log points into scales was also changed, with the marginal points (points at the extremes of the inner and outer radii) split into subsets of 3 to handle small radii sizes. This algorithm was developed to analyse trabecular structure of hand radiographs, which are much smaller than knee radiographs, and can be used to standardise the FD measurements between images of varying resolutions for all texture sizes.

Histogram based methods One of the main issues with FD methods is their susceptibility to image artefacts, such as noise, magnification and the x-ray projection angle. To adjust for these artefacts, methods which analyse pixels through sampling regions and intensity gradients [60] [8] have been designed. The paper [60] compares two image processing algorithms to define the fibrous structure of the trabeculae. The first, a Local Binary Pattern (LBP), uses a 3x3 neighbourhood with weights set perpendicular to direction of the fibres to get a clearer distinction of the structure in the image. The second uses horizontal and vertical Laplacian-based matrices. Each method generates a series of texture variables based on the enhanced pixel intensity and contrast values to split OA and non-OA regions. Bone density was estimated using the unprocessed mean and normalised pixel intensity across the region. The paper found the bone density thickens with the increase in KL severity; this finding is limited by the low inter-observer reliability. The structural analysis of the bone (LBP and Lapacian methods) produced the best repeatability scores, and indicated an increase of sclerosis throughout OA. Further to this, the paper also found some structural changes in the trabeculae of the femur, but overlap from the patella can cause difficulties in distinguishing the femur trabeculae from the patella.

The Signature Dissimilarity Measure (SDM) [8] removes the necessity of quantified texture variables and instead calculates the difference in trabecular texture in OA and non-OA images using Earth Mover's Distances (EMD) [61]. This compares normalised signature histograms of roughness and orientation per image and generates a distance value, which can then be split by a classification algorithm. The roughness is calculated

using a Gaussian kernel and gradients of the image intensities. The kernel is run over varying scales δ_n which is fixed at $\delta_n = 1.1^n$ with $n = 1, ..., 25$. Each scale is filtered to highlight the maximum and minimum intensity pixels in the sampled image, so that for each region, there are 25×2 images of gradient maxima and minima per scale. The roughness measure at each pixel $\mathbf{R}(x, y)$ then becomes the difference in maximum intensity differences at the δ_n scales. To remove noisy data from the image, the roughness is normalised with respect to the standard deviation of roughness values across the whole region (I). $\mathbf{R}_{norm} = \frac{\mathbf{R}_{(x,y)}}{StDev(I)}$. These values are then stored in a histogram.

The orientation is a calculation of the directions of highest edge roughness and the difference from the Principal Gradient Direction (PGD). The edge roughness is taken as the pixels along the edges of the maximum intensity regions from the binary image (obtained by adding maxima and minima images) and storing the angle $\theta(x, y)$ between the gradient vector at edge pixel (x, y) and the image horizontal axis. The PGD is then calculated as the angle with the most edge pixels. The histogram is constructed using the angles $\theta(x, y)-$PGD. Both histograms are normalised so all values sum to 1. The paper compares the SDM to an LBP [60] algorithm and a texture analysis method (WND-CHARM) that uses a series of image processing techniques [9]. The results show that the SDM achieves the best accuracy in detecting KL 0 vs. KL 2 and 3, with 85.4%, compared to WND-CHARM: 64.2%, and is comparable to the LBP method when analysing generated fractal images.

Joint Space Width

The measuring of joint space and joint space area has been automated since as early as 1989 [48] using a series of projected graph lines across the radiographs. Many methods focus on edge detection algorithms to delineate the edges of the tibial plateaus and femoral condyles [62] [63] [64] [65]. The most prominent work in this area is the semi-automated software by Duryea et al. [63]. The software requires a technician/operator to crop the knee from the image, with a vertical line dissecting the medial and lateral compartments. The program then finds the edges of the femoral condyles and tibial plateaus by finding the brightest gradient pixels, and applying a threshold and further edge detection to select the pixels forming a continuous edge along the joint space

outline. The measurement of the minimal joint space between the two bones was constrained to avoid the inter-condylar notch, which is not relevant to JSN during OA progression. The results were compared to manually measured radiographs in order to assess the accuracy of the program with differences of between 0.16mm and 0.18mm for normal and Osteoarthritic knees. The automated methods resulted in an increased accuracy of a factor of 2 over manual methods. This method was later expanded in [7] to include a series of joint space widths across both compartments (see Fig 2.25 below). The paper compared the fixed JSW to the progression of OA using KL grades. The mJSW was the best measure for KL ≤ 1, but for KL ≥ 2 they found that distances close to the inter-condylar notch were a better predictor for OA, not the minimal distance. Including extra JSW values doubled the time taken per image to acquire the data.

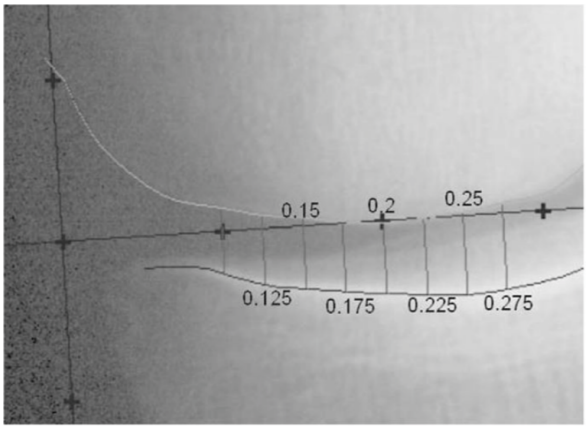

Figure 2.25: Fixed JSW measurements across the Medial compartment [7]

A fully automated algorithm, Knee Osteoarthritis Computer-Aided Diagnosis (KOA-CAD) [6], used edge detection algorithms with Canny [66] and Roberts [67] filtering to outline the tibia and femur in the radiographs. The paper reports inter-observer correlation of the features with OA using Spearman's correlation coefficients - where 1 is a strong positive correlation, -1 is a strong negative correlation, and 0 is no correlation. They analysed the mJSW in both the medial and lateral compartments and compared findings to the gold standard OARSI grades given, finding an inter-observer reliability of 0.54 and 0.53 (medial and lateral respectively), they also evaluated the AUC in detecting OA in a separate study [68] which achieved 0.728 and 0.5435 (medial and lateral). The semi-automated algorithm, Knee Images Digital Analysis (KIDA) [5] uses a similar process, but with two manually placed angled lines across the lowest

point on the femoral condyles and a second line on the base of the tibial plateaus. The algorithm measures the mJSW across the whole joint space and records the width at two reference points along the lines along the femoral condyles and tibial plateau. The algorithm compares the mJSW as well as the smallest compartment reference JSW. The correlations with KL grades (Spearman's correlation coefficient) are: mJSW - 0.57, medial JSW - 0.26 and lateral JSW - 0.15.

Knee Alignment

Knee alignment can be dependent on different features of the knee. Methods shift between basing angle on the bone shafts (diaphysis) [69] [6] or the space between the ends of the bones (epiphysis) [5] (see Figs. 2.26 and 2.27). The KIDA algorithm [5] focuses on the epiphysis angle. The method starts with two lines (altered by the observer) touching the lowest points of the femoral condyles and tibial plateaus. A series of four circles in medial and lateral compartments of the tibia and femur (16 in total) are placed along these lines based on distance from the margins. Using the central two points positioned from each compartment (4 points from the tibia and 4 from the femur) two straight regression lines are projected, one along the femur condyles and one along the tibial plateau. Where these lines meet is the alignment angle. This tibiofemoral angle achieved a fairly low correlation with KL grade (0.3), but good repeatability $3^\circ \pm 2.1$.

The paper [69] uses a semi-automated approach to measure the angle from the diaphysis of the tibia and femur. This positions two lines along the femoral condyles and tibial plateau, similar to the KIDA method [5]. The femur axis is then is calculated by taking the femur line and projecting a perpendicular line halfway between the two condyles towards the top of the image. The tibial axis is taken from two pairs of manually annotated points 1cm and 10 cm beneath the lowest point on the tibial plateau. Two lines are joined between the point pairs and a perpendicular line is drawn through the centre of each line towards the top of the image. The angle between the two axes (at the point they cross) is the angle of alignment of the two bones. The method improved inter-observer reliability from past methods, with correlation coefficients of 0.92 for the semi-automated method and 0.98 for the current method.

The KOACAD algorithm [6] also uses the diaphysis lines for tibiofemoral angle. The

algorithm uses a series of filters and edge detectors to find the medial and lateral outlines of the tibia and femur drawing a central line down the diaphysis of both bones to the inflection points (curves that conjoin the diaphysis to the epiphysis). A straight regression line is drawn along both central lines and the angle the lines create where they cross is taken as the tibiofemoral angle. The paper analyses 1979 images over varying flexion and x-ray angles, finding the repeatability ranges between 0.86-0.94, meaning the method is robust to variation in x-ray method and knee positioning.

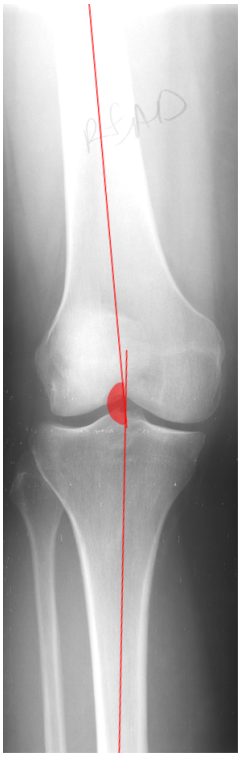

Figure 2.26: Angle between the diaphysis of the tibia and femur (angle and axes indicated in red).

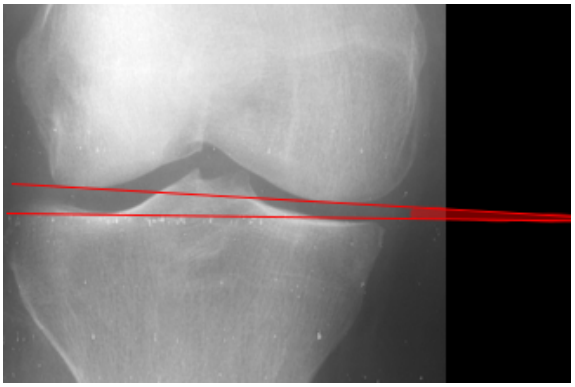

Figure 2.27: Angle between the epiphysis of the tibia and femur (angle and axes indicated in red).

Osteophytes

Osteophytes vary drastically in shape both during OA development and between different participants. The contrast of osteophytes in the radiographs also varies through bone thickness, joint rotation and location in relation to the x-ray projection angle. The methods in [70] and [5] overcome this problem by including operator input to

locate osteophytes along the margins. The KIDA [5] software places the circles (scaled to fit the average osteophyte size) close to the margins, the operators then shift them to surround the osteophytes. A series of points are then manually annotated on the outer edges of the osteophytes and program calculates the area between the manual edge and the central point of the circle. The paper analyses 20 normal and 55 OA knees, and measures reliability (correlation coefficients) of each of the four marginal osteophytes: lateral femur - 0.42, medial femur - 0.45, lateral tibia - 0.53 and medial tibia - 0.57.

The KOACAD algorithm [6] finds the medially prominent edge of the osteophytes extending from the tibial plateau by drawing a line along the medial tibial side using edge detection algorithms. The area is taken between the outline of the osteophyte and the tibial outline. The inter-observer correlation coefficient of osteophyte area is 0.54. This is comparable to the correlation found by the KIDA [5] medial tibia osteophyte area (0.57). The KOACAD was also run on 5950 images in a separate study [68], where the osteophyte area achieved AUCs of 0.691 when detecting OA (KL $\geq$ 2) in women. An improvement for both algorithms could be to include more osteophyte information: include more osteophyte areas in [6], and look at more aspects of shape/height in [5] and [6].

Tibial Spines

The tibial spines are a relatively unexplored area to determine OA development. The experiments in [31] report very weak associations if any at all in 950 images. The only automated method that analyses tibial spines is in the KIDA software. The algorithm places two circles on the image, to be moved by an observer so the bottoms of the circles are touching the top of the lateral and medial tibial spines. The software then measures the distance between the bottom of the circles and the line placed along the tibial plateau. Findings in the experiments were insignificant, with the lateral and tibial spines showing a correlation of -0.15 and 0.14 (respectively) with KL grade. Including more information about the angle and spiking of the tibial spines could create a stronger association, similar to the findings in [71].

Composite Feature Methods

Many algorithms combine features to strengthen the OA detection and inter-observer repeatability. These methods are roughly split into two groups: implicit features, these methods analyse raw image data (i.e. shape or texture) without distinguishing any specific OA characteristics but implicitly including them; and explicit features, which specifically measure OA features and then combine the features/outputs to analyse disease development.

Implicit Features These methods focus on analysing the shape or texture of the whole joint. The work by Shamir et al. [9] looks at a texture-based fully automated analysis of radiographic images. The algorithm first detects the knee through scanning a frame across each image. Once the relevant area of the image is detected, the algorithm extracts features from the data, based on shape and appearance (1470 in total). These are then reduced using Fisher's Discriminant [72], which finds the features that best separate data into the two classes. The classification of new images is done through a distance measure from other images belonging to each class. Each image that returns a strong weighting towards features relevant to Osteoarthritis will be closer to the images in KL grades 3-4. The study tested the method on 140 images, grades ranging from KL 0 to 3, and achieved an overall classification accuracy of 91.5%. The algorithm achieves a high accuracy, but the distance-based classification can be prone to errors when an image lies directly in between two classes (grade 1.5). The algorithm by Anifah et al. [73] uses shape information across the whole image, calculated from the texture features in the image. The algorithm first highlights contrast using Histogram equalisation. Then a series of methods are applied to extract edge orientation information and frequency of edges in the image. A series of textural features are extracted based on contrast, correlation and location of the edge information. A Self Organising Map (SOM) [74] is then used to spatially separate the image data and classify the images into the different KL grades. The paper uses 303 images, and achieved high accuracy (AUC) for the extreme grades (KL 0 and 4), but lower AUC for the middle grades: KL0 - 0.96, KL1 - 0.8, KL2 - 0.08, KL3 - 0.14, KL4 - 0.98. The low AUC for grades 2 and 3 could be due to the similarities between the grades, and the overall texture and shape information does not capture enough information

about specific osteophyte and JSN development in the images.

The experiments in [10] analyse shape information of the tibia and femur outlines using Active Shape Models (ASM) [75]. The model makes use of statistical model that maps the shape of the object and its deformations. The shape is based on a set of manually placed points annotating the key shape features of the objects. The model learns to place the points in new images based on texture information around each annotated point. To learn the shape features, the points are combined and aligned to the same co-ordinate frame. Once completed a Principal Component Analysis (PCA) method is applied to find the highest shape deformations across the training set, these are called shape modes and are ordered from highest to lowest. The paper found that 5 of the top 6 modes correlated with KL progression (107 images), with the highest mode predominantly showing compartmental JSN, modes 5 and 6 varied the shape of the tibial plateau and femoral condyles.

Explicit features These methods contain features which have been explicitly described in the independent feature sections above [6] [55]. The algorithm by Kraus et al. [55] combines FD based on the 2-D box counting method with a manual knee alignment angle and JSN score. The paper reports an improvement in accuracy (AUC) from 0.75 FD assessment to 0.79 when combining all features in their prediction of OA. These features focus on trabeculae structure and alignment angle. Adding extra features for osteophytes and further detail on multiple FDs, like in [49], could improve this accuracy further.

2.4.2 Pain Detection

The prediction and detection of pain related to Osteoarthritic features is still a debated topic, with many contradicting and weakly correlated findings. Automated methods have related quantified features to detecting and predicting pain and disability [32] [11]. A multi-feature assessment in [32] found an association with groups of radiographic feature measurements and the cross-sectional and longitudinal prediction of chronic pain. The model uses Duryea's semi-automated JSW measurements,

coupled with a series of manually assessed features: osteophytes, sclerosis, cysts, attrition, JSN, and Chondrocalcinosis (deposits of calcium pyrophosphate dihydrate which tend to cause damage in the knee during OA progression). The algorithm collected this data from the OsteoArthritis Initiative (OAI) dataset, using only the images with either all the semi-quantitative measurements or all quantitative JSW measurements available. This left 163 images with quantitative JSW measurements in one set, and 123 with semi-quantitative manual measures of OA features in the other. Three pain assessments were run, evaluating performance with AUC. The T0 comparison used case and controls at the time of chronic pain development (AUC 0.695 JSW, 0.62 semi-quantitative). T1y looking at images 1 year before (0.623 JSW, 0.62 semi-quantitative), and T2y which used images from two years prior to incidence of chronic pain (0.62 JSW, 0.61 semi-quantitative). They extracted the features for best detecting the outcome in each test, with JSW being the overall strongest feature in all experiments. The strongest semi-quantitative features found were osteophytes and Chondrocalcinosis. The semi-quantitative features make the data sets difficult to compare adding a lot less information and more susceptibility to error than quantitative measurements. Automating the analysis of the other features would allow comparison of the features on the same data, and allow analysis over a larger range of images.
The KOACAD [6] method applied explicit features (joint space area, mJSW, tibiofemoral angle) to detect painful from non-painful knees at a specific time point (cross-sectional data). The data found only a weak association with low mJSW and a high tibiofemoral angle, and no association with the linear progression of KL severity. This weakly supports the findings in [22] [32]. The KIDA algorithm [11] found osteophytes more correlated than JSW, when applying the various semi-automated features to clinical symptoms of OA (JSW, tibofemoral angle, osteophyte area, bone density and height of the tibial spines). The outcome was the WOMAC score, an assessment of pain and functionality in the respective knee. The paper assessed the time points of baseline (T0), 2 years before incidence (T2y) and 5 years before incidence (T5y). In all tests, osteophyte area showed significant correlation with the presence of pain. This correlates with the findings in [32]. Other significant features were: height of tibial spines (T0), tibeofemoral angle (T2y), and bone density (T5y).

2.4.3 Longitudinal Osteoarthritis Prediction

Early and pre-osteoarthritic signs of the disease are an important area to improve understanding and treatment of OA [76]. Automated methods have been developed that analyse properties of the bones in relation to the prediction of later onset disease development. The most prominent works look at 3-D bone shape [77] and surface area [78]. The following section will be split into MRI and radiographic methods that automatically analyse features to predict later onset disease from early and pre-Osteoarthritic images.

Magnetic Resonance Imaging Methods

A novel method by Bowes et al. [78] uses an Active Appearance Model (AAM) to detect Osteoarthritis through 3-D area change in the bones. The algorithm compares features of baseline non-OA patients to predict the outcome of follow-up visits two years later. The AAM detects the shape of the bones from the MRI images, defining the outline so that the area within can be calculated. The bone surface area was compared to Joint Space Width and cartilage thickness grades. They found that area change of the three bones (patella, tibia and femur) appeared before any other Osteoarthritic signs were visible.

A similar study by Neogi et al. [77], used 3-D bone shapes to predict later onset radiographic Osteoarthritis. The automated method also used an AAM to detect the individual shapes of the tibia, femur and patella. The knee examples were then separated into positive and negative cases, using Linear Discriminant Analysis (LDA), to find a linear threshold that best separates the data. The method then examined the abnormal bone shapes and the likelihood of OA developing in the 12 months following the baseline scans. The radiographic Osteoarthritis was graded at both points using the Kellgren-Lawrence method, with all baseline images included at KL0. They found that the best predictor came from combining the shape vectors of all three of the bones, patients with abnormal shapes in all three bones were three times more likely to develop OA in the 12 months after baseline. This prediction was found to increase in likelihood in follow-up visits longer than 12 months, predicting the future development of Osteoarthritis up to 2-4 years in advance. This is still a developing area of research, but as these changes have been detected in the bones of the joint, it is possible that

radiographs could also be used to predict later onset OA.

Radiographic

The work of Shamir et al. [12], analyses multiple texture features across the joint space (described in Sec 2.4.1) to predict the development of OA up to 20 years after the baseline images. All initial x-rays in the 246 images were classified with KL0. The method achieved fairly good results, predicting images would progress from KL 0 to KL 2 in 20 years with 62% accuracy, and KL 0 to KL 3 with 72% accuracy. The main issue with this paper is the gap between visits, as the participants could have developed OA at any point during the 20 years after the initial x-rays, conducting analysis over multiple time point predictions, like in [32] and [11], or reducing the length between the follow-up visit could add more information about the disease development and improve prediction accuracy.

The work by [79] used features from the KIDA algorithm to measure later onset OA (KL $\geq$ 2) from baseline images with KL $\leq$ 1. The model achieved a good accuracy when predicting OA 5 years from baseline, with an AUC of 0.74 in the combined feature analysis (gender, Body Mass Index (BMI), mJSW and osteophyte area).

Joint Space Narrowing Prediction The prediction of later onset OA has also been expanded to focus on the development of JSN in radiographic images, as loss in joint space is a key factor in the progression of OA [2]. The combined feature analysis in [55] which made use of explicit FSA, JSW and the area of the joint space (JSA) predicted JSN and decrease in JSA using the direction of the trabeculae in the image. This prediction was predominantly based on a shift of trabeculae becoming more vertical, and predicted a 5% change in JSW with an AUC of 0.85 and 5% change in JSA with an AUC of 0.81.

This is supported by the work in [8], which used the SDM method (described in Section 2.4.1) to predict a decrease in joint space from knees with early OA KL $\leq$ 1 (n=135), knees with late OA KL $\geq$ 2, and using all images in both medial and lateral trabeculae regions. The prediction was assessed using AUC and achieved: all images - 0.74 (0.67,0.82) medial and 0.68 (0.62,0.75) lateral, early OA - 0.74 (0.67,0.82) medial and 0.72 (0.64,0.8) lateral, and late OA - 0.76 (0.68,0.84) medial and 0.68 (0.60,0.77)

lateral.

The prediction of loss in joint space is useful for measuring the specific progression of OA, however, the disease features are reported to vary widely between patients [18]. Predicting only JSW could be limiting the algorithm, targeting only people who develop OA through joint space loss. Also, using FSA to predict JSN could just be emphasising the relation between joint reloading (from JSN, alignment change and remodelling that happens concurrently) and a response in trabeculae to mediate the focal pressure, stated in [15] and [16]. The change in JSN seen in radiographs will not be measured with perfect accuracy, so the trabeculae shift could be compensating loading that is already present.

2.4.4 Summary

As can be seen, automated methods have greatly improved Osteoarthritis analysis, allowing detailed analysis of features that were previously ungraded in manual methods (trabeculae, alignment, osteophyte area, joint space area, and multiple JSW measurements), adding flexibility for multiple features to add implicit information, improve OA grading, and speed up analysis over large sets of radiographs. However, manual methods still achieve a better accuracy for OA classification than their automated counterparts. The KOACAD achieves the best independent feature AUCs with: osteophytes 0.645, medial mJSW - 0.728, and lateral mJSW - 0.5435. This could be improved by combining features to detect OA, such as in [55] which combines FD and JSN to achieve an AUC of 0.79. The KIDA algorithm also achieved fairly good results, correlating the inter-observer repeatability to KL grade with: osteophytes - lateral femur 0.42, medial femur 0.45, lateral tibia 0.53 and medial tibia 0.57, alignment 0.3 and mJSW 0.57. These results are comparable to the manual grading inter-observer repeatability (0.36-0.8). A more comprehensive analysis could be done if the features were combined or compared to individual OARSI grades.

Furthermore, the combined explicit feature models analyse various features, but none have looked at analysing explicit and implicit combined models. OA is a disease that affects the entirety of the joint in varying feature development, extending the analysis to include overall shape or texture features and explicit measurement of OA features could expand understanding of OA and create a stronger detector and predictor of

future onset of the disease.

Another issue apparent from the literature is the dependence on operator input, requiring input and interaction for every image. This slows the process down and removes the ability to run the algorithms over large sets of data, without being extremely expensive and time consuming. Expanding these algorithms so that explicit and implicit features are analysed using a fully automated system would allow for further analysis of OA features and allow for additional methods to be applied on top without the need for operator input.

The current automated methods all use two primary functions: 1) image processing - methods to analyse the radiographs and gather the necessary information for further processing, 2) feature analysis - from the gathered data, the main features associated with the disease and non-diseased joints are processed further before being used to train the relevant classification/regression algorithms.

Following these stages, the first problem to approach is how to first segment the relevant information from the image. During OA the shape of the knee changes with the development of osteophytes, widening and flattening of the plateaus (attrition), JSN and alignment and a general change in normal shape properties (bone remodelling). So to accurately distinguish change, automated methods must first accurately find the outline of the bones. This concept is termed Object Segmentation, and is an extensively researched problem within the area of Computer Vision.

2.5 Object Segmentation

Object segmentation is an important field of study in Computer Vision and Machine Learning, and is used to detect and track objects in a given image. It is a process of applying algorithms to discriminate the object from a background, and locate the relevant properties and boundaries. The algorithm itself requires three things to be able to run: a model that learns the most defining features of an object, an optimisation algorithm that will best fit the learned parameters to a new image, and a set of images to form a template that is used to compare to the new data.

The first approaches to segmenting objects were to apply a simple rigid structure of

the object, similar to a fixed template of the object shape, which uses a correlation measurement to detect similarities between the rigid object and the texture in the image. However, applying the model to deformable and organic structure, such as faces, hands, and bones, means that the models must be able to adapt over a diverse set of transformations the object can vary between. Training each of these deformations into a model can be time consuming, and often has unwarranted effects. The learning algorithms applied must be able to determine the true nature of the object through its key characteristics. These characteristics are based upon observed features, defined in the initial stages of the creation of the model. The model then learns the characteristics during training, using a set of example images with different forms and positions (training data).

The different methods of detection have spanned various solutions: from simple thresholds, to matching a pre-defined template of the object to the new images. Previous algorithms that attempted to locate deformable objects [80] [81] looked to making a base template object and allowing a certain degree of freedom on each of the defining features, so making the models 'elastic'. This involved the model points moving over a region around their current position, trying to find the best fit irrespective of the logical transformations of the object. The work of Cootes et al. [75] aimed to remove this 'elasticity' by constraining the deformations to those based on statistical measurements of the object (Statistical Model). A statistical model maps the characteristics, such as shape and appearance, which best describe the object as a whole. The features will span a set of many variations across one object. This model can then be applied to new images by optimising a quality of fit function.

2.5.1 Statistical Model Methods

Statistical models are a method for mapping the observable variation of an object across a given set, and representing the examples as the mean example plus some linear combination of the modes of variation. Statistical Shape Models (SSM) map this variation through the shape of the object. The shape is represented through a series of annotated points and connecting curves. The SSM then takes the shape of each object

in the training set and learns the variation using PCA. This process constructs eigen-vectors and corresponding eigenvalues that describe the variation of the object and amount the shape features vary across the data. The eigen-vectors are then ordered according to eigenvalue from highest to lowest. Each shape can then be represented using the equation:

$$x = \hat{x} + \mathbf{Pb} \tag{2.1}$$

Where x is a vector representing the shape, $\hat{x}$ is the mean values across all features of the object, $\mathbf{P}$ is the eigenvectors that describe different forms of variation and $\mathbf{b}$ is the set of object specific parameter values.
The number of modes (or features) depends on the number of eigenvectors kept. This is done by including a percentage of the variation, i.e. 100% shape variation = all modes, 50% variation = shape modes that make up 50% variation. The amount kept depends on the how noisy the data is and how much variation is needed for the problem. Appearance models work similarly to SSMs, but instead of shape, they use smoothed texture from the data to find the best change in intensity regions and gradients from the training data.

Various statistical models are used in detecting bone objects: ASM [10], Active Appearance Models (AAM) [78], and Random Forest Constrained Local Models (RFCLM) [82]. The work by Lindner et al. [83] compares a series of statistical models in detecting the outlines of proximal femurs from 839 radiographic images. The paper compares an ASM, AAM, and Constrained Local Models (CLM). For the CLM models, the points are initially placed, before the CLM optimises the positions, using three different methods. The first two are based on comparing textural regions using correlation and probabilistic models, the third is a Random Forest (RF) regression voting model (RFCLM [84] model). The methods were compared using a point-to-curve error between the model output points and connecting curves, and the manual (ground-truth) annotations. When testing the algorithms, the RFCLM outperformed all other methods with a mean point-to-curve error of 0.9mm for 99% of the images. The accuracy of the RFCLM algorithm has lead to its use in many similar problems,

such as segmenting spinal vertebrae [85], and knees [82].

2.5.2 Random Forest Constrained Local Model

The RFCLM is split into two parts, the first is a RF regression-voting algorithm to search patches in the image and predict the placement of the points, the second part is the application of a Constrained Local Model to fit the most likely points to the observed shape boundaries from the training data.

Random Forest Regression-Voting

The Random Forest Regression-Voting algorithm [84] works on the principle that a large number of separate independent votes will result in a majority vote on the correct answer. The algorithm applies multiple decision trees to solve a problem, combining the votes of each tree. A decision tree is a predictive model, trained by finding features that best split a given sample of data. The resulting 'leaf' nodes (where no more splits can occur) are then outcomes of the data, decided by the majority of class samples in the node. In relation to the current algorithm, the decision trees form a prediction on the location of the shape model points in an image. The trees form decisions on sections of texture detailing the shading and edges in the image using Haar-like features, similar to the Viola-Jones [86] model. The tree outputs a displacement from the given patch of the image, based on learned displacements from the training set, along with a weight on how likely it thinks the prediction is. The mean and standard deviation of the displacements are stored in the 'leaf' nodes during training. This will give the response (mean displacement) and how accurate (standard deviation the result is during testing. Each of the trees are individually trained using a boot-strapped subset of samples and a constrained number of features. The predicted displacement results in a series of votes and accumulated weights per shape point, called a response image ($\mathbf{R}_i$). The stronger the weight of any one displacement, the higher the likelihood the displacement is accurate. The response image is then passed to an optimiser (CLM), to shift the shape model to the best fitting points given by the Random Forest.

Data: Image patches.

Result: Response images, $\mathbf{R}_i$.

for *each point* i **do**

  Create and zero a response image $\mathbf{R}_i$;

  for *Each tree in Forest* t **do**

    for *Each patch centred on* (p_x, p_y) **do**

      Predict displacement (dx, dy) and weight w for point i.;

      Increment votes in $\mathbf{R}_i$: $\mathbf{R}_i(p_x + dx, p_y + dy) += w$;

    end

  end

end

Algorithm 2: Estimate Response Images

Constrained Local Model

The CLM uses a more complex model for the statistical shape, with an added function to match the overall global transformation of the image to the model. The CLM uses a similar method of acquiring a statistical model as the ASM, with a model based on the learned shape characteristics of the object. The CLM then optimises where the points are shifted using the highest weighted votes in the response images, with an added constraint of the shape model to keep the points within realistic bounds.
The algorithm uses an adapted version of Equation 2.5.1 (see Equation **??**) to assess the points quality of fit.

$$x_i = T(\hat{x}_i + \mathbf{P}_i\mathbf{b}; \mathbf{t}) \tag{2.2}$$

With the cost function:

$$Q(\mathbf{b}, \mathbf{t}) = \sum_i \mathbf{R}_i(x_i) \tag{2.3}$$

Where $T(;t)$ is a function mapping the global similarity transform of all the points in the image, and $Q(\mathbf{b}, \mathbf{t})$ optimises the shape parameters and global transforms using the response images from each point.

This optimisation is applied to all of the shape points, shifting the shape and position variables to optimise the most likely position and the shape constraints of the model. The pseudo code below (see Algorithm 3) shows the basic steps to converge the shape model, where the initial point estimates are acquired from the weighted positions of the response image. The algorithm searches around the feature positions in a slowly decreasing radius (narrowing the search area to speed up convergence), calculating the shape variation and global pose of the object. If the chosen 'most likely' points fail to fit within the object bounds, then the points are shifted to the closest valid location in the radius. Regularisation of the points is then applied to transform the points into the reference frame. If the search radius is not at its pre-set minimal value, then the search is iterated with a reduced radius. Once the search is finished, the resulting point relocations are applied to the model points.

Data: Image template and Estimated points.

Result: Converged shape model points.

while $r \geq r\ min$ **do**

  Search in a radius r for best weighted points.;

  Estimate shape variation value and pose $\mathbf{b}$, $\mathbf{t}$ to fit the selected points.;

  If shape is outside the constraints (in relation to the shape points as a whole), move $\mathbf{b}$ to the nearest valid point in the radius.;

  Regularise the points to the reference frame $x_i \rightarrow T(\overline{x_i} + \mathbf{P}_i\mathbf{b}, \mathbf{t})$;

  Reduce the radius;

end

Map results to the image frame;

Algorithm 3: Fitting the shape model to the response images

The RFCLM is both efficient and robust in finding objects within an image. The algorithm improves on problems in the earlier methods, such as the search getting stuck in local minima (AAMs), containing more information about the object in question (ASMs) and applying efficient optimisation relative to the global object shape. The algorithm has been applied to many detection problems and has been used to detect spines, knees and hips from radiographs, making this an ideal algorithm to apply to knee joint detection problems.

2.6 Summary

Osteoarthritis is a prevalent and disabling disease with variable pathomechanics. There exists no known cure or preventative treatment. There is a need for further research of progression and features of the disease. Current grading methods are mostly based on semi-quantitative measurements, which lead to subjective opinions and generalised severity scores applied to the disease. This is shown in the variable inter-observer reliability scores, ranging between 0.36-0.8 for KL grading. Automated methods have been developed to create quantified measures of OA, provide further insight and classify large numbers of radiographs to help clinical trials and patient assessments (reducing the need for expensive manual grading). Improvements are still required in the current automated methods, namely to remove the need for operator input [69] [65] [7] [5], expand features to focus on a broader range of OA characteristics [6] [87] [65], and automatically analyse osteophytes [6] [5].

Automated methods have demonstrated that detectable features can predict the later onset of the disease and pain. These methods are limited in radiographs, but MRI studies have shown that the features are relevant to the bone shape and area. The work by Shamir et al. [12] uses an implicit feature method that analyses texture features over the whole joint space, however, the outcomes are limited by the data. The images acquired only show baseline and then 20 years later follow-up. Including follow-up outcomes from visits one or two years after baseline could create a better analysis, similar to the pain prediction methods in [32] [11]. Applying an all encompassing explicit and implicit radiograph model over a large number of radiographs could potentially improve on knee Osteoarthritis detection and prediction. One of the first issues to tackle when designing a method to approach this problem, is to accurately extract the relevant data from the radiographs. The work of Lindner et al. [83] has shown the best accuracy was from an RFCLM algorithm, when comparing multiple object segmentation algorithms for segmenting proximal hips from radiographs.

In the remainder of the thesis we document: the fully automated system based on an RFCLM to analyse all aspects of the disease, using both implicit and explicit

OA features; the application of the automated model to a large set of radiographs, detecting current and later onset disease; and the accuracy of the radiographic feature model on other clinical factors, such as the occurrence of current and later onset pain.

The methods in the remainder of the thesis use techniques based those found in the literature, the project hopes to combine optimal features for all radiographic features associated with OA. Something which has not been done by methods found in the literature. For comparison, the table below illustrates the various methods used in the current automated methods (see Table 2.3).

Table 2.2: Comparison of Current Automated OA Methods

			Features Evaluated						
Method	Dataset (outcome)	Numbers	JS	OSA	Tr	Im	FTA	TS	Other
mHot [88]	N/A (OA/non-OA)	N/A			x				
VOT [49]	N/A (OA/non-OA)	N/A			x				
AVOT [58]	OAI hand (OA/non-OA)	40 (k)			x				
SDM [8]	N/A (OA/non-OA)	41 (k)			x				
xJSW [7]	N/A (KL 0-4)	217 (k)	x						
KOACAD [68]	ROAD (OA/non-OA)	3040 (p)	x	x			x		
KIDA [5]	N/A (KL 0-4)	75 (k)	x	x			x	x	
CLAHE [73]	N/A (KL 0-4)	308 (k)				x			
WND-CHARM [9]	N/A (KL 0-3)	350 (k)				x			
ASM [10]	N/A (KL 0-4)	107 (p)				x			
Trabecular morphomerty [55]	POP (KL 1-3)	138 (p)			x		x		bmc.

Table 2.3: JS = Joint Space, OSA = Osteophyte area, Tr = Trabeculae, Im = Implicit OA features, TS = Tibial spines, FTA = femorotibial angle, Other = Extra radiographic or clinical features included, bmc. = bone mineral content. Numbers = number of participants (p) or knee images (k) used in the studies. POP = Prediction of Osteoarthritis study. N/A = no reference to the values found in the paper. CHECK = Cohort Hip and Cohort Knee. ROAD = Research on Osteoarthritis Against Disability.

Chapter 3

Data and Methods

This chapter describes the data used throughout the project and the various methods to extract features from knee radiographs. The algorithms have been chosen following the key analyses from the literature (see Section 2.4). The chapter is split into five parts: a description of the data used throughout the project; shape localisation, to find the knee in the image; shape methods, which extract features of shape from the image; texture methods, which extract features based on areas of pixel intensities; and classification, the algorithms used to separate the data based on the extracted features and evaluate the results of the experiments.

3.1 Data

To analyse Osteoarthritis (OA) features and evaluate the accuracy of our methods, we used a series of OA and non-OA images with associated grades from the OsteoArthritis Initiative (OAI) [89] dataset. The OAI is a multi-centre, prospective observational study of Osteoarthritis, run across four different sites over the US. The different sites have a single standardised approach to acquiring radiographic images, however, these are likely to differ because of the facilities, operators and patients used for each scan. This presents a set of data with varying contrast, rotational and positional error which may be present. The OAI was chosen because of the high number of participants, and the centrally assessed manual OA grades that are available in the data. The study recruited 4796 people at baseline, ages ranged between 45-79 and 58% of the participants were female. The study is still on-going and has taken x-rays of the participants

at baseline, to 108 months follow-up. Centrally assessed images are available for the visits at 12, 24, 36 and 48 months. The centrally assessed radiographs were graded by experienced radiologists blinded to the grades attributed by the sites that collected the images. An experienced adjudicator clarified any disagreements between the two gradings. A subsample of 150 participants was used to test inter-observer of the manual grading, on the baseline images the blinded radiologists scored a weighted kappa of 0.70, and a 95% Confidence Interval (CI) of 0.65-0.76. All the x-rays were acquired at fixed-flexion and 10 degree beam angle. For this study we have taken images from baseline to 48 months follow-up.

3.2 Shape Localisation

A shape localisation algorithm is used to fit a series of shape points to the knee in the radiographs and localise the feature extraction methods. In this project the methods are based on an RFCLM [84] algorithm, chosen because of the high accuracy in analysing similar 2D bone shapes [85] [82] [90]. An object detection algorithm is run first to find the approximate global parameters of the knee. The object detection, or global model, uses a RF regression voting technique (see Section 2.5.2) to find two points central to the knee (Figure 4.6). The position, scale and orientation of the knee can be approximated using these points. These approximated parameters are then used by the Random Forest Cinstrained Local Model (RFCLM) to find the outline of the bones. The RFCLM algorithm uses manually annotated landmarks to train the displacements from patches of texture around the points, and the variation of the point locations to train the shape model.

A sample of images were analysed to understand the variation in shape and orientation of the knees. Using this information a set of points were placed to describe the key shape characteristics of the tibia and femur i.e. the corners of the two bones, the points of the tibial spines, and the edges of the femoral condyles and tibial plateaus.

3.2.1 Annotating images

Annotated shape points represent the relevant characteristics of the object to aid both object detection and feature analysis. Ideally the shape should be defined through a concise set of points, containing only necessary features to reduce wasting memory and processing time. Each point increases the number of regression-voting trees required by the algorithm. To aid in placing the points, the annotations were split into two iterations: guide points and detail points. The guide points are used to describe the main shape characteristics on the bone, such as the corners and minimal points on curves. These points are placed first and then a set of equally spaced detail points are placed along curves between the guide points. The detail points fill out extra shape needed on the segments of the object with few distinguishing characteristics. In this project, the points were placed around the outer edge of the tibia and femur, connected with straight-line curves to outline each bone (see image 3.1). The annotations were revised several times to optimise the locations and number of points in the set. The final set took approximately 6 minutes to annotate on each image, and contains 37 points for both the tibia and femur models (74 points in total).

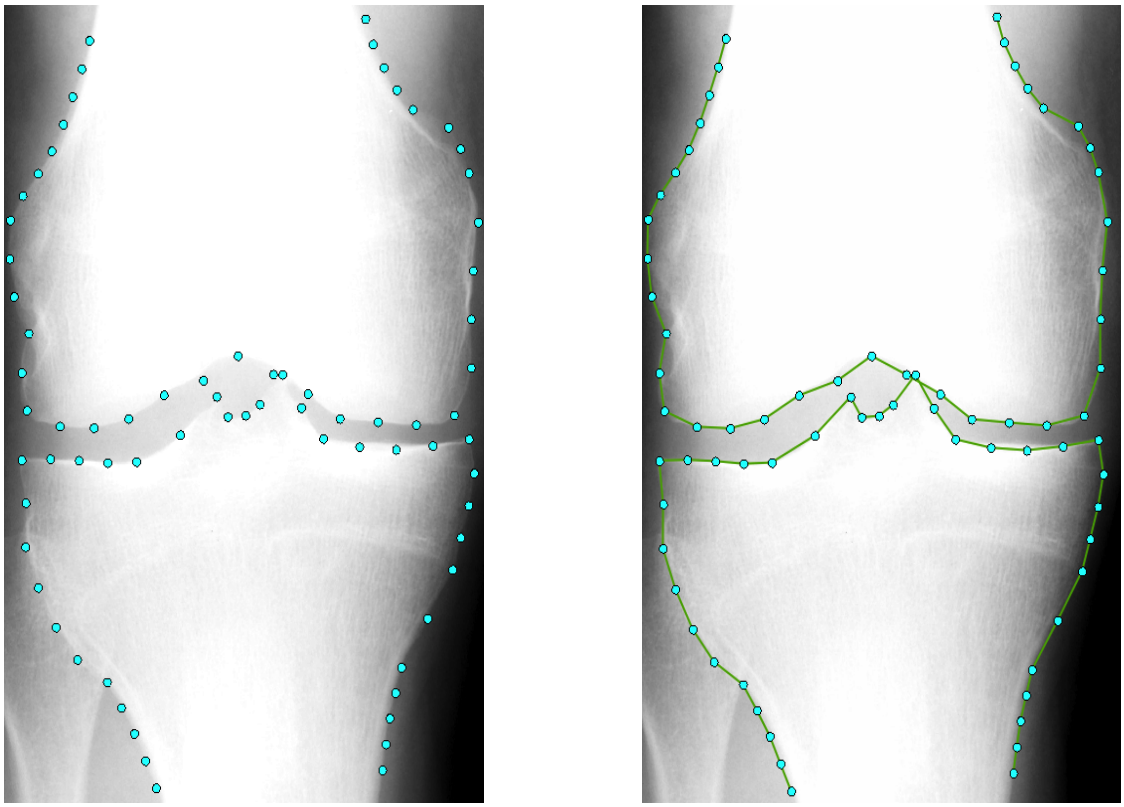

Figure 3.1: Full knee points (left) and points with curves (right).

3.2.2 Random Forest Constrained Local Model Algorithm

The manually annotated points are used to train and test a Random Forest with a Constrained Local Model (explained in Section 2.5.2). The manual annotations enable the algorithm to learn the shape and displacements of the points, and can be used as a gold standard when evaluating the RFCLM output. The application of learning methods to find the optimal shape point locations before fitting shape models has been done in the past, with Multivalue Neurons [91], a neural network non-linear learning method; and k-Nearest Neighbours [92]. The RFCLM method has been chosen for this project because of the previously reported high segmentation accuracy on similar radiographic image problems; the points fitted to the image are easier visualise changes and to apply further texture region segmentation; and the availability of the software within the faculty.

3.2.3 Training and Testing the Algorithm

The RFCLM construction follows a similar process to that in Lindner et al. [93], with a single global model to find a small subset of points central to the joint and a series of local models to find the final 74 points. The global model fits two points to either side of the knee and by fitting a mean example of the points gives an estimate orientation, scale and location of the knee. The local models are a series of RFCLM models built to iterate through increasing resolutions of the image, fitting the points to the best location at each stage. Each resolution iterates the RF search for the displacement of each of the 74 points, a CLM then fits each point to the optimal displacement that complies with the global pose and shape parameters $\mathbf{t}$ and $\mathbf{b}$ of the points (see Equation 3.2.3 below).

The local model is trained on a set of manually annotated images transformed to the same reference frame (image resolution). A Statistical Shape Model (SSM) is constructed from the aligned shapes of the training set. The points are shifted towards a selected mean image, this is iterated with a new image selected as the mean until there is a minimal shift in the images since the last iteration. The model is then scaled to fit the reference frame (set prior to training). The object segmentation algorithm is

split into two stages: firstly, the Random Forest performs the initial point displacement predictions (see Algorithm 2 in Section 2.5.2); and the second step is the shape fitting, which fits the points to highest weighted response positions that agree with the shape model constraints (see Algorithm 3 in Section 2.5.2).

The Random Forest (RF) is trained on patches of texture around each model point. The pose and orientation is estimated by minimising the difference between the SSM points and the manual annotated points, with respect to the global pose of the object ($\mathbf{t}$). During testing the pose is estimated by the global model. We find the pose parameters, $\mathbf{t}$, to minimise:

$$|T(\hat{x}; \mathbf{t}) - x|^2 \tag{3.1}$$

Where $\hat{x}$ is the mean SSM point locations, x is the current image points, and $\mathbf{t}$ is the global pose of the object. $T(:, \mathbf{t})$ applies a scale, rotation and translation encoded in parameters, $\mathbf{t}$. $\mathbf{t}$ is selected to best fit the mean points to the pose of the new points.

To improve the accuracy and efficiency of the local models we use a coarse to fine approach. By changing the number of pixels in the reference frame we can modify the level of detail captured by the model. Using a reference frame with fewer pixels in the early stages enables a rough estimate of the shape to be found quickly. Later models then use more pixels in the reference frame, giving more detail and thus more accurate results.

The model trains a RF for each of the 74 points to be able to locate the point in a new image. This is done by first sampling each image into the reference frame. For each point we sample regions of image at displaced positions around each model point, recording the image patch and the displacement. The displacement is set to be within a predetermined range from the true point position to keep the texture patches localised to the shape points. Once all the patches and displacements from the training set have been gathered, we then train the RF on the gradient information (in this case Haar-like features) contained in the patches. The trees in the Random Forest are given a bootstrapped sample of these features to be trained on. The trees

then split the samples by a threshold t_f that compactly splits the data according to the 'best' feature in the sample they have been given. This is done by finding t_f to minimise the following equation:

$$G_T(\mathbf{t}) = G(\{\mathbf{d}_i : f_i < t_f\}) + G(\{\mathbf{d}_i : f_i \geq t_f\}) \tag{3.2}$$

Where f_i is one feature from sample i and $G(..)$ is a function evaluating the set of vectors. The function finds the threshold that best splits the sample features by the resulting variance of the displacement values $\mathbf{d}_i$ that result from the split. A forest is trained to predict the displacement of one shape point from the given subset of texture features (see Fig. 3.2). These features are fixed and are later used to predict the position and the likelihood of the displacement (weight) of the point in new images.

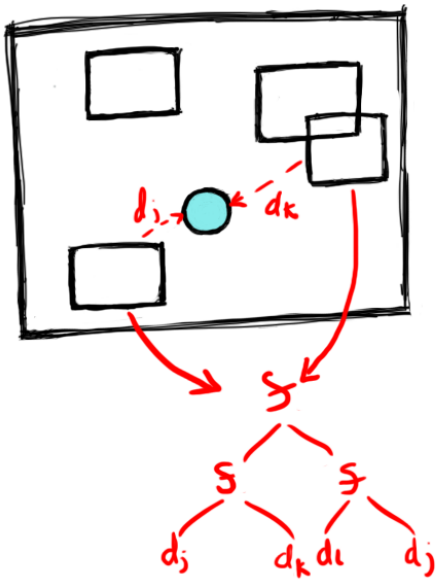

Figure 3.2: Illustration of a single tree in the Random Forest learning the features (f) to split the texture data to predict the correct displacements (d).

Once a Forest has been trained for each point in the shape model, the features and thresholds in each tree are fixed. When given a new image, similar displacements around the estimated mean shape points are taken. The trees apply the trained thresholds to the Haar-like features and form a predicted displacement of where the shape point should lie in the image (see Fig. 3.3).

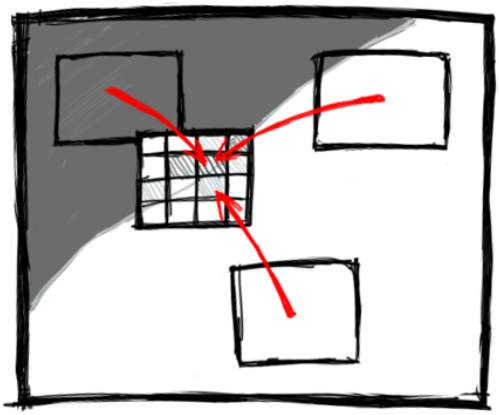

Figure 3.3: Illustration of a Random Forest predicting displacements (red arrows) from patches of image. Displacements and weights are stored in the Response Image.

The predictions form the weighted votes of the Response Image for that point. The Constrained Local Model (CLM) then finds the shape and pose parameters which maximise the total number of votes at each model point. The optimal point positions per reference frame stage are found through the CLMs slowly decreasing radius around each of the 74 points. This will find the best position that fits with the global pose and shape of the rest of the model, shift all points to the optimal, then decrease the radius around the shifted points and search again. This is iterated until a minimal shift or minimal radius size is reached.

3.3 Shape Analysis

This section covers all the methods to extract shape information from the output points found by the RFCLM. Each shape method utilises a SSM to encode the shape using shape parameters $\mathbf{b} = \mathbf{P} \times T(x - \hat{x})$.

3.3.1 SSM

Statistical Shape Models [94] have been used in previous automated methods [10] [77]. They represent the shape of the relevant object as a linear sum of vectors (modes) representing the main ways in which the shapes vary. The modes are found using the

Equation 2.5.2 in Section 2.5.1. These vectors can then be used to train a classifier to separate the image data.

3.3.2 Contour Extraction using Dynamic Programming

In the literature, one of the main algorithms used in automated OA analysis is edge detection. This is used to delineate edges of key OA features such as joint space [7] [62] [64], osteophytes [6], and the bone diaphysis [6] [69]. Previous algorithms scan edges for gradient change [62] [64] and draw the contours along the brightest edges. The RFCLM is dependent on fitting the points to a shape model and so often misses unnseen or unusual variations in the object, such as bone remodelling and boney spurs (osteophytes). A per point edge detection was chosen to detect these unseen shapes following methods used in the literature. The algorithm consists of two parts, first is the extraction of the gradient change across the relevant area of bone, the second is fitting an optimal curve to the gradient variation so that it fits the outline of the feature. Using the points output from the RFCLM, we select the points relevant to the area we want detailed edge information. As the RFCLM is trained to find a sparse set of points to keep processing time minimal, these points must be expanded. This is done using Bézier curve interpolation [95] to expand the points by estimating the new positions based on the properties of the known data points. We interpolate N points between sections of shape model points output from the RFCLM stage.

Gradient Sampling

Once the interpolated points have been found, the region is sampled to find the intensity variation. Each of the interpolated points have normal vectors projected perpendicular to the line between the current point and the next point in the set. Along this normal the algorithm samples the average gradient at a set number of (M) equally spaced distances. The profiles are defined as the lines $n_j = x_i + d_i.u_i$, where $j = 0, .., N$ $j = 0, .., M$, x_i is a point on the main bone contour, u_i is the unit normal to the curve at that point and d_i is the distance along the profile. This is illustrated in the Figure 3.4 below.

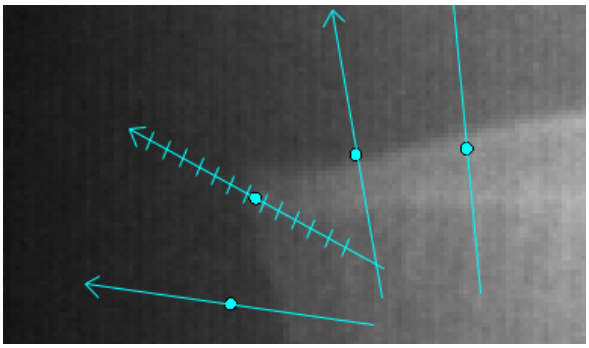

Figure 3.4: Illustrating the projected profiles from the interpolated points (blue circles), and the gradient profiles at set distances along the point profiles.

The gradient between pairs of intensities along the normal project lines are taken and stored in a (NxM) matrix G and passed to the optimisation algorithm.

Dynamic Programming

To find the best continuous contour we find the points at distance d_i along each profile from the bone model contour, which minimise the following cost function:

$$\mathbf{Q} = \sum_{i=1}^{n} -|g_i(d_i)| + \alpha \sum_{i=1}^{n-1} (d_i - d_{i+1})^2 \tag{3.3}$$

where $g_i(d_i)$ is the intensity gradient at distance d_i along the profile i and the second term encourages a smooth shape. The α variable is a weight to control the strength of the second term, a large α means a straighter line is fitted. This equation can be solved using Dynamic Programming (DP). DP is an optimisation algorithm that can fit the interpolated shape points to the outline of the bone (see Fig. 3.5) based on the optimal values of $\mathbf{Q}$. The resulting points can then be encoded using shape models to train and test classification algorithms.

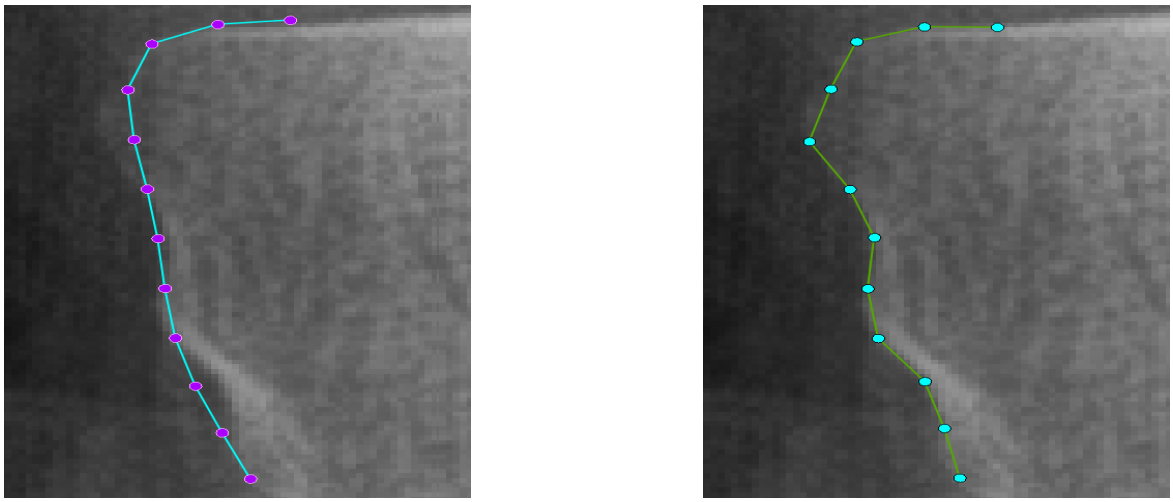

Figure 3.5: Interpolated purple (left) and DP optimised green (right) points.

3.4 Texture Analysis

These methods evaluate the variation of pixel intensities from patches of image. To analyse the texture, a patch must be extracted from a specific area of the image. Problems can arise when sampling texture where the object is not consistent in location, orientation or scale. To avoid this problem we sample texture in a fixed region on each knee, using the RFCLM output points. The positions of two points are sufficient to define the position, scale and orientation of a local reference frame, in which a rectangular region can then be constructed (see Figure 3.6). This means that any region selected will be rotated, scaled and displaced relative to the global properties of each knee.

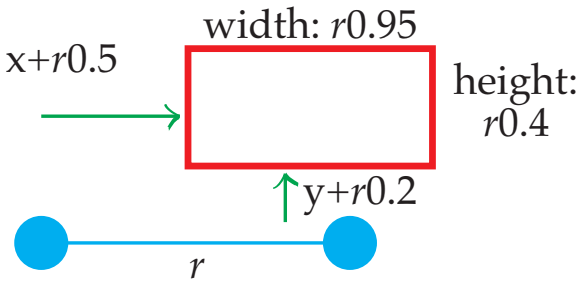

Figure 3.6: Projecting the red region by the vector r between the two blue points.

3.4.1 Fractal Signature Methods

Texture analysis algorithms used in OA studies tend to focus on the Fractal Signature (FS), mapping the roughness, spacing and orientation of the trabeculae as a Fractal Dimension (FD) [49] [57]. The best reported algorithms for analysing OA are the Variance Orientation Transform (VOT) [49] and Augmented Variance Orientated Transform (AVOT) [58] methods.

Variance Orientation Transform

The VOT method analyses the FD of an image by the variation in pixel intensities across a fixed number of distances per angle. The method, described in Section 2.4.1, uses this roughness and orientation to measure trabeculae changes on patches of bone texture. The method extends the modified Hurst Orientated Transform (mHOT) algorithm (see the Algorithm 1) by adding more complexity to the scales and variation used. The differences can be seen in Algorithm 4.

Data: Texture from ROI

Result: fractal dimensions for each scale per direction.

foreach *pixel* (x, y) *in Image* **do**

  Set region $\mathbf{C}_{x,y}$ at new centre (x, y).;

  Calculate mHot pixel differences along each angle.;

  Remove all directions with ≤ 4 pixel distances.;

  foreach *Angle* θ_i *with distances* $\leq d_{0j}$ **do**

    foreach *Missing distances* d_j *in* θ_i **do**

      Search in sampling region at distance d_j;

      if *Pixel* (x_{ij}, y_{ij}) *not used in other angles* **then**

      $\mathbf{R}(\theta_i, d_j) += I(x, y) - I(x_{ij}, y_{ij})$;

    end

  end

end

Calculate variation of intensity differences for each distance per angle;

foreach *Angle* θ_i **do**

  Plot each distance vs. pixel intensity variation in a log-log plot;

  Split distances into N subsets (scales) ;

  foreach *Scale* s *per angle* θ_i **do**

    fit a line using logistic regression;

    H_{i_s} = gradient of the line $\beta/2$;

    FD= $3 - H_{i_s}$;

  end

end

Algorithm 4: VOT algorithm to calculate FDs over an image patch

The Figure 3.7 below shows the log-log plot of a single angle θ_i plotted with the data split into scales. The Hurst orientations can then be plotted using a polar/rose plot to show the orientation of roughness in the image (see Fig. 3.8).

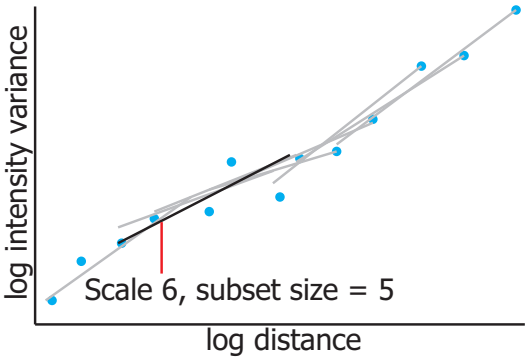

Figure 3.7: Plotted log-log distance vs. variance, with highlighted subset.

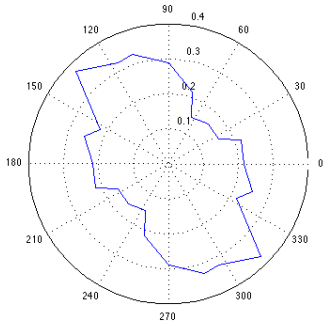

Figure 3.8: Rose plot of the Hurst coefficients across all angles at scale 4.

A limitation of this algorithm is that it only looks specific angles, only expanding θ_i with ≥ 4 distances. An improvement we explored was expanding these directions to cover a wedge of the circular region $\mathbf{C}_{xy}$ so to include all pixels within the sampling region.

Wedge Variance Orientation Transform

This algorithm uses the same process as VOT, but instead calculating the intensities along rays from the centre, the region is split into θ sized 'wedges' (see Figure 3.9). The angle between the pixel and the region centre is rounded up to the nearest θ angle. An increasing number of pixels are used to compare to the central pixel as the distance from the inner radii increases (see Figure 3.10). The increased values are taken into account when calculating the variation of intensity differences. The log-log plots are then split into the respective scales and plotted the same as the VOT algorithm [49].

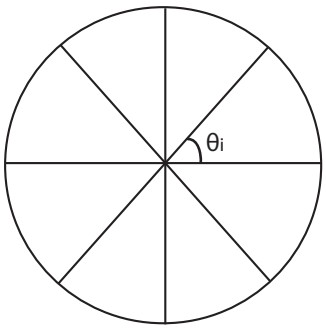

Figure 3.9: Circular region split into θ wedges

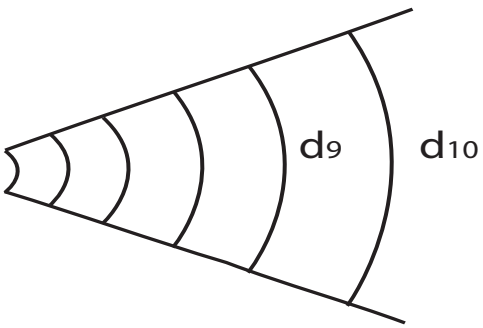

Figure 3.10: Pixels rounded to the nearest distances d in one wedge.

Augmented Variance Orientation Transform

The AVOT algorithm, explained in Section 2.4.1, uses a similar process as the VOT method. The AVOT differs through adapting the algorithm to handle variable texture sizes, this was done to calculate FS in hand radiographs [58], which could not be analysed using VOT method due to the fixed radii and log-log scales. The method changes the sizes of the circular region radii depending on the size of the texture region FS is being calculated from. The inner (r_1) and outer (r_2) radii are calculated using the equations $r_2 = floor(\sqrt{min(\mathbf{R}_w, \mathbf{R}_h)})$ and $r_1 = floor(r_2/4)$. Where $\mathbf{R}_w$ and $\mathbf{R}_h$ indicate the region width and height. Coinciding with this change, the scales of the log-log points (see Figure 3.7) are changed to handle smaller number of pixel distances. The VOT method uses fixed size subsamples of 5 log-log points. The AVOT method splits the three marginal points into subsamples, before splitting the remaining points (see Fig. 3.11). This allows smaller regions to be plotted and adds extra detail in the number of scale gradients per angle.

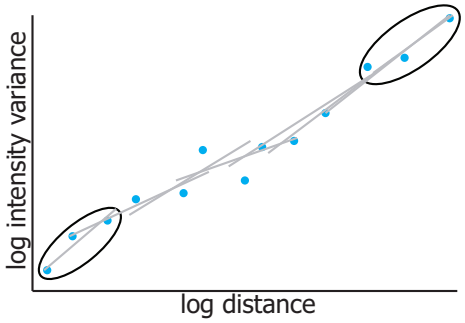

Figure 3.11: The marginal three points on either side (circled) as the start and end scales, remaining subsets contain 5 points.

3.4.2 Pixel Ratios

Other texture methods utilise sampling regions and intensity gradients to evaluate OA [60] [8]. Following these methods, we developed a new Raw Pixel Ratios (RPR) method. The RPR uses a square pixel sampling region but instead of applying filters and contrast enhancements, we extract raw pixel intensities. Samples of fixed pixel size are taken randomly from the texture patch. The method uses N samples per image i using each sample as a separate example to train the classifiers. The training set is made of N samples per all training images $x_{i,j}$ where $j = 1..N$. The RF classifier trains

on ratios between randomly selected intensity pairs in each sample. When testing the classifiers, a single image i is taken and N samples of image are collected x_j, the result of the classification is taken as the mean over each sample extracted from the image: $\frac{1}{N}\sum f(x_j)$. The number of samples across the region is flexible, but large numbers of samples will slow processing time.

3.4.3 Signature Dissimilarity Measure

This algorithm, explained in Section 2.4.1, uses a Gaussian sampling region across a series of scales to acquire pixel gradient maxima and minima values. The paper [8] reports the best number of scales is $N = 25$ with $n = 1..N$ and each scale shifting the region $\delta_n = 1.1^n$. The sampled gradients are then taken to extract features of roughness and orientation of roughness from the entire image region. Intensity gradients are sampled in smoothed regions of the image by taking the first and second order derivatives of the data:

$$\mathbf{D}_1(x, y, \delta_n) = \sqrt{L^2_{x,norm}(x, y, \delta_n) + L^2_{y,norm}(x, y, \delta_n)} \tag{3.4}$$

$$\mathbf{D}_2(x, y, \delta_n) = L_{xx,norm}(x, y, \delta_n) + L_{yy,norm}(x, y, \delta_n) \tag{3.5}$$

Where δ_n is the current scale of the Gaussian sampling region, $L_{x,norm}$ and $L_{y,norm}$ are the first order derivatives, and $L_{xx,norm}$ and $L_{yy,norm}$ are the second order derivatives of the intensity gradients at pixel (x, y). The N smoothed images are stored at the varying scales. The roughness is then the variation of maximum gradient values (maximum value across all scales of each pixel) of $\mathbf{D}_2$. The maximum and minimum of the $\mathbf{D}_2$ values across all scales are taken, and the summed images with a threshold on 0 forms a binary image of the regions of the image - with white being the gradient peaks and the black regions the gradient troughs. The outlines of the peaks are used to indicate edge gradients to calculate the roughness orientation and the Principal Gradient Direction (PGD). The roughness and orientation roughness are stored in two separate histograms, which are used as features for subsequent classifiers.

3.4.4 Haar-feature Analysis

A widely used feature for analysing patches of pixel intensities is Haar-like features [86]. The Haar-like features are formed from weighted sums of the pixels within two or

more rectangular regions (see Fig. 3.12 below). The regions vary in size, orientation and pattern. The pixel intensity values underneath the different colour sections are summed and the difference is taken between the two. This value is then the feature of that region under that specific Haar-like feature, and is used to categorise subsections of the image. A classifier can then learn to separate the data using such features. These differences will correspond edges or shadows in the image that vary between the classes, for example, shape and texture change over OA.

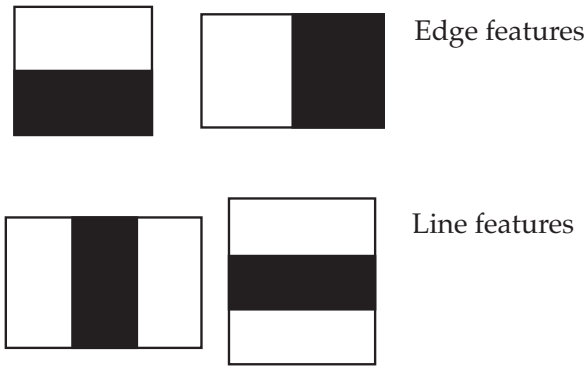

Figure 3.12: Examples of Haar-like features [96].

3.4.5 Texture with Implicit Shape

The method developed by Shamir et al. [9] (publically available from [97]) analyses implicit features to study the progression of OA. The method, explained in Section 2.4.1 uses a series of texture analysis methods, across varying image scales, to extract a large number of features across the whole joint region. The region is placed by comparing a set of 20 selected images of the central joint space (see Fig. 3.13) to the radiograph, the centred region is fit by optimising the equation:

$$d_{i,w} = \sqrt{\sum_{y=1}^{15}\sum_{x=1}^{15}(I_{x,y} - W_{x,y})^2} \tag{3.6}$$

Where $d_{i,w}$ is the distance of the selected region w from the patch i of the image being searched, $W_{x,y}$ is the intensity value in the selected region at pixel x, y, $I_{x,y}$ is the intensity of the pixel x, y in the new image. The 15×15 search region is then shifted and compared to each of the 20 representative joint space images. The best d is chosen as the correct location of the joint space in the new image. The texture extraction algorithms applied to the whole joint are: Zernike features [98], which describe image

properties through a series of orthogonal polynomials; multiscale histograms, which represent intensity information in a series of discrete bins; variation of intensity values taken across four orientations (0°, 45°, 90°, 135°); Tamura texture features [99], these calculate the coarseness, directionality and contrast in the image; Haralick features [100], which record the matching intensity values that occur within a set distance of each other; and Chebyshev statistics [101], these approximate the image intensity regions across varying distances and angles. The 1470 extracted features (210 features across seven image scales) are then put into a Fisher's Discriminant Analysis (FDA) method to find the features which best split the data into the given classes.

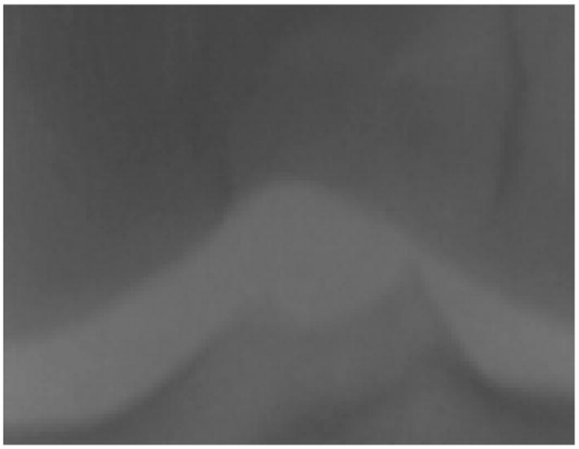

Figure 3.13: An example of one joint space to find the knee in the new image [9].

3.5 Classification

3.5.1 Random Forest Classifiers

A classification algorithm is used to separate the data depending on the features relevant to each of the classes. In this project we use a Random Forest classifier, similar to the Regression-Voting algorithm (explained in Section 3.2.3), but instead of predicting displacements, the algorithm will separate the data into discrete classes. Random Forests can be more cumbersome than simpler methods, but produce accurate results and can be adapted to classify data into many classes. The forests contain multiple trees which classify the data on subsets of the features, finding the best split at each branch by calculating the Mutual Information (MI) or mutual gain. MI uses entropy to calculate the gain of selecting each feature. During training the features f_i are

selected by calculating whether more information is gained by the data classes in the new nodes for each outcome of the split, these features are then fixed for the introduction of new data (test samples). The equation below illustrates a simplified measure of entropy on the data $\mathbf{X}$ when choosing a binary feature f_i:

$$H(\mathbf{X}|f_i = 1) = -\sum_{x \in \mathbf{X}} p(x) log p(x) \tag{3.7}$$

Where $p(x)$ is the probability of data in each class x after splitting given $f_i = 1$. This is taken for both sides of the binary split and a weighted average taken of them both:

$$H(\mathbf{X}|f_i) = \frac{n_1}{N} \times H(\mathbf{X}|f_i = 1) + \frac{n_0}{N} \times H(\mathbf{X}|f_i = 0) \tag{3.8}$$

Where n_{val} is the number of samples in that node, and N is the total samples across both nodes. The entropy is also taken for the data before the split, $H(\mathbf{X})$ and the MI for choosing f_i to split the data becomes $MI(\mathbf{X}; f_i) = H(\mathbf{X}) - H(\mathbf{X}|f_i)$. This value is calculated for each feature, and the feature with the maximum information gain is selected. Once the branching stops, when the nodes contain data that cannot be split any further, the leaf nodes are then evaluated. The probability of each class arriving at each leaf is estimated from the training samples.

In the case of the features extracted from the shape and texture data, the values are typically float numbers, so instead of a simple on-off for the entropy, a threshold is used to split it into the two nodes. This threshold is shifted per feature to find the best split and is compared to the best threshold on the other features that could be chosen at the branch.

For the multi-class problems, i.e. splitting the data by KL or OARSI grade, the calculation is the same, but instead of a probability of two classes, each node contains a histogram of the numbers of each data sample per class (see Figure 3.14 below).

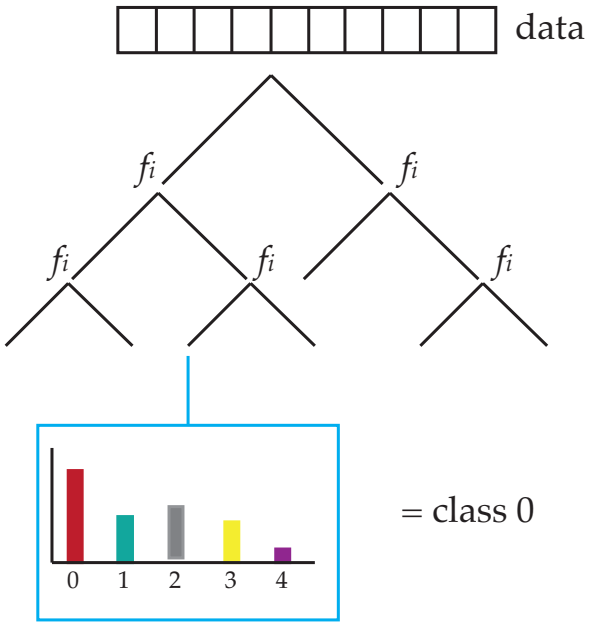

Figure 3.14: Illustration of a tree from the RF splitting the data and classifying the leaf node depending on the histogram distribution.

3.5.2 Cross Validation

Cross validation was used to measure how well the RFs classify unseen data. The method selects a subsample of test data and trains the classifiers on the remaining examples. This process is iterated, choosing different test examples each time, until all samples have been selected for testing. The examples are chosen by splitting the data into k folds, so that one in every fold is selected for the test set. The project used 5-fold cross validation and repeats the method with permutated data to give a mean accuracy, Area Under the ROC Curve (AUC) and standard deviation across the two iterations. The Receiver Operating Characteristic (ROC) curve plots the performance of the binary classification (between positive and negative classes) by plotting the true positive rate (number of correctly classified true images) against the false positive rate (number of misclassified true images).

3.5.3 Statistical Analysis

To analyse the results from the RF and cross validation methods we used a series of statistical analyses to adjust for bias, calculate confidence intervals on the results, and compare the agreement to the gold-standard assessments.

Confidence Intervals Taking the output from the cross validation method, logistic regression was run using the class outcome as a dependent variable. This measures the relationship between the estimated output and the dependent variable and generates a 95% Confidence Interval (CI) by assuming a normal error distribution. The CI represents the range within which 95% of samples would lie.

Contralateral Knees The project used images from the OAI data. To include more data both knees per participant were selected. This can cause biased estimates, as the contralateral knee of a single participant is more susceptible to OA development and associated symptoms [102]. Using data from both knees violates the assumption that all observations are independent, and therefore this needs to be accounted for in the analysis. Using random-effects panel logistic regression on the cross validation outputs and associated classes we adjusted for this bias by assuming all pairs of knees are correlated. The analysis first generates a value of this correlation between all pairs of right and left knees per participant, and then adjusts the final AUC output to remove potential bias.

Comparison With Gold Standard To assess how well the results compare against the gold-standard (manually assigned grades) in the project we used weighted kappa [103] to compare the automated outputs against the manually assessed grades. Weighted kappa takes into consideration the disagreements between both assessments, and weights the disagreement linearly depending on distance from the true answer. The kw is a value between 0 and 1, where 0 means the agreement is equivalent to chance, and 1.0 is a perfect agreement.

3.6 Summary

This section covers a range of different methods to analyse specific features of shape and texture. All methods are based on the object detection RFCLM algorithm, and each method has flexibility to be tailored to specific bone features. The next step compares these methods in evaluating the different features of OA to find the best individual and then combination of features to classify OA.

The next three chapters compare the various methods, and use the best combination to analyse current and future onset of OA and related outcomes of the disease.

Chapter 4

Comparison of Methods

This chapter summarises experiments comparing the methods in Chapter 3 and evaluates the features using images from the Osteoarthritis Initiative (OAI) dataset [89]. The main comparisons will focus on: trabecular structure features, comparing shape and texture to analyse osteophytes, and finding the best combination of all extracted features to detect Osteoarthritis (OA). The Figure 4.1 shows the layout of the system with a radiographic image first being segmented and then analysed for each of the radiographic features using the shape and texture methods. The optimal features are then compared to manual Kellgren Lawrence (KL) grading and current state-of-the-art automated methods.

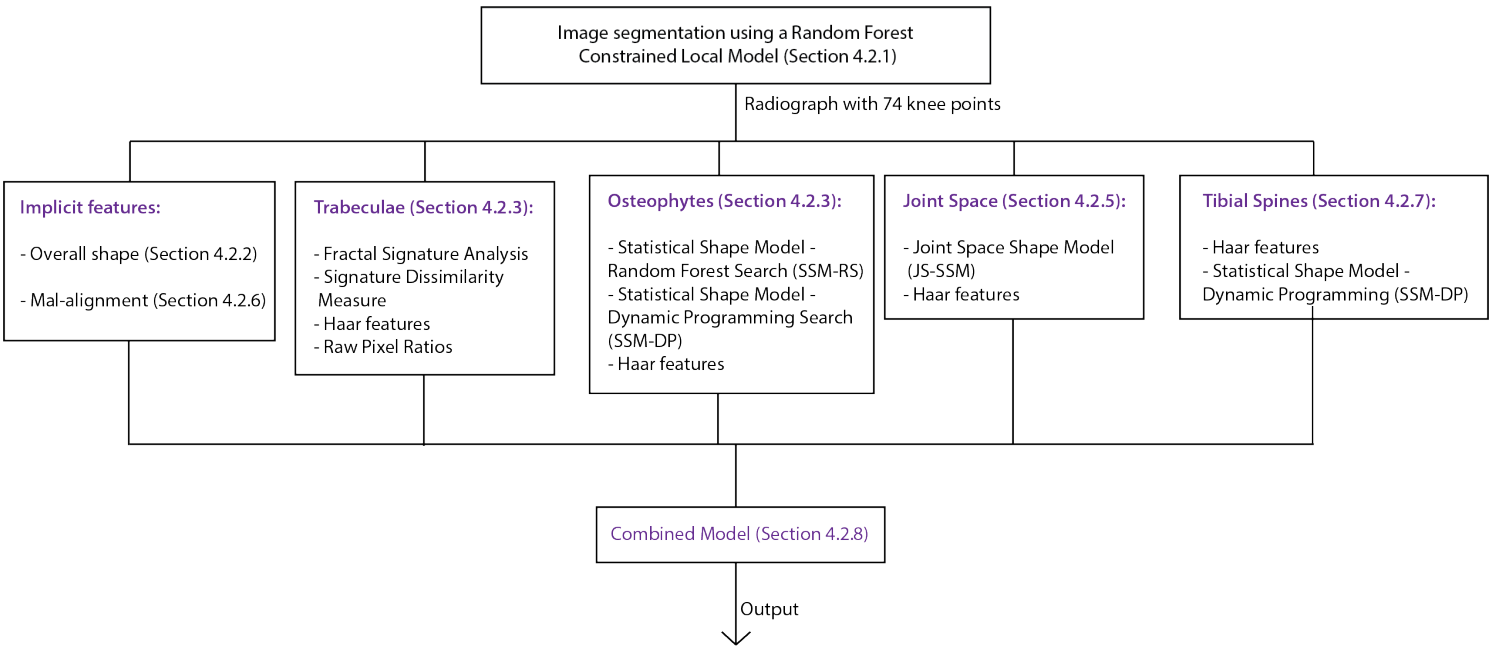

Figure 4.1: Chapted layout explaining the system and referencing method sections.

4.1 Data

Experiments were run using images from the OAI dataset (see Section 3.1). To compare the feature extraction methods four datasets were chosen to test for various outcomes. These include: i) object detection accuracy, with OA and non-OA images from male and female participants; ii) the presence of osteophytes, using manual OARSI osteophyte grades; iii) the presence of JSN, using manual OARSI Joint Space Narrowing (JSN) scores; and iv) OA/KL grades, which contains a range of KL grades that can be split into the two class (OA vs. non-OA) and multi-class problems (splitting by KL grade). The experiments using OARSI grades were used as partial testing for the feature specific methods (osteophytes and joint space narrowing). The data was taken from the subset of OA and non-OA images (747 OAI images) and was believed to be sufficient to compare the feature specific methods. The two-class detection uses the KL grade to split the images into OA (KL ≥ 2) and non-OA (KL ≤ 1). The statistics of the three sets are as follows:

- **Object detection** : 500 knees with a range of KL grades: KL0 - 111 (22.2%), KL1 - 135 (27%), KL2 - 90 (18%), KL3 - 121 (24.2%) and KL4 - 43 (8.6%). This divides into 246 non-OA, and 254 OA, of these 61.3% are female.

- **OARSI osteophytes** : 640 knees with a range of OARSI osteophyte grades (see Table 4.1) across the four sites: medial tibia, lateral tibia, medial femur, and lateral femur.

Table 4.1: OARSI osteophyte dataset statistics

OARSI (% samples)	medial tibia	lateral tibia	medial femur	lateral femur
0 (44.7%)	173	301	303	368
1 (30%)	327	231	95	114
2 (11.3%)	90	57	73	70
3 (14%)	50	51	169	88

- **OA/KL detection** 747 images, with KL grades: KL0 - 169 (22.7%), KL1 - 203 (27.2%), KL2 - 134 (18%), KL3 - 176 (23.6%), KL4 - 64 (8.5%). The OA vs. non-OA then becomes : OA - 374, non-OA - 373

– **OARSI JSN** : The 747 images were reduced to 704 images with JSN grades, the grades are split into each side of the knee (medial and lateral joint space). As the 747 contain predominantly medial OA knees, the lateral JSN scores are fairly low - see Table 4.2.

Table 4.2: OARSI JSN dataset statistics

OARSI (% samples)	Medial JS	Lateral JS
0 (44.7%)	315	671
1 (30%)	182	18
2 (11.3%)	142	13
3 (14%)	65	5

4.2 Methods

The methods from Chapter 3 can be applied to various radiographic features. This section details how the methods extract the relevant implicit and explicit OA features for the classification experiments. The section is split into object detection (RFCLM), the overall shape model built from the RFCLM, the individual radiographic features, and combined and comparison methods. In this chapter, we make reference to pilot experiments undertaken to acquire optimal parameters for the models. These experiments used 500 OA and non-OA images (250 samples per class). All shape and texture models ran through multiple iterations of the same 250 train, 250 test experiments on varying shape mode variation (iterating from 70% to 99%) or varying the number of samples included in the texture models (10 samples - 2000). The optimal results were found by plotting the Area under the ROC curves. The experiments were not included in this thesis to reduce space and because the experiments were primarily used for parameter optimisation and added little to the overall project aim of combining multiple OA features to create a stronger classifier.

4.2.1 Random Forest Constrained Local Model

All methods are based on the RFCLM output points (see Section 3.2.3). The algorithm is trained on manually annotated images, a series of annotation models were tested

during preliminary experiments. The models varied in amount of points and point locations (see Figures 4.2-4.5 below).

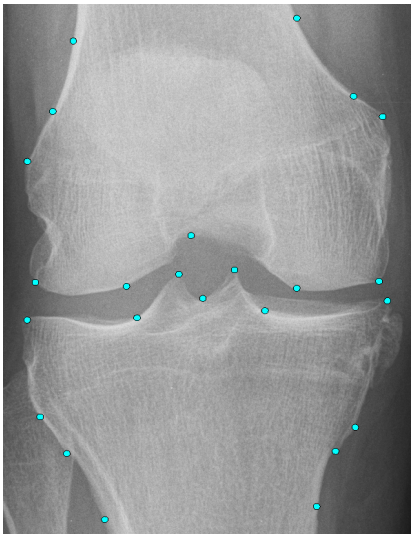

Figure 4.2: 24 point model of key shape features.

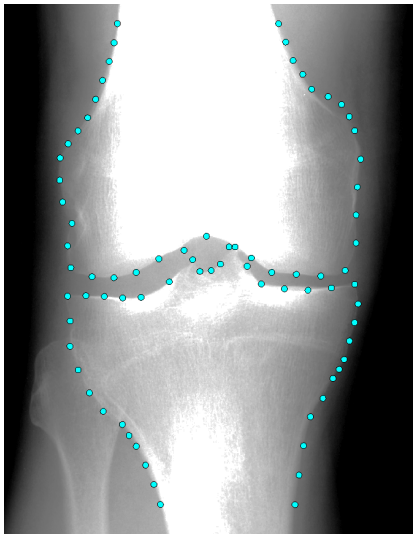

Figure 4.3: Original full model, 78 points.

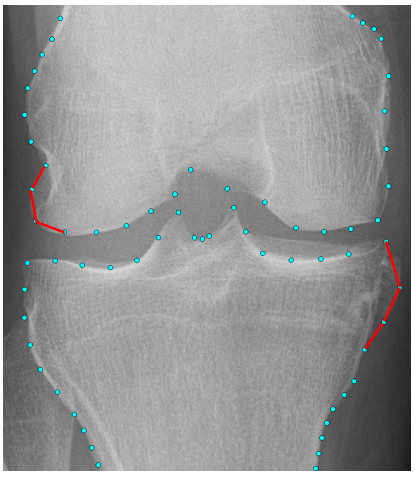

Figure 4.4: Revised 74 point model with deformations annotated.

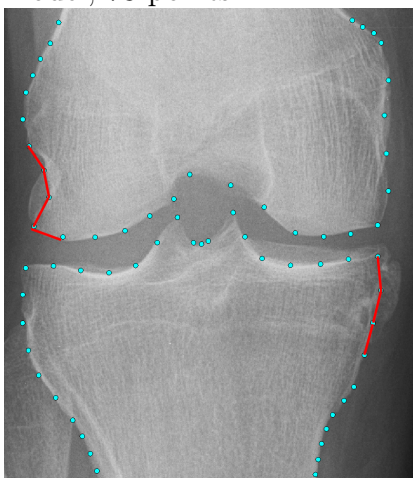

Figure 4.5: Revised 74 point model outlining 'normal' bone (see Figure 3.1).

The optimal points were found to be the 74-point model, which outlines the 'normal' shape of the tibia and femur. The annotations ignore bone remodelling and osteophytes where possible and outline the base knee (see the difference in Figures 4.4 and 4.5 above). This reduces errors (from a mean distance error of 0.47% to 0.39%). The RFCLM model contains a single global model to find two points along the joint

margins of the femur, and four local models that find 74 points around the knee outline (see Figures 4.6).

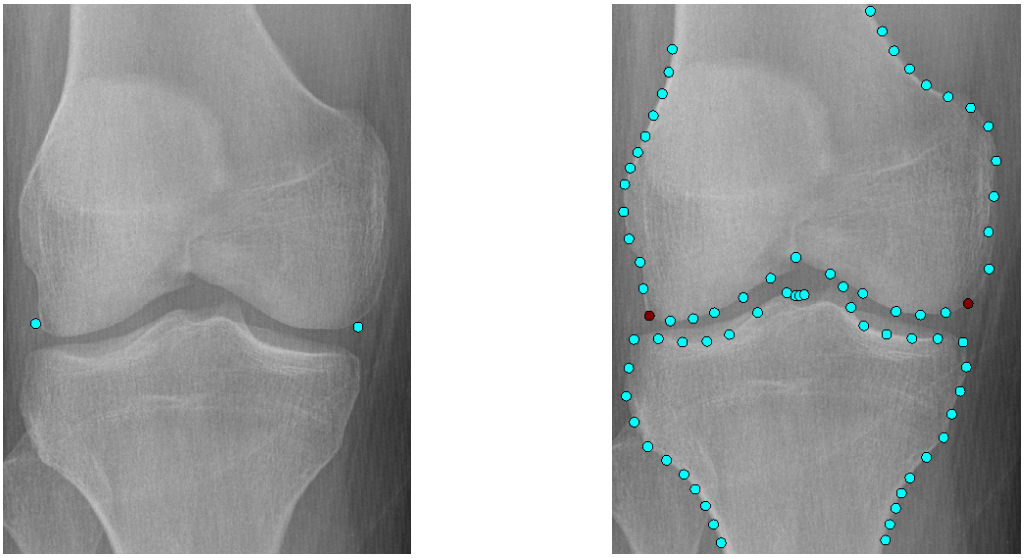

Figure 4.6: Global searcher points (left). Local searcher points, global initialised points are highlighted red (right).

The global model uses a Random Forest trained to find two points on the medial and lateral margins of the femur. The model estimates the global parameters to initialise the local Random Forest Constrained Local Model (RFCLM) searchers. The points were chosen on the femur margins to localise the search to centre of the knee joint, and estimate the scale by finding points across the width of the knee.

The local RFCLM models are split into three stages, with each stage initialised using the point positions from the previous stage. The frame widths were scaled each stage: 50 (coarse), 100 (medium) and 200 (fine). The fine stage was split to find the femur and tibia separately, fitting both points to the image before evaluating the fit. This allowed for the fine tuning of the tibia and femur separately (see Fig. 4.7). Extra parameters were optimised for each of these stages, these controlled: the search radius around the points, to set the search region for the Constrained Local Model (CLM) optimisation; and displacements of the point model in training, to displace the initial points during RFCLM training.

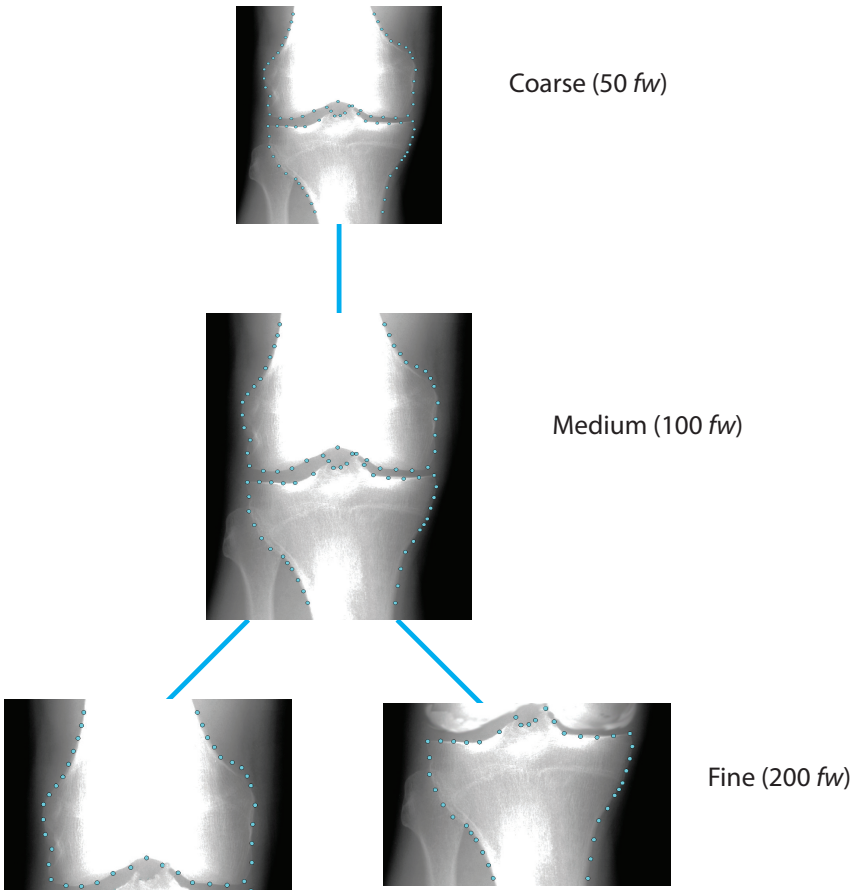

Figure 4.7: Illustration of local RFCLM searches with the models iterating over the various frame widths (fw).

4.2.2 Overall Shape

In the literature, methods such as [9] and [10] focused on implicitly capturing OA features. This analysis does not specifically target a certain part of the knee but extracts texture and shape features from the whole joint. This allows the detection of bone changes such as, alignment, JSN, attrition, and subtle changes in the overall shape/texture. A Statistical Shape Model (SSM) built on the RFCLM output points was chosen to capture similar information of overall shape change.

The SSM is built using a Principal Component Analysis (PCA) algorithm, the number of shape modes extracted is controlled by the amount of shape variation specified. For the overall shape experiments a 99% variation was found to be optimal, which equalled 44 shape modes. This parameter was optimised through pilot experiments on a smaller

subset of training and testing images.

4.2.3 Trabeculae

Trabeculae are narrow and tightly packed in small regions of subchondral texture. Current methods that analyse these features focus on extracting pixel intensity and gradient variation values (see Section 3.4.1). The methods used to extract the trabeculae features are: Augmented Variance Orientation Transform (AVOT), Variance Orientation Transform (VOT), wedgeVOT, Signature Dissimilarity Measure (SDM), Haar-like features, and Raw Pixel Ratios (RPR).

Region Sampling

To extract the relevant trabeculae features a region of interest (ROI) must first be selected in the image. We use the region sampling explained in Section 3.4 with two points of the RFCLM output and the vector r projected between them. The location of the fibula (lateral side of the tibia) overlaps with the trabeculae structure beneath the lateral tibial plateau and can produce noisy texture features (see Fig. 4.8). The algorithms focus on the medial side of the tibia to analyse trabeculae, targeting medial OA in the knee, which has a much higher prevalence in the OAI dataset and overall population [104] [105].

Figure 4.8: Fibula overlapping the lateral side of the tibia

Experiments were run on various locations (avoiding the lateral tibia), including the femur medial and lateral sides, and central tibia (see Figures 4.9-4.11). These areas all achieved lower accuracies than the medial tibia region, and added no extra information when combined with the medial tibial texture in preliminary experiments.

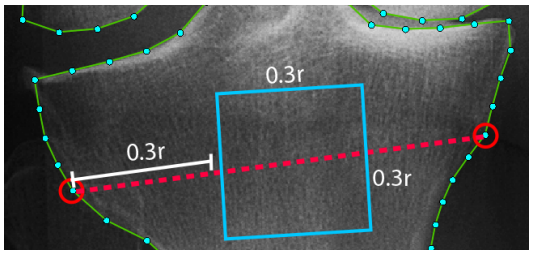

Figure 4.9: Central placed tibia ROI.

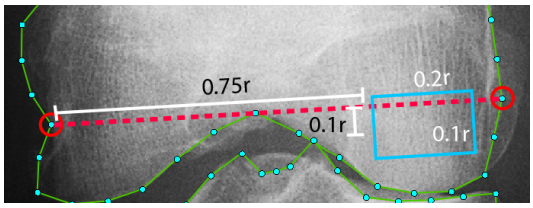

Figure 4.10: Femur medial ROI.

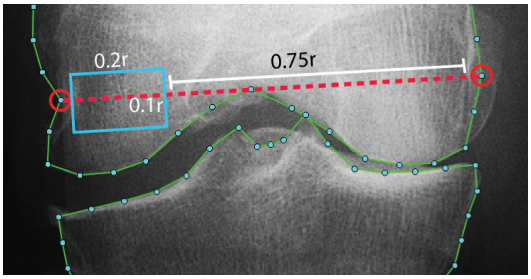

Figure 4.11: Femur lateral ROI.

Following the literature, the region was placed below the tibial plateau to avoid thickened bone texture from sclerosis. The region was placed relative to the angle of the tibia, and was shifted $0.8r$ along and $-0.1r$ above the vector. The region size varied depending on the texture sampling method (see Figures 4.12, 4.13 below), with the FSA and SDM methods following the best reported parameters from the literature: 256×256 size pixel region, with a region size of $0.2r \times 0.2r$. For the methods which used raw pixel intensity data (RPR and Haar features) a region of 256×125 pixels, and a region size of $0.2r \times 0.1r$ was found to be optimal.

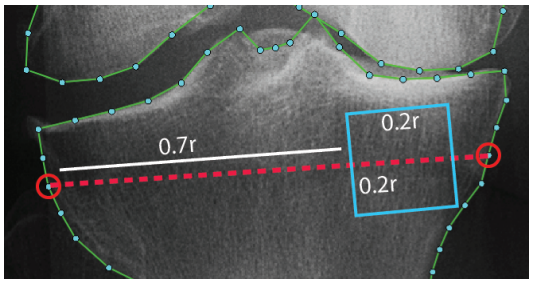

Figure 4.12: Larger ROI used for FSA methods.

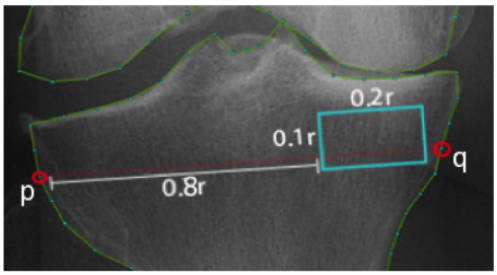

Figure 4.13: ROI used for Haar features and RPR.

Fractal Signature Methods

The Fractal Signature methods are: VOT (Section 3.4.1), wedge VOT (Section 3.4.1), and AVOT (Section 3.4.1). All three extract features of pixel intensity variation and produce a series of Hurst coefficients from fitted line gradients for the scales (subsets of log-log points per angle).

The wedge VOT algorithm had a variable number of angles to split the circle sampling region; to keep the methods consistent 24 angles of 15° was used. The distances were kept the same as the VOT algorithm parameters from the literature (13 pixel distances per angle). The VOT and AVOT each sampled pixels along 24 angles; the VOT used fixed radii size of 4 and 16 pixels. The AVOT had variable radii sizes depending on the size of the texture region, however, as the texture regions were fixed at 256×256 the sizes were the same as the VOT method.

Signature Dissimilairty Measure

The SDM method (described in Section 3.4.3) builds histograms of trabeculae roughness and orientation. Roughness is constructed using a Gaussian sampling (see Figures 4.14-4.16 below), with the orientation taken from the edges of the trabeculae (see Fig. 4.17). All parameters were taken from the literature [8], using a sampling size of $\delta = 1.1^n$ where $n = 1, ..., 25$. The histograms of trabeculae roughness and orientation are used to train and test the classifiers.

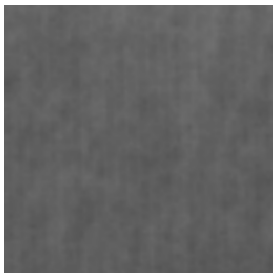

Figure 4.14: Original trabeculae image.

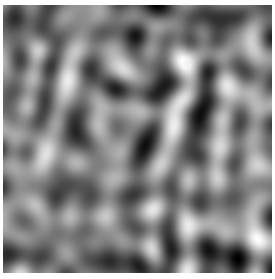

Figure 4.15: D_2 minimum gradients image.

Figure 4.16: D_2 maximum gradients image.

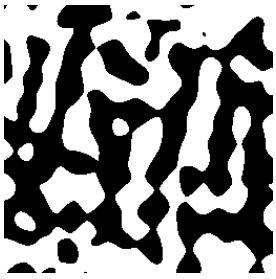

Figure 4.17: Binary image of trabeculae regions.

Haar features

Haar features (explained in Section 3.4.4) measure intensity patterns based on differences between the mean intensity in nearby rectangular regions of the image. The classifiers are trained on the features gathered from the optimal ROIs beneath the tibial plateau.

Raw Pixel Ratios

The RPR method (explained in Section 3.4.2) extracts multiple samples of raw pixel intensities from each image. The classifier trains on the ratios of random pixel pairs in the samples extracted (see Fig. 4.18).

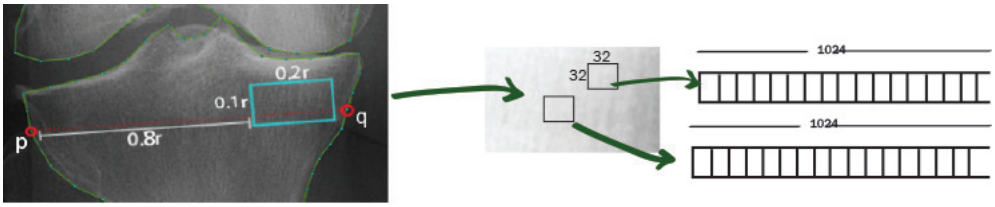

Figure 4.18: Illustration of RPR taking 32×32 pixel samples. The method finds a ROI beneath the medial tibial plateau, from that segmented region we sample 670 samples (of size 32 x 32 pixels) per image. These samples are saved as 1024 pixel vectors to train the Random Forest classifiers.

The number of samples taken per region (670) and the size of the regions (32×32) was optimised during preliminary experiments. Each of the 670 samples were trained as separate examples, in testing the mean output (over all samples) is chosen as the image classification.

4.2.4 Osteophytes

Osteophytes vary in size and shape between participants irrespective disease severity, due to this there is a need for algorithms that are sensitive to variable shape change that are not limited to detect objects that follow a typical progression. To extract the relevant features we use the shape methods explained in Section 3.3, and Haar features to measure edges and gradient change.

To compare the analysis to OARSI grades we extract osteophyte features that occur along the joint margins (see Figure 4.19).

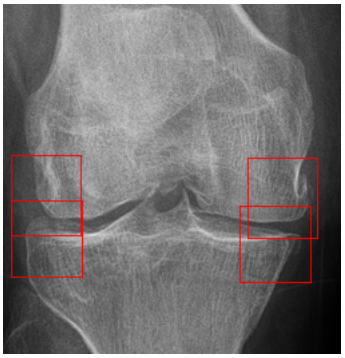

Figure 4.19: The red squares illustrate the four regions to detect marginal osteophytes.

The experiments compare features extracted from: a Statistical Shape Model built from a RFCLM search (SSM-RS), a SSM built from DP detected contours (SSM-DP), and Haar-like features (Haar).

Statistical Shape Model - Random Forest Constrained Local Model Search (SSM-RS)

To build an osteophyte SSM-RS model the images have manually annotated points along the joint margins. A new set of points was placed on the images from the RFCLM object detection experiments. The osteophyte point model included 44 points added to capture the shape variation of the osteophytes. The points were added along the corners of the joint space and the medial and lateral sides of the tibial and femoral epiphysis (see Figure 4.20).

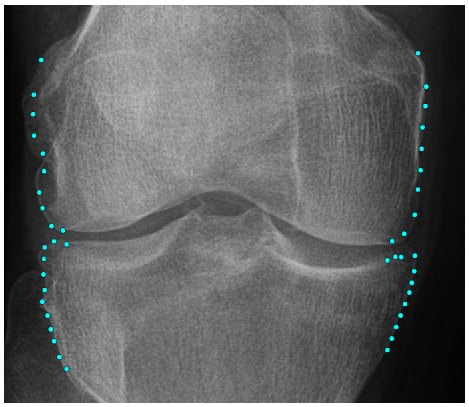

Figure 4.20: The 44 points outlining the joint margins.

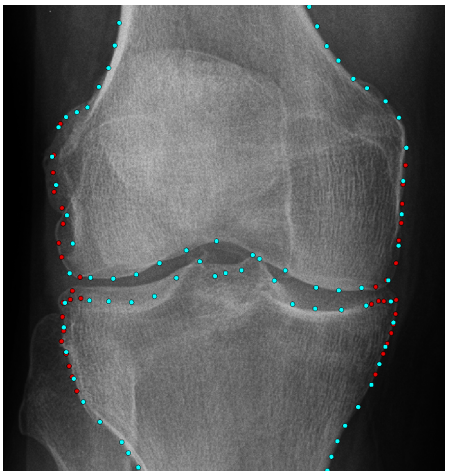

Figure 4.21: The 74 base knee points with the 44 SSM-RS points highlighted in red.

The 44 points were attached to the 74-point model and a new RFCLM set up to find 118 points. The 44 osteophyte points are extracted from the output. Preliminary experiments applied a sequence of models to optimise osteophyte detection. The optimal used the original 74-point model found in the coarse stage, and the larger 118-point model placed in the medium and fine stages. The points are split per object in the fine stage, with 59 on the tibia object and 59 on the femur (see Fig. 4.21). When training the SSM on the extracted osteophyte points 99% of the shape variation contained the optimal amount of features (50 shape modes).

Statistical Shape Model - Dynamic Programming Search (SSM-DP)

This model builds an SSM from the shape points found using the Bézier curves and Dynamic Programming optimisation explained in Section 3.3.2. The algorithm first interpolates a set of points between a set control points from the RFCLM output, for this we use points either side of the areas marginal osteophytes develop (see Fig. 4.22). Each of the four regions had 12 points interpolated to capture enough shape detail over the margins (see Fig. 4.23).

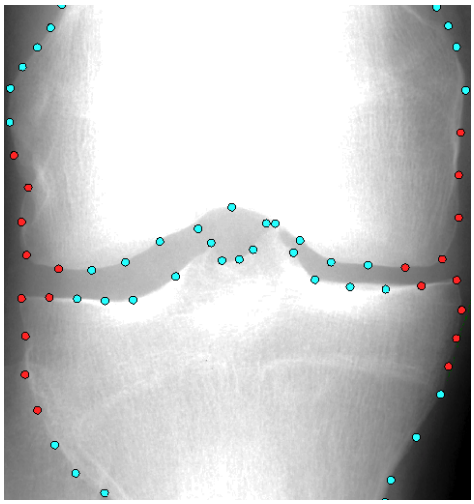

Figure 4.22: Control points (red) selected to interpolate new osteophyte points.

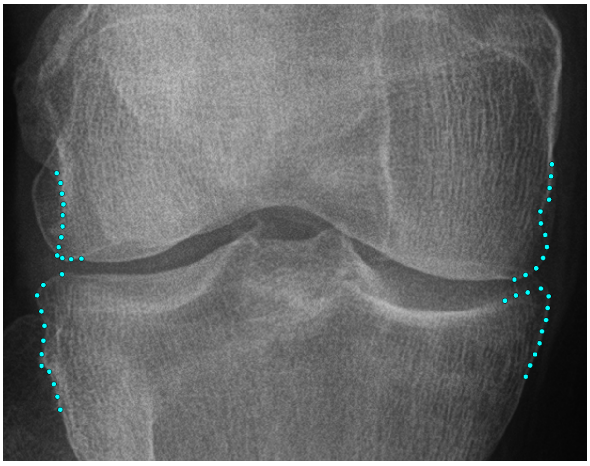

Figure 4.23: The 48 points found by the SSM-DP algorithm.

Dynamic programming then fits the points to the strongest edge by optimising the Equation 3.3. Through prior experiments we found that the best value for the curve constraint α was 0.2 as the osteophytes varied in height rapidly and often extended across the corners of the joint (see Figures 4.24-4.25 below).

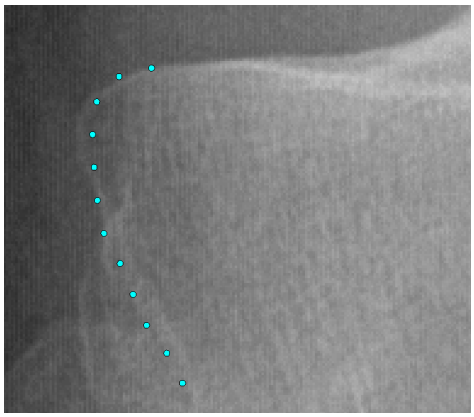

Figure 4.24: Contours with k = 0.9.

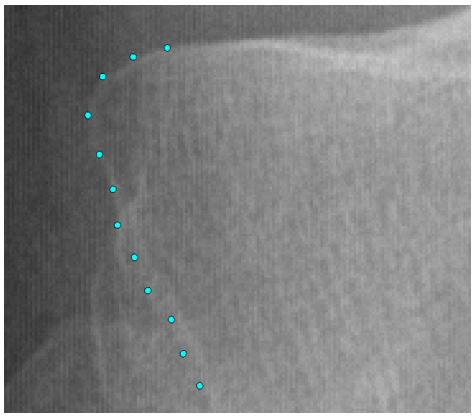

Figure 4.25: Contours with k = 0.1.

The SSM was built on the 48 contour points (see Fig. 4.23) and used 85% shape variation (30 shape modes).

Haar Features

Similar to the trabeculae feature extraction (see Section 4.2.3) Haar features (explained in Section 3.4.4) are used to extract edge and shading information across the osteophyte regions. The regions were placed over the four margins (medial tibia, lateral tibia, medial femur and lateral femur), the method used two points close to the margin to position and scale the regions (see Fig. 4.26). The region was shifted so the centre lies over the corner points ($x = -0.5r$). The size used is 25×25 pixels taken using the vector r to scale the texture sample.

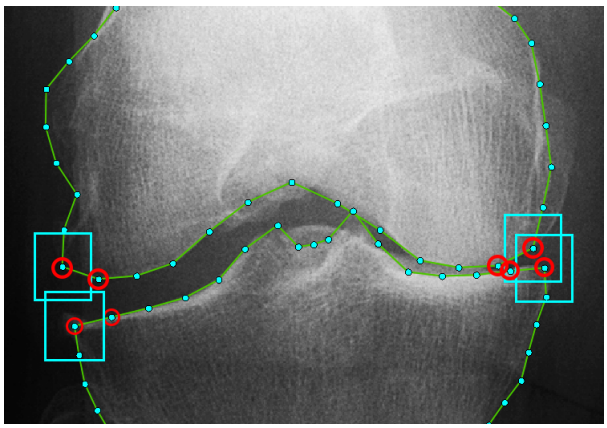

Figure 4.26: The four regions placed to sample osteophyte texture. The points (circled in red) are used to shift each region relative to the margin.

4.2.5 Joint Space

JSN and Joint Space Width (JSW) change is a prevalent factor of OA development [22]. The work by Duryea et al. [63] looks at the distances from the front line of the plateau, quantifying OA change using a series of joint space widths (xJSW) along set distances of the tibial plateaus. The measurements analyse the widths from the front brightest edge of the plateaus to keep the calculations consistent and less susceptible to knee rotation. Our methods expand on this to include all of the joint space shape by building a Joint Space Statistical Shape Model (JS-SSM), and extracting Haar features in the region.

Joint Space Shape Model

The JS-SSM uses the same edge feature extraction as the osteophyte SSM-DP, interpolating points between fixed points on either side of the femoral condyles and tibial plateaus (see Fig. 4.27).

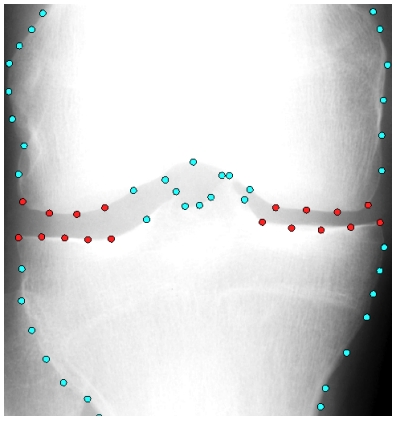

Figure 4.27: The control points (red) used to interpolate new edge points.

The optimal number of points across the femur and tibia surfaces was 10, with $\alpha = 0.2$ (see Fig. 4.28). This α achieved the highest in preliminary experiments, with AUCs of 0.851 $\alpha = 0.4$, to 0.858 $\alpha = 0.2$.

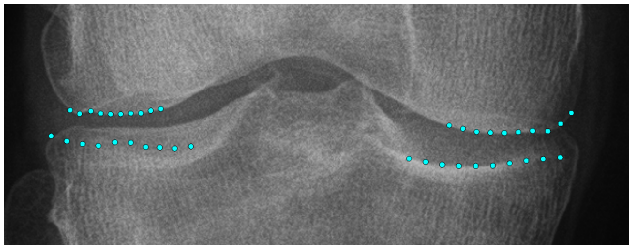

Figure 4.28: Contours fitted to all four joint space surfaces.

We then built a SSM from the four curves. The SSM used 99% shape variation, 18 shape modes. The overall shape model (see Section 4.2.2) includes some JSN information, the SSM-DP was included to interpolate extra detail along the joint space.

Haar features

To analyse extra joint space information, texture samples for Haar features (explained in Section 3.4.4) were also extracted. The regions were placed using the start and end

control points from the JS-SSM contours (see Figure 4.27). Two regions (see Fig. 4.29 below) of 25x15 pixels were sampled to span the medial and lateral joint spaces.

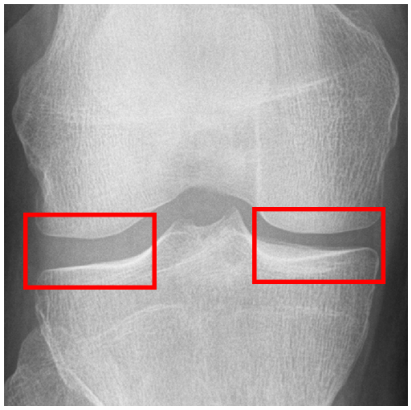

Figure 4.29: Medial and lateral joint space regions (red boxes) extracted.

4.2.6 Mal-alignment

The joint mal-alignment measures the angle of the tibia and femur either in projected lines across the joint space or the angle between the intersections of lines projected down the diaphysis of the two bones. The mal-alignment is also apparent in the positions of the two bones. This information we extract implicitly (see Figures 4.30-4.31) in other feature extraction methods i.e. overall shape (Section 4.2.2) and joint space (Section 4.2.5).

Figure 4.30: Overall shape SSMs with femur mal-alignment.

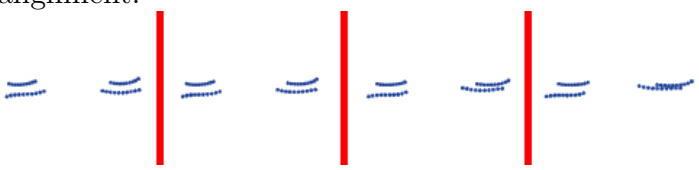

Figure 4.31: JS-SSM with JSN and mal-alignment.

4.2.7 Tibial Spines

Literature on the association of tibial spines with OA is conflicting. These features are included in the experiments to expand on current analysis that compares height and spike angle [31] [71]. To extract features from the tibial spines we use: DP extracted contours (SSM-DP) and Haar features.

Haar features

Haar features (explained in Section 3.4.4) are used to extract edge and shading information across the tibial spines and intercondylar notch. The region is placed using two points either side of the spines (see Fig. 4.32), with 19×19 pixels sampled. The intercondylar notch was included from the overall of tibial spines with the notch in knees with severe JSN (see Fig. 4.33).

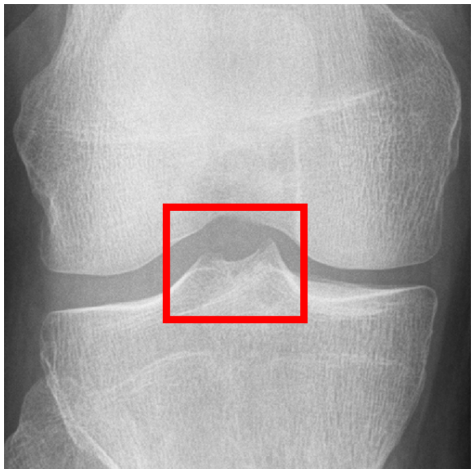

Figure 4.32: ROI texture sampling of tibial spines.

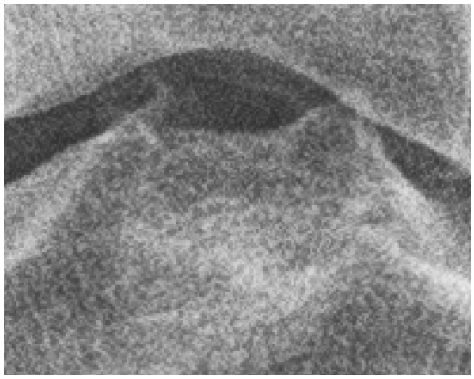

Figure 4.33: Tibial spines overlapping the intercondylar notch.

Tibial Spine Contours

The SSM-DP extracts contours along the tibial spines using similar points to extract the Haar features (see Fig. 4.34). A set of 24 points were found to be optimal in capturing the tibial spine shape (see Fig. 4.35), the α was set to be 0.1 so that the contour would follow the edges over the peaks of the spines.

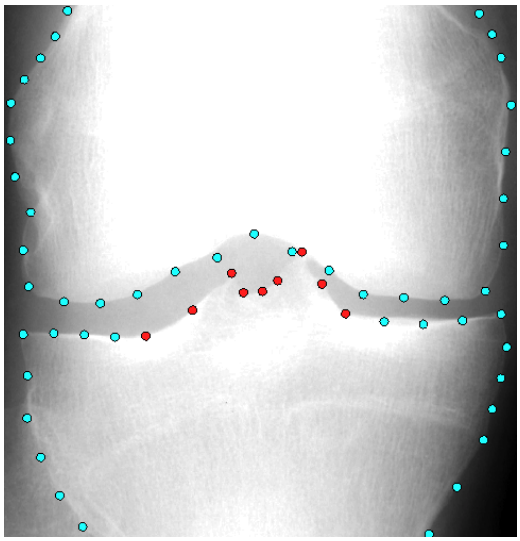

Figure 4.34: Control points for interpolating new tibial spines edges.

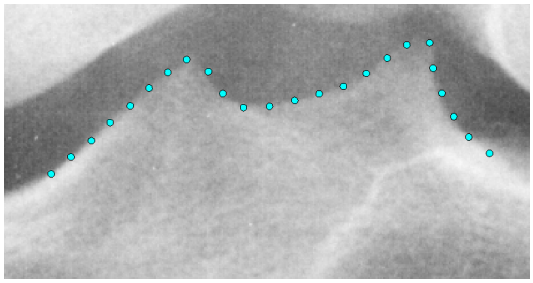

Figure 4.35: SSM-DP found tibial spines points, k = 0.1.

A SSM is built from the tibial spine points extracted from the RFCLM output. This included 9 points between the tibial plateaus (see Fig. 4.34). Specific tibial spine annotations were used during preliminary experiments, but the extra points had with no significant increase in classification accuracy. The SSM built from the points contained 99% variation (8 shape modes).

4.2.8 Combined Model

The combined model uses the optimal methods from each experiment. This includes features that are best evaluated by combining shape and texture methods, or using a single method that achieves comparable classification results. The optimal OA features are combined one at a time, to find the combination of radiographic features that achieves the highest detection accuracy of OA. We then compare the optimal model with the WND-CHARM algorithm by Shamir et al. [9] (explained in Section 3.4.5). The parameters for WND-CHARM were taken from the literature.

4.3 Experiments

The experiments compare the various extracted features using the accuracy of the Random Forest classifiers (described in Section 3.5.1). The validity and mean accuracy

of the classifiers is reported using 5-fold cross validation (see Section 3.5.2). The symbol ** will be used in all the tables to state that the best accuracy is significant compared to the other results reported in the same table. The significance will be based on the difference between the mean values and the standard deviation of the highest and second highest result in the table.

For the two-class shape model experiments a Linear Discriminant Analysis (LDA) was run. This outputs a shape mode that spans the shape space between the two classes (OA and non-OA). These LDA shape modes have been included in the relevant sections.

4.3.1 Random Forest Constrained Local Model

The RFCLM object detection uses the set 500 images split into 250 training and 250 testing examples. The results are in point-to-curve Euclidean distance error (see Fig. 4.36). This error was chosen over point-to-point and curve-to-curve, to both minimise the penalty of finding the shape but not the exact point positions (point-to-point) and to reduce computation time over comparing points along each curve between the manual and automated points (curve-to-curve).

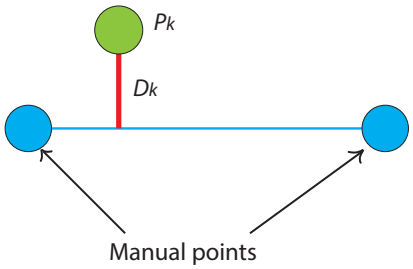

Figure 4.36: The point-to-curve distance error D_k is found between the point P_k and the closest part on the curve (blue line between the manual points).

The RFCLM models achieved a high accuracy across the 250 images: mean: 0.39% (0.29mm)$\pm$0.14mm, median: 0.34% (0.26mm) and 95th percentile: 0.72% (0.54mm). The error Cumulative Distribution Function (CDF) in Figure 4.38 shows that 95% of the images have a mean point-to-curve distance of less than 1% of the relative distance (see Fig. 4.37). Following [93] the mean point-to-curve distance was converted to mm

by assuming a mean knee width of 75mm. These results are similar to those presented by Lindner et al. [93], though on a different dataset.

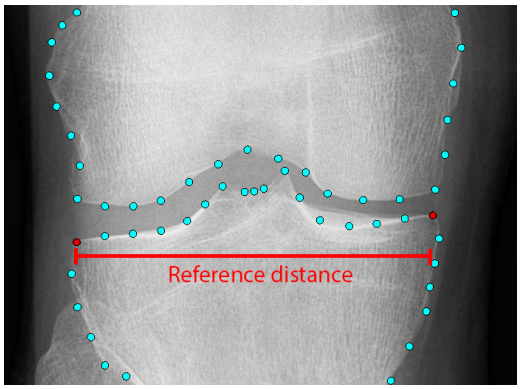

Figure 4.37: The error is taken as a percentage of the reference distance (between the two red points).

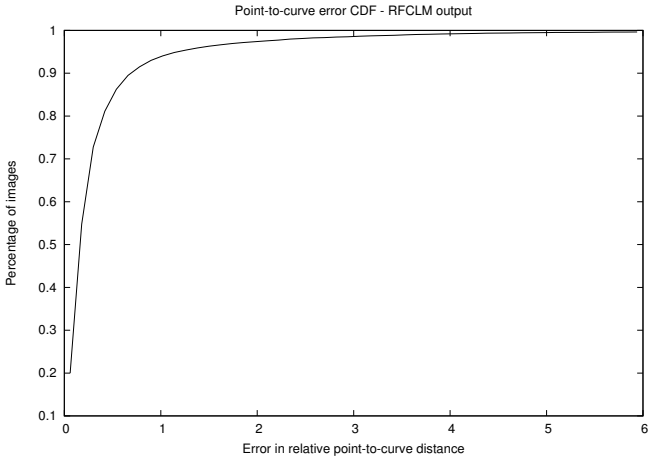

Figure 4.38: The CDF shows the relative distance error of all 250 images.

The images that fall in the last 5% are likely to be knee shapes that are untypical or unseen in the 250 training images, examples of the RFCLM poor output points can be seen in Figures 4.39 below.

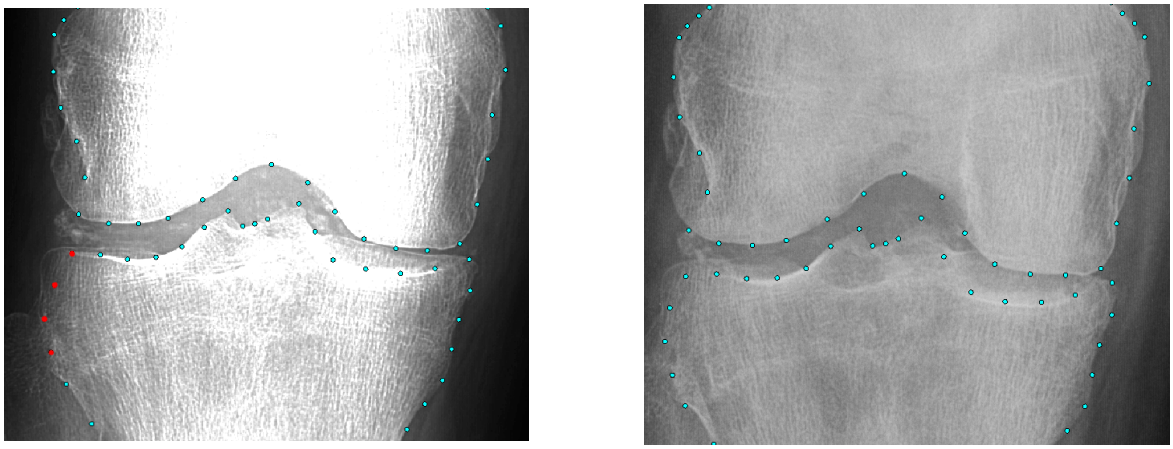

Figure 4.39: Examples of poor RFCLM output.

4.3.2 Overall Shape

The shape model taken from the RFCLM output points was analysed using OA/KL grades, splitting the experiments into OA vs. non-OA (see Figure 4.55) and KL grades (see Table 4.9). The LDA shape information illustrates the effect of moving along a

vector projected between the two classes in shape space (see Figure 4.40). The shape change shows JSN and spiking of the tibial spines in the knee. The AUC for the two-class problem is 0.84, whilst the KL grades achieved a mean accuracy of: 33.9% ±1.8%.

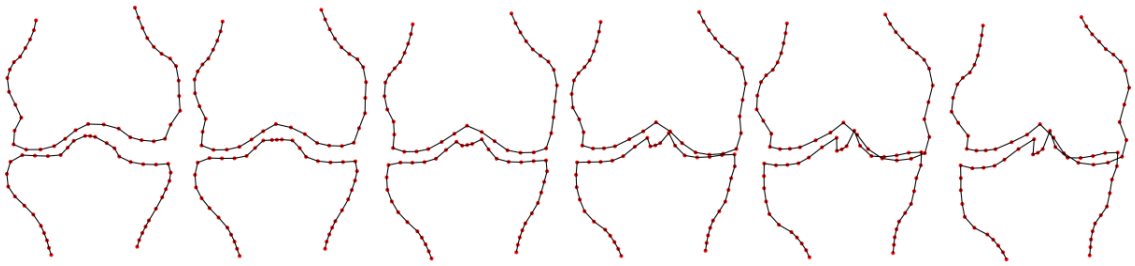

Figure 4.40: LDA shape model of the differences between the non-OA (left) and OA (right) classes.

4.3.3 Trabeculae Comparison

All trabeculae methods were compared using the OA/KL grade 747 images. The comparisons were all run on detection of OA vs. non-OA images (see Fig. 4.41) and multi-class features (see Table 4.3). The results show that the best method was the RPR method in splitting the data, with an AUC of 0.703. Combining the RPR method with the best FS method (AVOT) added nothing to the accuracy with 0.703 RPR, and 0.702 combined texture AUC. The multi-class results for all texture features are fairly poor (27.7% - 30.8%). This is likely because the KL grades are based on shape analysis. For the optimal trabecular model, the RPR was chosen because of the strong two-class AUC.

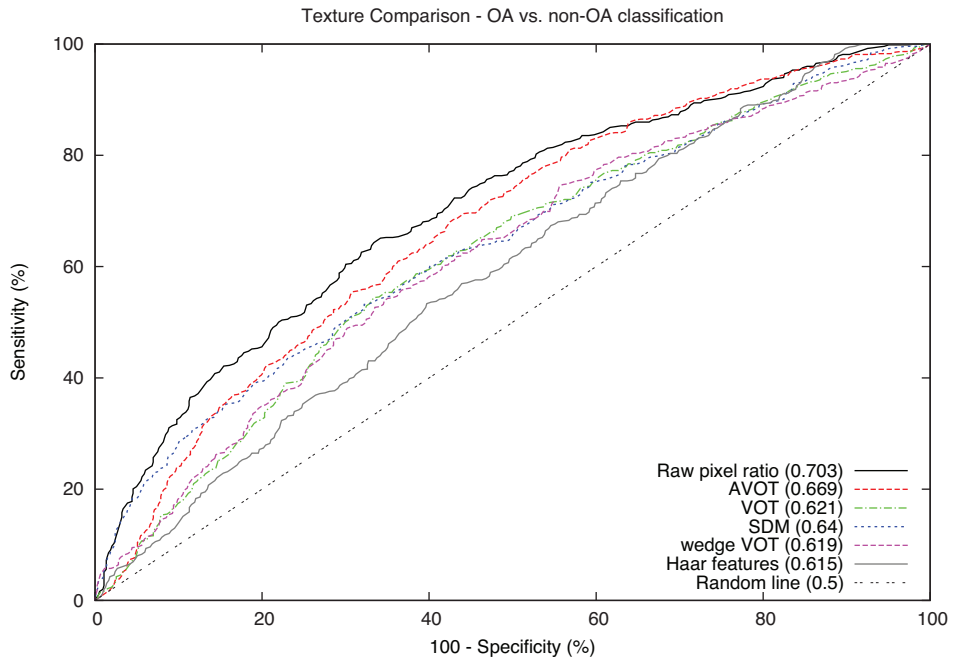

Figure 4.41: Comparison of the various texture methods in detecting OA vs. non-OA images.

Table 4.3: Trabeculae AUC and multi-class performance

Analysis Method	AUC (stdev.)	multi-class % (stdev.)
Haar	0.615 (0.012)	28.1 (0.5)
wedge VOT	0.619 (0.018)	29.1 (0.3)
SDM	0.64 (0.009)	**30.8 (0.7) ****
VOT	0.621 (0.003)	28.8 (0.8)
AVOT	0.669 (0.008)	28.8 (0.8)
RPR	**0.703 (0.004)****	25.7 (0.2)

All FSA methods were compared with comparable parameters, each used 24 angles and pixel distances (d_i) between $4 \leq d_i \leq 16$ (13 in total). Out of the three methods the AVOT achieves the highest AUC (0.669 AVOT, to 0.623 VOT and 0.619 wedgeVOT). These results show that measuring intensity differences along specific angles, and not a wider sampling region achieves a better accuracy. The main difference in the AVOT algorithm, with the texture regions fixed at 256×256, is the number of scales the method splits the distances per angle into: AVOT using 13 scales (pixels 4 to 16), and VOT using 11 scales (pixels 6 to 14). This improved the AUC from 0.623 (VOT) to

0.669 (AVOT).

4.3.4 Osteophyte Comparison

To test the accuracy of the osteophyte methods, we ran various experiments:
(i) RFCLM object detection on finding extended 118 point model
(ii) Detecting osteophytes via OARSI grades (0, 1 vs. 2, 3)
(iii) Detection of OA and non-OA images (374/372)
(iv) Multi-class classification into grades (KL0-4) using the best method from the OA vs. non-OA experiments.
We report independent and combined results for each method: Statistical Shape Model - RFCLM Search (SSM-RS), Statistical Shape Model - Dynamic Programming contour detection (SSM-DP), and texture analysis using Haar-features (Haar).

Statistical Shape Model - Random Forest Constrained Local Model Search (SSM-RS) Object Detection

Using a similar split to the RFCLM experiments (explained in Section 4.3.1) we used the 500 images with a 250/250 (train/test) split. The overall point-to-curve distance error is worse than the base point model with 0.39% (74 point model) to 0.836% (118 point model). This error is likely to be caused by the variability of the extra osteophyte shape (see Fig. 4.42).

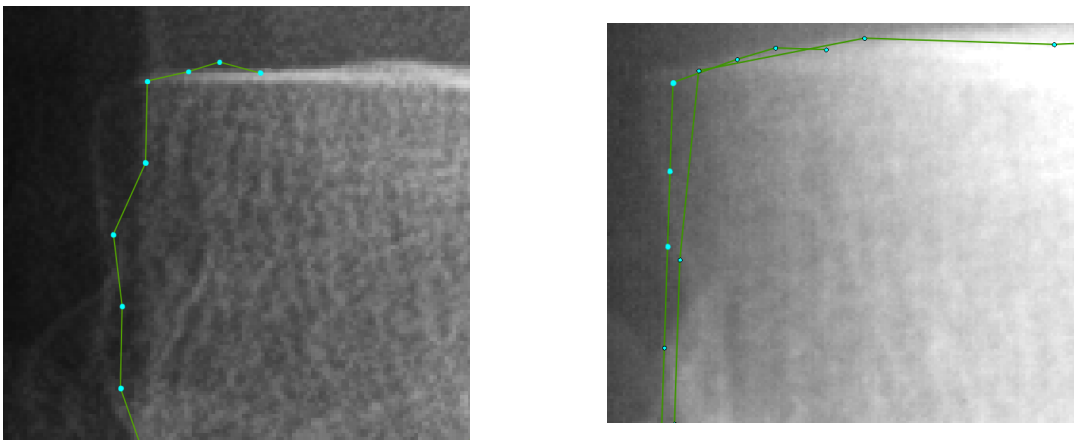

Figure 4.42: Examples of the SSM-RFCLM not finding the correct edges of the osteophytes.

Osteophyte Detection

We ran each method over the four regions (lateral femur, medial femur, lateral tibia and medial tibia) detecting osteophytes in each region separately. The mean AUC from each region per method is reported in Table 4.4 below. The best independent classifier is the Random Forest (RF) based on Haar-features with AUC 0.77. By combining all methods we achieve AUC 0.85.

Table 4.4: Osteophyte detection performance

Analysis Method	mean AUC (stdev.)	mean multi-class % (stdev.)
SSM-RS	0.756 (0.056)	47.8 (0.9)
SSM-DP	0.663 (0.072)	46.3 (1.4)
Haar features	0.771 (0.053)	54.1 (2.0)
SSM-RS + SSM-DP	0.769 (0.065)	50.9 (2.0)
SSM-RS + Haar	0.826 (0.015)	55.7 (1.6)
SSM-DP + Haar	0.806 (0.015)	**60 (1.6)****
All methods	**0.846 (0.014)**	55.8 (1.9)

The multi-class automated KL grades compared to the gold standard achieved weighted kappas of: 0.12(0.07-0.17) lateral femur, 0.41(0.35-0.47) medial femur, 0.26(0.2-0.32) lateral tibia, 0.24(0.18-0.3) medial tibia. The mean weighted kappa (kw) of these is 0.26, which is lower than the manual OARSI osteophyte grading (0.64-0.78) [41] [40]. This lower accuracy could be due to the limited numbers in the 2-3 grades, with lateral marginal osteophytes having between 51-88 in grades 2 and 3. Medial marginal osteophytes have more samples (50-169) but are still lower than the number of samples with grades 0 and 1 (95-327).

Osteoarthritis vs Non-Osteoarthritis

The cross validation mean AUC results are shown in Table 4.5 and ROC curves in Figure 4.43. The best extracted features are the Haar features, with an AUC of 0.92. Combining this method with both the SSM-DP and SSM-RS method slightly improves the AUC to 0.929.

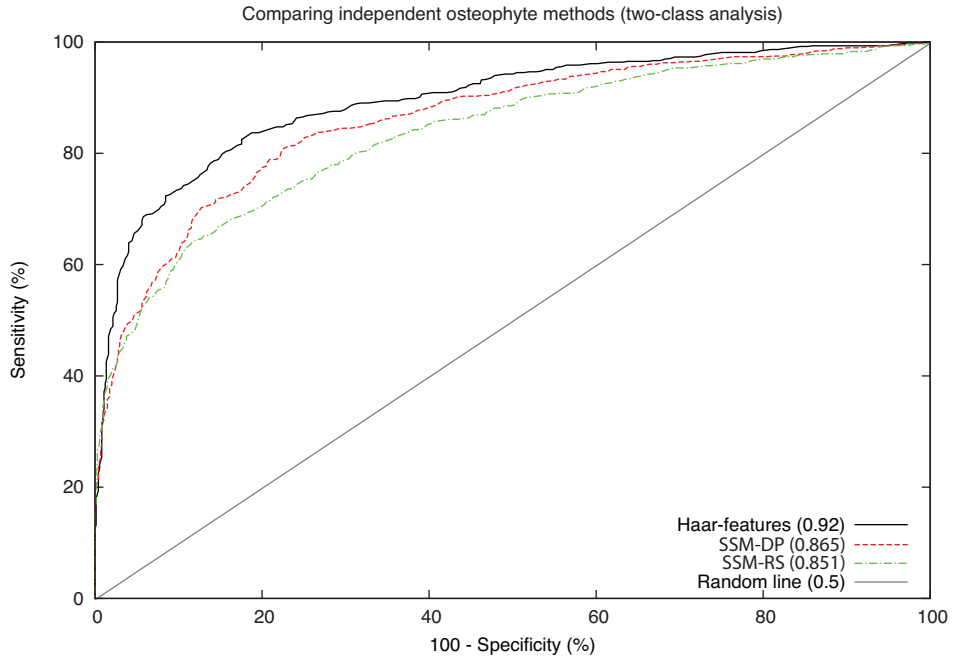

Figure 4.43: Comparison of the various osteophyte feature extraction methods in detecting OA vs. non-OA images.

The methods were combined to improve the detection from an Area Under ROC Curve (AUC) of 0.92 Haar features, to 0.929 fully combined (see Table 4.5 below). For further comparison, the overall shape model was included with the osteophyte features. The SSM-RS features were found to add less accuracy than the overall shape with AUCs of 0.929 (Haar, SSM-DP and SSM-RS) to 0.933 (Haar, SSM-DP and overall shape). The optimal osteophyte features are Haar features and SSM-DP contours (see Fig. 4.44). The SSM-RS was removed because the features captured focus primarily on the shape of the bone (see Fig. 4.45). These features are captured with extra detail about the JSN and alignment in the overall shape model.

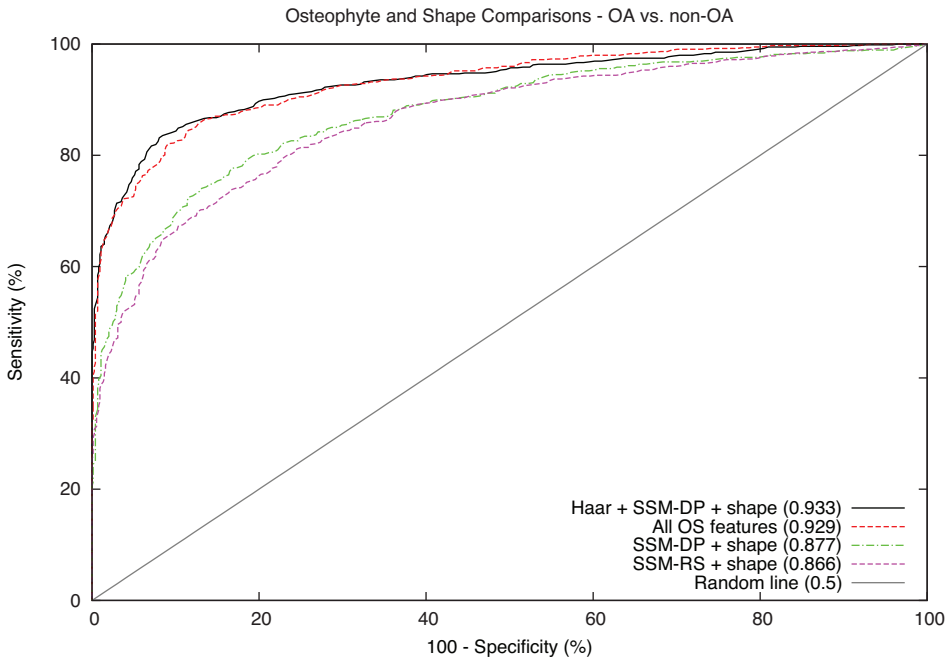

Figure 4.44: Comparison of combined osteophyte and shape features.

The LDA shape vectors of the SSM-RS (see Fig. 4.45) and SSM-DP (see Fig. 4.46) show that the shape change is mainly shown in the JSN of the knee, with some marginal osteophytes and bone remodelling in the severe OA cases.

Figure 4.45: LDA shape model of the osteophyte SSM-RS between the non-OA (left) and OA (right) classes.

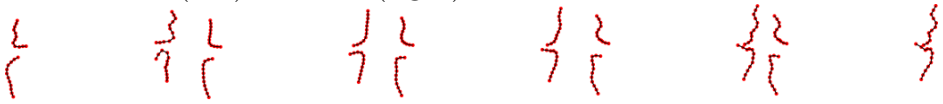

Figure 4.46: LDA shape model of the osteophyte SSM-DP between the non-OA (left) and OA (right) classes..

Multi-class Experiments

The features were combined by training three separate RF classifiers on the osteophyte features, and taking the mean of the outputs as the final classification. The

features were concatenated to train and test a single classifier in preliminary experiments with no significant improvement. The combination of all features improved the multi-class overall probability from: 47.8% ±0.3 (Haar features), 40.7% ±0.5 (SSM-DP), 43% ±0.8 (SSM-RS), to 50.2% ±0.5 (combined features). Table 4.9 shows the per-class accuracy and the overall probability that the correct class is chosen. Replacing the SSM-RS features with the overall shape features had minimal effect when taking into account the standard deviation (stdev.) of the values, with the combined osteophyte features and SSM-RS replacement with overall shape achieving similar accuracies (50.6% ± 1.1 to 49.6% ± 1.7).

Table 4.5: Osteophyte AUC and multi-class performance

Analysis Method	mean AUC (stdev.)	multi-class accuracy (stdev.)
SSM-RS	0.851 (0.001)	43 (0.8)
SSM-DP	0.865 (0.003)	40.7 (0.5)
Haar features	0.92 (< 0.001)	47.8 (0.3)
Shape	0.843 (< 0.001)	33.9 (1.8)
SSM-RS + SSM-DP	0.872 (0.002)	39.7 (0.7)
SSM-RS + shape	0.866 (0.001)	41.9 (0.1)
SSM-DP + shape	0.877 (< 0.001)	42.3 (0.1)
SSM-RS + Haar	0.926 (0.002)	48.5 (1.5)
SSM-DP + Haar	0.929 (0.001)	48.3 (1.6)
All OS	0.929 (0.002)	50.2 (0.5)
SSM-DP + Haar + shape	**0.933 (0.001)**	49.6 (1.7)
All OS + shape	0.93 (0.002)	**50.6 (1.1)**

The algorithms did miss some shape variation in both SSM-DP (see Fig. 4.47) and SSM-RS (see Fig. 4.42). SSM-RS tends to stick to the edges of the bone; this is likely because the RFCLM uses the mean shape seen in the training set. SSM-DP finds the strongest and brightest edges, often missing some faint osteophyte edges and affected by varying image contrasts (see Fig. 4.48).

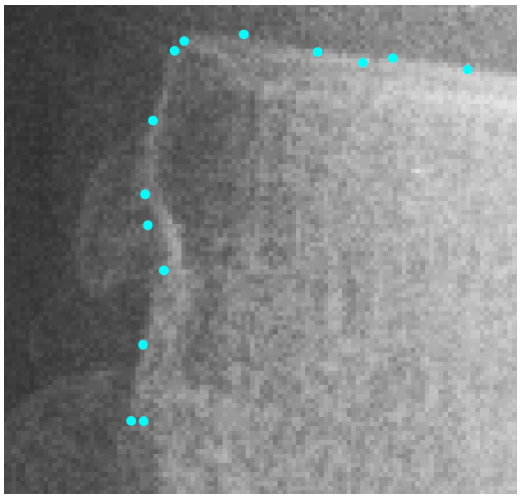

Figure 4.47: SSM-DP missing the osteophytes along the edge.

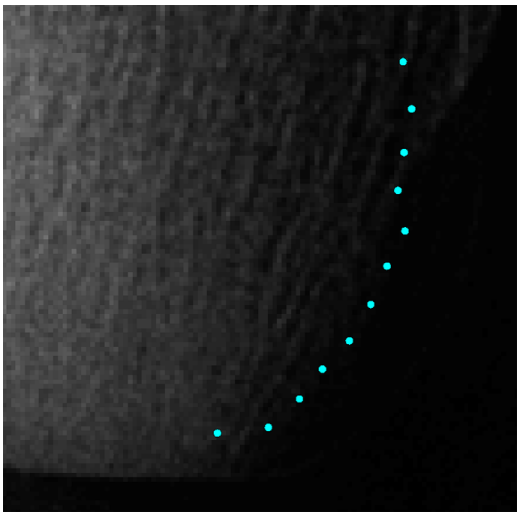

Figure 4.48: SSM-DP missing edge from image contrast.

4.3.5 Tibial Spines

For the tibial spines we compare the Haar, SSM-DP contours and SSM-RS contours extracted from the overall shape points. The Haar features achieved the highest accuracy with 0.824 AUC (see Fig. 4.55), the combination of the shape and texture features (see Fig. 4.49) did not improve this accuracy with an AUC of 0.785 (combined), to 0.824 (Haar), 0.647 (SSM-DP), 0.651 (SSM-RS). The increased accuracy of the Haar features is attributed to the overlap with the intercondylar notch, this can add features about JSN and osteophytes in the region. The Haar features achieve an overall multi-class accuracy of 34% ±0.9% (see Table 4.6).

Table 4.6: Spines AUC and multi-class performance

Analysis Method	mean AUC (stdev.)	multi-class accuracy (stdev.)
SSM-DP	0.647(0.003)	41.6 (0.4)
SSM-RS	0.651 (0.004)	27.2 (2.7)
Haar	**0.824 (0.006)****	**45.8 (0.4)****
SSM-RS + SSM-DP	0.68 (0.009)	29.1 (1.9)
SSM-DP + Haar	0.811 (0.003)	33.9 (0.5)
SSM-RS + Haar	0.807 (0.005)	33.9 (1.1)
All features	0.785 (0.003)	29.9 (0.9)

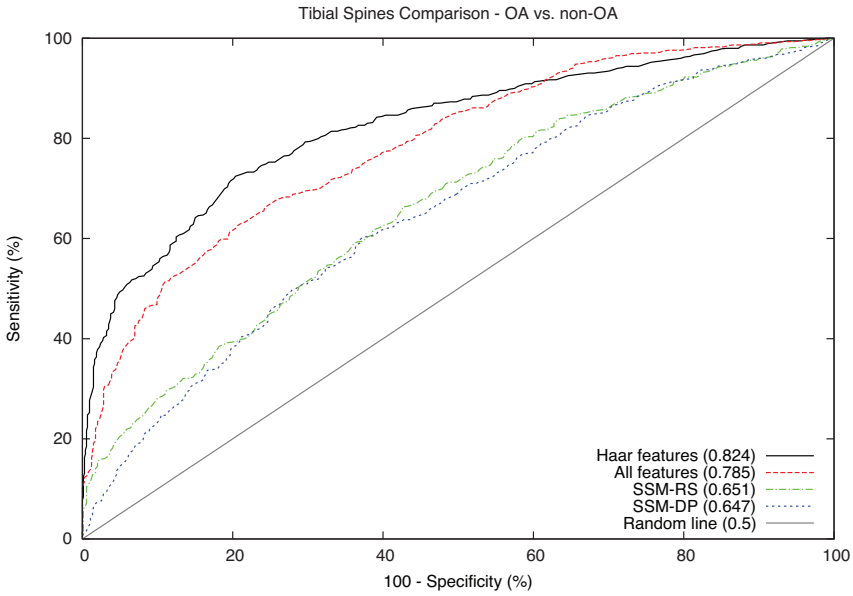

Figure 4.49: Comparison of tibial spine Haar, SSM-DP, and combined features .

The LDA shape modes show no discernable change in shape (see Fig. 4.50). The shape models in Figure 4.40 show a change in the overall shape model, this could mean that tibial spine change is only apparent in a few cases of OA and so not strongly associated with disease development.

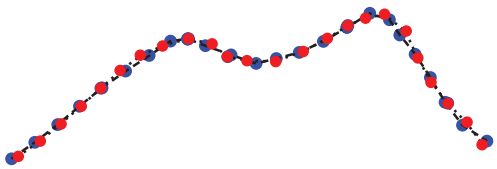

Figure 4.50: Tibial spines SSM-DP LDA model, the red circles represent the cases, blue points represent the control class mean shape.

4.3.6 Joint Space

The JS-SSM is compared to xJSW [7] measurements, available in the OAI the dataset, in the next chapter (Chapter 5). The experiments are separated due to the xJSW measurements available in the data.

This section analyses the detection of JSN using the 704 images with OARSI JSN grades (0-3), and the joint space shape change across the OA - splitting the experiments into OA vs. non-OA (see Figure 4.55) and KL grades (see Table 4.9).

JSN Detection

The JS-SSM was run over the medial and lateral compartments of the knee separately (the shape models were split into two halves) analysing the OARSI JSN grade in each compartment. The OARSI grades categorise the narrowing into grades (0-3) depending on severity, with 0 = no JSN and 3 = severe joint space loss. The AUC and multi-class JSN accuracy is shown in the Table 4.7 below. The experiments also compared the JS-SSM, JS Haar features and the combined joint space features (combined JS). The combined model achieved the best results for the medial JSN, with an AUC of 0.977. For the lateral side the JS Haar achieved the best results with an AUC of 0.947.

For overall multi-class repeatability when compared to the gold standard the method achieved a kw of 0.64 (0.6-0.69) medial JSN, and 0.17 (0.01-0.32) lateral JSN. The lateral accuracy is low because of the limited numbers in lateral JSN ≥ 1. The medial kw (0.64) achieves an accuracy that is within the reported manual JSN grading (0.48-0.86) [41] [39].

Table 4.7: OARSI JSN detection performance

	Medial		Lateral	
Analysis Method	AUC(stdev.)	multi-class(stdev.)	AUC(stdev.)	multi-class(stdev.)
JS-SSM	0.972(0.003)	61.4(1.5)	0.918(0.009)	94.7(0.6)
JS Haar	0.96(0.003)	64(0.1)	**0.947(0.012)**	95.1(0.4)
Combined JS	**0.977(0.001)****	**66.1(0.6)****	0.944(0.001)	**95.3(0.2)**

Osteoathritis Classification

These experiments evaluate the two-class and multi-class problems using JS-SSM and JS Haar features. The experiments are also expanded to show the effect of combining JS features (combined JS), and the combination with osteophyte Haar features and overall shape model (see Table 4.8). The JS Haar features achieved a higher accuracy, with 0.887 (Haar), 0.867 (JS-SSM). However, when both features were combined with osteophyte Haar features, the results were comparable 0.929 (JS-Haar + osteophytes), 0.930 (JS-SSM + osteophytes), indicating that the extra information in the joint space Haar features overlaps the osteophyte texture regions.

Table 4.8: Joint Space AUC and multi-class performance

Analysis Method	mean AUC (stdev.)	multi-class % (stdev.)
JS-SSM	0.867(0.004)	41.6 (0.4)
JS Haar	0.887 (0.001)	40.7 (0.5)
OS Haar	0.92 (< 0.001)	47.8 (0.3)
Shape	0.843 (< 0.001)	33.9 (1.8)
JS-SSM + shape	0.884 (0.002)	44.4 (0.7)
JS-SSM + Haar	0.901 (0.001)	46.8 (0.2)
JS Haar + OS Haar	0.929 (0.003)	47.7 (0.2)
JS-SSM + OS Haar	0.93 (0.002)	**49.4 (1.8)**
Combined JS + OS Haar	0.931 (0.001)	47.7 (0.1)
JS-SSM + OS Haar + shape	**0.933 (0.001)**	47.7 (0.7)

Adding the JS-SSM to the overall shape model increases the AUC (see Fig. 4.51) from: 0.843 and 0.867 to 0.884 (shape, JS-SSM, combined) and a multi-class of 41.6% ±0.4 JS-SSM, to 44.4% ±0.7 overall shape + JS-SSM.

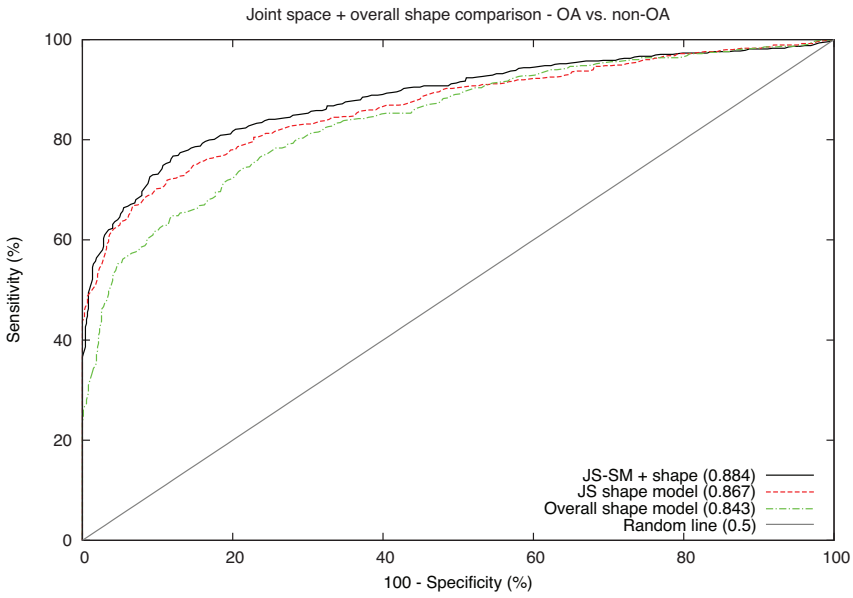

Figure 4.51: JS-SSM, overall shape and combined feature ROC curves.

The LDA shape modes from the two-class experiments (see Fig. 4.52) show that the main shape feature used to split the data is JSN in the knee, with some pockets and rotation of the lateral plateau occurring in the OA class shape.

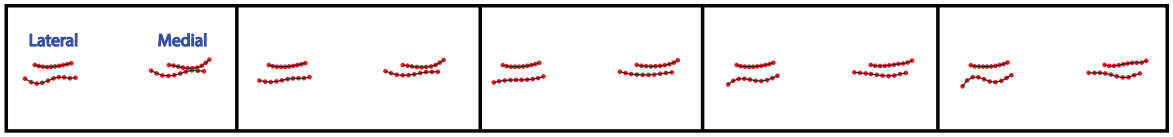

Figure 4.52: JS-SSM LDA model of the differences between the OA (left) and non-OA (right) classes.

The lower α from the DP optimisation had trouble finding the correct edges in some cases (see Figures 4.53-4.54 below).

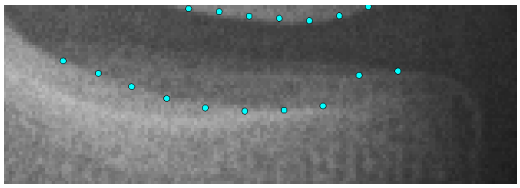

Figure 4.53: Contours switching between plateau lines.

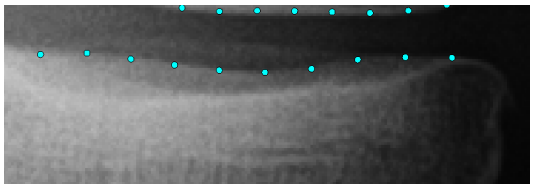

Figure 4.54: Contours following rotated edge of the plateau.

4.3.7 Combined Methods

The combined model takes the best features and method combinations from the individual feature evaluations. The features are combined to improve on the OA vs. non-OA and KL grade.

Osteoarthritis vs. Non-Osteoarthritis

Experiments found that all combined features and all features without the osteophyte SSM-RS achieved the best detection for the two-class experiments, with 0.939 (AUC) fully combined, and 0.939 (AUC) all without SSM-RS. The difference between the two is minimal (< 0.001). From the independent features, the osteophyte features achieved the best detection accuracy, with an AUC of 0.93 (see Fig. 4.55 below). The fully combined features were compared with the WND-CHARM [9] algorithm built on the same data (see Fig. 4.56), we found our algorithm achieved a higher AUC of 0.94 (combined model) to 0.82 (WND-CHARM).

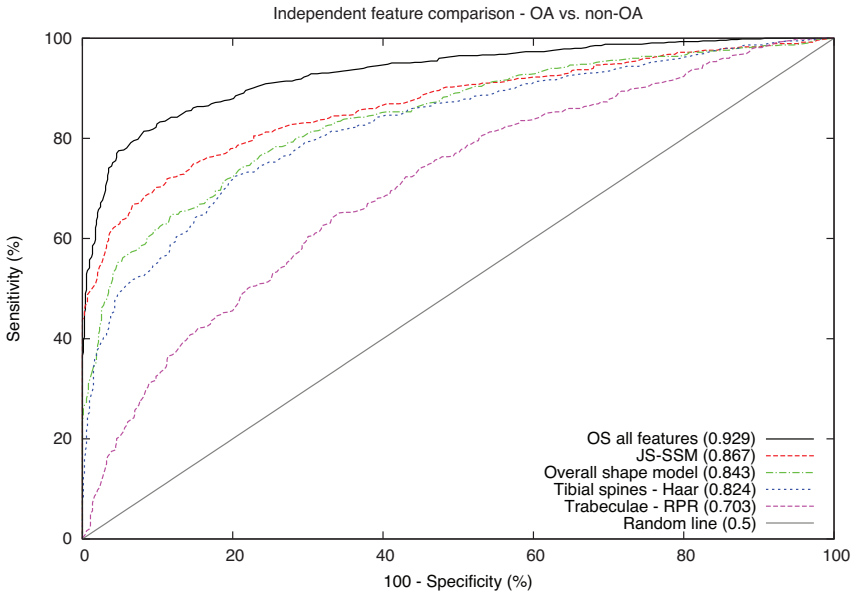

Figure 4.55: ROC curves of all features: osteophytes (SSM-RS + SSM-DP + Haar), Joint space shape models (JS-SSM), overall shape SSM , tibial spines (Haar features), trabeculae (RPR) in detecting OA vs. non-OA images. Osteophyte features have been abbreviated to OS.

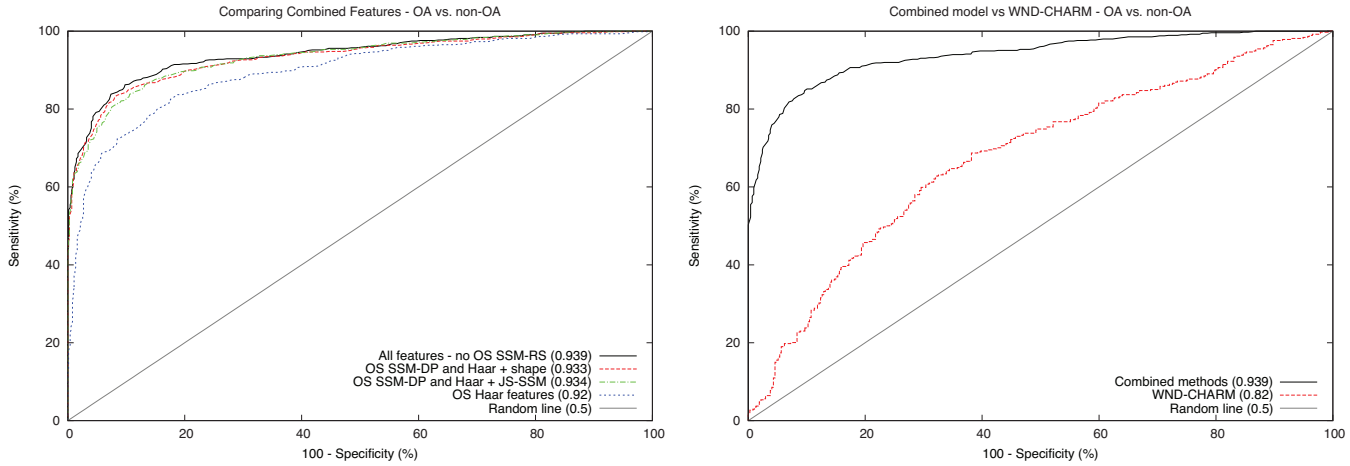

Figure 4.56: ROC curves comparing: combined features (left), combined feature model with the WND-CHARM model (right).

All methods are combined through taking the mean (unweighted) outputs of all random forests trained on the separate extracted features. Preliminary experiments varied weights on the classifier outputs before taking the mean, but this weakened the detection on larger datasets. Experiments were run on training the RFs on concatenated feature vectors, this achieved similar accuracy to the mean outputs (0.937

combined outputs vs. 0.937 concatenated features), but slowed the RF training from splitting larger feature sizes.

Multi-class

The best combination of features for KL classification was found to be: overall shape (shape), osteophytes (OS), and tibial spines (spines) (see Table 4.9). The trabeculae and joint space (JS) features were removed because they weaken the accuracy from 51.8% (optimal method) to 42.7% (fully combined), and 48.1% (optimal + JS-SSM). This is likely because of noise introduced from the features, the trabeculae provide no clear distinction between the specific KL grades (25.7%), and JS-SSM features have a weak KL2 accuracy (13.7%). All two-class and multi-class experiments are shown in the table below (see Table 4.9). The osteophyte SSM-DP + Haar features is abbreviated to OS×2.

Table 4.9: Accuracy of the KL grading, and AUC of the OA vs non-OA experiments

	Accuracy (%)					kw (CI 95%)	AUC (stdev.)
Analysis Method	KL 0	KL 1	KL 2	KL 3	KL 4	Overall (stdev.)	multi-class	Two-class
Overall Shape	36.2	37.7	19.0	37.5	37.5	33.9 (1.8)	0.29(0.24-0.34)	0.843 (< 0.001)
Trabeculae (RPR)	34	34.8	19.3	19.5	5.3	25.7 (0.2)	0.05(0.02-0.1)	0.703 (0.004)
OS Haar	43.5	49.8	21.9	64	63	47.8 (0.3)	0.51(0.47-0.56)	0.92 (< 0.001)
OS SSM-RS	44.1	47.5	18.9	46	68.8	43 (0.8)	0.44(0.39-0.49)	0.851 (0.001)
OS SSM-DP	46.5	41.9	21.3	43.2	55.5	40.7 (0.5)	0.42(0.37-0.48)	0.865 (0.003)
Spines Haar	35.9	35.5	20.5	38.1	41.4	34 (0.9)	0.32(0.28-0.37)	0.824 (0.006)
JS-SSM	46.3	35.3	13.7	55.7	57.1	41.6 (0.4)	0.41(0.36-0.45)	0.867 (0.004)
Combined features								
OS all features	45.9	**53.5**	20.1	68	65.6	50.2 (0.5)	0.56(0.52-0.6)	0.929 (0.002)
OS×2 + shape	42.4	51.2	**22.4**	69.9	64.8	49.6 (1.7)	0.56(0.52-0.6)	0.933 (0.001)
OS×2 + JS-SSM	48.6	45.5	9.4	65.2	**70.6**	47.6 (0.1)	0.47(0.42-0.51)	0.934 (0.007)
OS×2 + spines + shape	45	53.2	**22.4**	**73.3**	68	**51.8 (0.5)**	**0.58(0.54-0.62)**	0.934 (0.005)
OS×2 + spines + JS-SSM	50.2	43.5	6.6	66.5	69.8	47.3 (1.4)	0.49(0.44-0.53)	0.937 (0.001)
Combined (no trabeculae and OS SSM-RS)	**51.1**	45	7.5	67.1	67.5	48.1 (0.5)	0.48(0.43-0.52)	0.937 (0.001)
Combined (no OS SSM-RS)	50.2	45	6.6	72.2	69	48.9 (0)	0.49(0.45-0.54)	**0.939 (0.001)****
Fully combined	32.6	47.5	17.5	58.6	54.5	42.7 (1.8)	0.37(0.32-0.42)	**0.939 (0.003)**

4.4 Discussion

This chapter has evaluated various shape and texture methods orientated around specific radiographic OA features. The experiments were run on a small subset of 747 OAI images to evaluate OA classification (both two-class and multi-class), with an even distribution of OA to non-OA knees. Further experiments were run to classify OARSI grades of osteophytes and JSN using the joint space and osteophyte features. The OARSI grades were available on a subset of the 747 images, which left some class unbalance in the experiments. This is recognised as a limitation when evaluating the methods for classifying OARSI grades, but was felt to be sufficient for comparing the feature specific methods. The experiments in this chapter have shown that combining various shape and texture features from plain radiographs of the knee create a better OA and KL classification. Osteophytes provide the best independent detection of OA. The best features for the two-class (OA vs. non-OA) experiments were a combination of all features (overall shape, osteophytes, joint space, trabeculae and tibial spines). The multi-class KL experiments found that the optimal model contained all but the trabecular and joint space features. The reliability of the optimal combined multi-class model has a weighted kappa (kw) 0.57, which is within the range achieved by human grading (0.36 - 0.8) [35]. When compared to the inter-observer reliability of the OAI dataset, our algorithm is worse, with 0.57 (automated method) to 0.70 (OAI subset). The optimal features for all methods in the final model can be seen in the Table 4.11 below.

From the literature, it might be expected that the osteophyte and joint space models would be the optimal multi-class model, as KL grade is mainly based on osteophyte and JSN development [2], however the optimal model was in fact the combined osteophyte, spines and overall shape features. The decrease in accuracy mainly comes from KL2, which is 9.4% with the JS-SSM model and 22.4% with the shape model. This could indicate noisy data from the JS-SSM, or that the shape captures more information from the alignment and JSN relating to KL grade than the specific JS-SSM measurement.

The multi-class experiments have shown that all methods are relatively poor at detecting KL2 grades. This may be due to the similarities of the central KL grades (1-3), with the distinctions between "doubtful" (KL1), "definite" (KL2) and "multiple definite" (KL3) being less comprehensible by the automated method. The manual grading is reliant on osteophytes and JSN being visible to the observer, meaning the automated method could have detected smaller "definite" features, where a manual grader might classify them as "doubtful". This is supported by the majority of the KL2 images (83.8%) classified in the range KL1-3 (KL1 42%, KL2 22.9%, and KL3 18.9%). Further to this, the automated method classifies most KL2 images as KL1 (42%) meaning that the distinction between "doubtful" and "definite" features is not as clear for the extracted features.

The automated method was compared to the WND-CHARM algorithm by Shamir et al. [9]. We trained the algorithm on the same 747 images and found that our fully combined model achieved a higher OA detection AUC. The implicit feature method by Anifah et al. [73] (tested on a different dataset) the fully combined features achieved a higher AUC compared to the mean AUC of 0.592 over all KL grades. No two-class AUC was given in the literature.

Experiments show that the osteophytes achieved the best AUC of the independent features and of the combined osteophyte features, Haar features added the most information. This is likely because the SSM-DP will find the edges with the highest gradient differences, this can often ignore less well-defined edges in the radiographs caused through under-exposed regions. SSM-RS can be biased towards the mean shape within images missing the osteophytes, which can exhibit a wide range of shapes.

Our combined osteophyte features achieve a good OARSI osteophyte detection, and a mean OARSI multi-class probability of 55.8% across all four regions. The comparison to the gold standard OARSI grades achieved a low mean kw of 0.26 (0.12-0.41), where 0 = agreement equivalent to chance and 1.0 = perfect agreement. This is lower than the manual repeatability (kw = 0.64-0.78), but could be improved by expanding the datasets to include an even distribution of OARSI osteophyte grades in all marginal

regions.

The osteophyte features achieved a better result than the KOACAD [68] algorithm (on different datasets) in detection on the medial tibial osteophytes only (0.895 osteophyte Haar, 0.65 KOACAD). This may be because of important information in the texture surrounding the osteophytes. However, there may also be some bias towards our algorithm that will inherently measure some joint space narrowing information in the extreme cases of OA.

The joint space experiments show that combining Haar features over the same area adds accuracy to the JSN shape, but increases accuracy from overlap with osteophytes. This can be seen in the AUCs from the two-class experiments: 0.93 Haar joint space + Haar osteophytes, 0.93 JS-SSM + Haar osteophytes, 0.93 JS-SSM + Haar JS and Haar osteophytes.

The JS-SSM achieved a high JSN detection, with an AUC of 0.92 (lateral) and 0.97 (medial). The comparison to the gold standard OARSI JSN grades from the medial joint space is comparable to manual inter-observer OARSI grading kw: 0.6 JS-SSM, (0.48-0.86) manual JSN. The lateral comparison to the gold standard is lower than this (kw: 0.31), which is likely due to the few lateral JSN grades ≥ 1 to train on, JSN $\geq 1 = 33$. This was due to the focus on medial OA participants from the restrictions of the trabecular features (trabeculae on the lateral side of the knee are obstructed by the fibula). More lateral OA samples trained in the classifier should produce similar accuracies to the medial side.

The JS-SSM features achieved a higher AUC than the best reported method in the literature [68]. The KOACAD algorithm was run on separate data (5950 images) and achieved AUCs of: medial mJSW - 0.728 $\pm$0.003, lateral mJSW - 0.544 $\pm$0.032, medial JSA - 0.685 $\pm$0.043, lateral JSA - 0.53 $\pm$0.03. No combined joint space AUC was given, the mean scores were taken across the separate male and female participant results given in the paper.

Trabecular texture is best captured through RPR rather than gradient and FS analysis; however, this only adds useful information in OA vs. non-OA experiments, with weak accuracy for KL classification. This could be from the KL grade being based on shape features and trabeculae structure only varying in the extreme KL grades, i.e. KL0 and KL3. The two class experiments detect this change, but the discrepancy of trabecular change in the between grades (KL1, 2 and 4) appear to weaken the accuracy of the model overall.

The experiments on the tibial spines, show that the features do have some association with OA, but the Haar features achieve a better accuracy. This increase could be attributed to the detection of faint edges, missed by the DP optimal edges, but is more likely to be from the added osteophyte and JSN information from the intercondylar notch. The two-class finding a fairly moderate AUC, although this is the weakest of all the independent features. The multi-class results produce a similar correlation to the literature with [71] showing a weak correlation (close to 0) of -0.15 and 0.14, and a kw of 0.08 for the tibial spines shape.

In the following chapters the experiments on the JS-SSM are expanded to analyse further features of OA: detection of pain and the prediction of later onset OA (KL ≥ 2) and later onset pain in follow-up visits from the OAI dataset. Also, the JS-SSM features are compared in all experiments, including the detection of OA, to classifiers trained on the multiple JSW measurements (xJSW) found through the method by Duryea et al. [63] available in the OAI data.
The thesis follows this with an analysis of the fully combined feature model on the pain detection and prediction of later onset pain and OA experiments.

4.4.1 Important Findings

The optimal model for detecting OA contains features across all areas of the knee:

- **Overall shape** Using a SSM built on RFCLM output, these features capture JSN and mal-alignment of the bones.
- **Joint margins** Focuses on features of marginal osteophytes and some JSN. The

optimal model uses Haar features and a SSM built on DP optimised contours.

- **Joint space** The features analyse: JSN, some attrition and mal-alignment. The optimal model uses SSM features from DP contours.
- **Tibial spines and intercondylar notch** Change to the tibial spines, osteophytes and some JSN is analysed using Haar texture features.
- **Subchondral tibia** Trabeculae structure is gathered beneath the tibial plateau using RPR features.

Table 4.10: Optimal Features for the Fully Combined Model

	Parameters									
Features	n pts	n samples	modes (var.)	RF trees	α	ROI	img. patch size	disp.	fw	Section
RFCLM coarse	74	250 imgs.	20 (95%)	10	-	-	20 x 20 px.	20	50	4.2.1
RFCLM medium	74	250 imgs.	30 (99%)	10	-	-	20 x 20 px.	15	100	4.2.1
RFCLM fine	74	250 imgs.	30 (99%)	10	-	-	20 x 20 px.	15	200	4.2.1
Overall shape	74	-	44 (99%)	50	-	-	-	-	-	4.3.2
Trabeculae	2	640 samples	-	10	-	256 x 125 px.	32 x 32 px.	-	-	4.2.3
OS SSM-DP	48	-	30 (85%)	15	0.2	-	-	-	-	4.2.4
OS Haar	8	-	-	15	-	25 x 25 px.	-	-	-	4.2.4
JS-SSM	40	-	18 (99%)	20	0.2	-	-	-	-	4.2.5
Tibial spines	2	-	-	10	-	19 x 19 px.	-	-	-	4.2.7

Table 4.11: img. = image, var. = variance (%), ROI = Region Of Interest, disp. = max displacement of image patches around the points in training (RFCLM), fw = frame width (resolution of the RFCLM stage), n pts = points, px. = pixels, modes = shape modes extracted from the PCA of the annotated points, n samples = number of images or patches of image used to train the model (this is left blank for models that used the cross validation across all images to train and test), r = reference length between the selected RFCLM points (typically the width of the tibia), OS = Osteophytes, JS-SSM = Joint Space Statistical Shape Model

Chapter 5

Joint Space Method Comparison

This chapter summarises the analysis comparing the Joint Space Statistical Shape Model (JS-SSM) (described in Chapter 4) with the widely used xJSW approach by Duryea et al. [63]. Following the literature, experiments have been expanded to evaluate current knee pain [6] [29] [32] [11], later onset knee pain [32] [11], and later onset OA [79] [7].

5.1 Data

The images and outcomes are taken from the Osteoarthritis Initiative (OAI) dataset, explained in Section 3.1. The data is split into the classification of current disease outcomes and the prediction of later onset disease outcomes. All images are restricted to knees with pain, xJSW and Kellgren and Lawrence (KL) grades recorded at the specific visit.

The current disease features include: Osteoarthritis (OA) detection, taken from the previous experiments in Section 4.1; and pain detection, using all images from the baseline visit with pain, KL grade and xJSW. The later onset outcomes include: OA prediction, to detect images with no OA (KL ≤ 1) that develop later onset OA (KL ≥ 2); and pain prediction, detecting participants with no pain at baseline developing pain during follow-up visits. The outcome of pain is based on the calculated binary variable, reported as any occurrence of pain, aching or stiffness experienced in the last 30 days in the respective knee. This measurement is calculated from a clinical questionnaire taken at each visit (baseline and follow-up) for the participant. The pain

assessment [106] measures the knee pain using various scoring mechanisms i.e. Western Ontario and McMaster Universities Arthritis Index (WOMAC) and Knee injury and Osteoarthritis Outcome Score (KOOS) [107]. The binary variable is then taken as any occurence of pain, aching or stiffness reported using the pain variables. The statistics for each dataset are:

Current features

- ***Current OA*** 547 images (from the 747 images in Section 4.1), with KL grades: KL0 - 83, KL1 - 141, KL2 - 103, KL3 - 157, KL4 - 63. The two-class OA vs. non-OA is then: OA - 323, non-OA - 224.
- ***Current pain*** 5932 images taken from baseline visit. The images are split: pain - 2084, no pain - 3846. The KL grades are: KL0 - 1177, KL1 - 877, KL2 - 2355, KL3 - 1230, KL4 - 293.

Later onset features

- ***Later onset OA*** contains 2060 images with KL grades ≤ 1: KL0 - 1178, KL1 - 882. The case (participants who develop OA) and controls (participants who do not develop OA) over follow up are split: case - 463, controls - 1597.
- ***Later onset pain*** contains 3840 images with KL grades: KL0 - 918, KL1 - 587, KL2 - 1560, KL3 - 675, KL4 - 100. The images are split into: 1541 participants who develop pain and 2299 participants who do not develop pain.

5.2 Methods

This section describes the methods used to extract xJSW and JS-SSM features for the experiments. The Random Forest Constrained Local Model (RFCLM) for the JS-SSM method is the same as in the previous experiments (see Section 4.2.1).

Duryea's Joint Space Width Measurements (xJSW)

The xJSW measurements, explained in Section 2.4.1, are a series of joint space widths at fixed distances along the joint space. The measurements are taken from the OAI

data and were collected using a semi-automated method by Duryea et al. [63]. The distances are measured relative to the whole knee by projecting a line (x) across the femoral condyles (see Fig. 5.1). The measurements are split between the medial and lateral compartments of the joint space, missing the tibial spines and intercondylar notch. The medial joint space has 7 widths between $x = 0.15$ and $x = 0.3$ with increments of 0.025. The lateral joint space has 9 widths at $x = 0.7$ to $x = 0.9$ in 0.025 increments. The distance at the minimal Joint Space Width (mJSW) of the medial compartment is also measured.

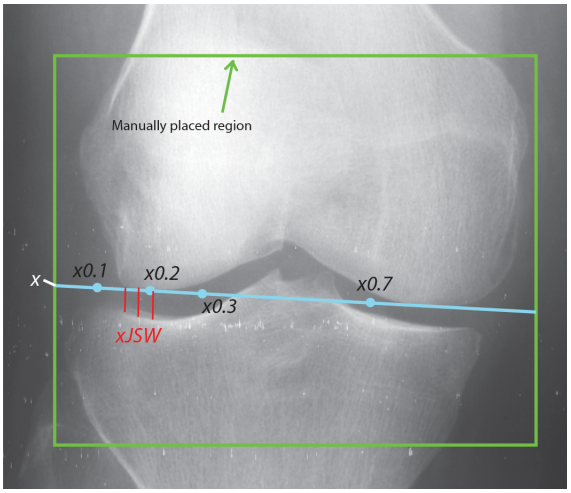

Figure 5.1: xJSW measurements at points along projected line x within the manual placed region (green box).

The xJSW distances are recorded at points where a line can be drawn from the plateau to the condyle above, the JSW values are left blank in cases where the joint is mal-aligned or the plateaus are shorter than the condyles. To fill the missing values we used "-1" values to null the features. This was found to be the optimal value during preliminary experiments. The Random Forests are trained and tested using 17 variables, JSW and mJSW concatenated into a single vector per image.

Joint Space Shape Model

The method from Section 4.2.5 is used to place 40 points across the medial and lateral joint space compartments. The shape variation for the SSM was fixed at the optimal value from the preliminary experiments (99%), this equated to 18 modes for all datasets except later onset OA which used 20 modes. Prominent shape modes from

the preliminary experiments are seen in Figures 5.2-5.4 below.

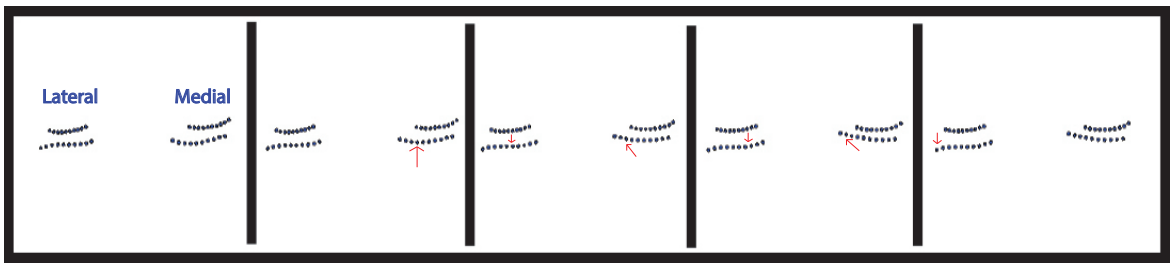

Figure 5.2: Shape mode showing JSN and medial plateau shape change.

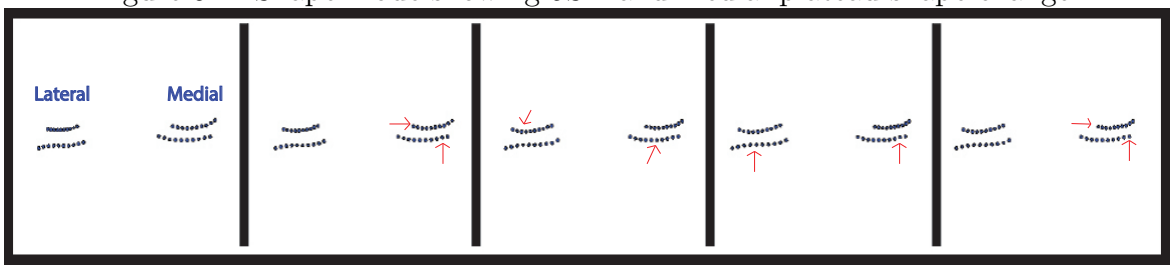

Figure 5.3: Shape mode showing slight shift of the femur and tibia.

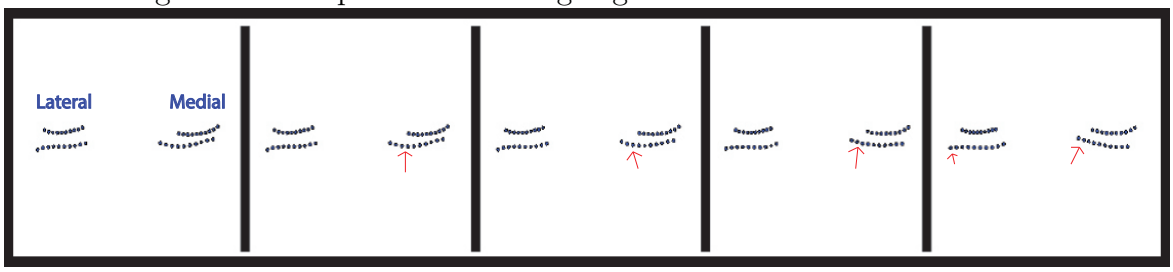

Figure 5.4: Shape mode showing some medial plateau shape change.

5.3 Experiments

The experiments compare the utility of JS-SSM, xJSW, and the combined joint space features (combined JS). The features are compared using the accuracy of Random Forest (RF) classifiers (described in Section 3.5.1) trained and tested using 5-fold cross validation with one repeat (see Section 3.5.2). The data was analysed using the methods in Section 3.5.3. The results reported are adjusted for potential bias from the correlation between the participants knees, as the difference between Area Under ROC Curves (AUCs) in treating the knees as independent and correlated was < 0.01. The features were compared using the four datasets: 1) Current OA (two-class and multi-class) 2) Current pain 3) Later onset OA 4) Later onset pain. Results are

reported as: mean AUC for all two-class experiments, and mean overall and per-class accuracy (OA multi-class). Linear Discriminant Analysis (LDA) shape modes have been generated from the JS-SSM features for each experiment. The symbol ** will be used in all the tables to state that the best accuracy is significant compared to the other results reported in the same table.

5.3.1 Current Disease Outcomes

The following experiments detects images depending on the current disease outcome (OA or pain). The images are taken across any time point (baseline or follow-up visits).

Current Osteoarthritis

The xJSW and JS-SSM were compared using the 547 OA detection images. The experiments split the data either into OA vs. non-OA (two-class) or KL grade (multi-class). The results show that the JS-SSM achieves the best two-class accuracy (see Figure 5.5) with an AUC of 0.849 (JS-SSM) to 0.798 (xJSW). The combination of the two features increases the two-class AUC to 0.859, see Table 5.1. The multi-class experiments (see Table 5.2) show the two features are comparable with accuracy rates of 45.6% (xJSW) to 44.5% (JS-SSM). The combined JS features show a significant improvement in the multi-class experiments with an overall accuracy of: 44.5% (JS-SSM) and 45.6% (xJSW), to 51% (combined JS).

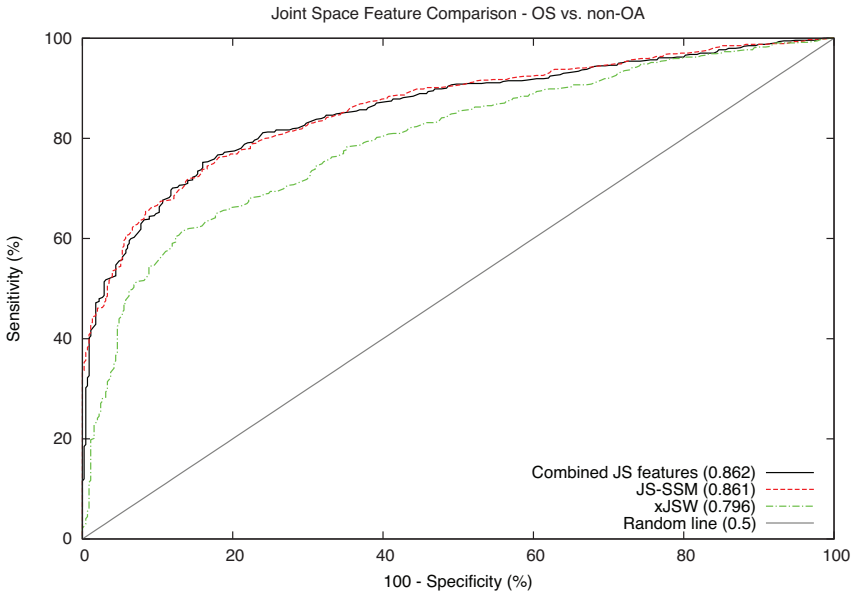

Figure 5.5: Comparing ROCs for the JS features: xJSW, JS-SSM and combined JS.

Table 5.1: Joint Space Two-class OA Detection Results

Analysis Method	AUC (stdev.)
xJSW	0.798 (0.76 - 0.83)
JS-SSM	0.849 (0.82 - 0.88)
Combined JS	**0.859 (0.83 - 0.89)**

Table 5.2: Joint Space Multi-class OA Classification Results

	Accuracy (%)						kw (CI 95%)
Analysis Method	KL 0	KL 1	KL 2	KL 3	KL 4	Overall (stdev.)	
xJSW	**33.0**	41.5	15.5	69.4	70.6	45.6 (1.6)	0.47(0.42-0.53)
JS-SSM	28.3	43.0	25.6	64.7	53.9	44.5 (0.3)	0.49(0.44-0.54)
Combined JS	28.1	**43.4**	**28.1**	**78.2**	**70.8**	**51 (1.5)**	**0.58(0.53-0.62)**

Combining the two JS features significantly increases the accuracy of KL grades 2 and 3. The increase could come from the specific measurements of xJSW adding to the shape and alignment information captured by JS-SSM features. The LDA shape modes show the same variation as the previous OA detection experiments (see Figure 4.51).

Current Pain

The current pain experiments compare JS-SSM and xJSW across 5932 baseline images (see Fig. 5.6) when used with classifiers to predict whether the participant reports pain in the respective knee. The JS-SSM achieves a higher AUC with 0.609 (JS-SSM) to 0.594 (xJSW). The combined JS features achieve the best AUC with 0.621 (see Table 5.3).

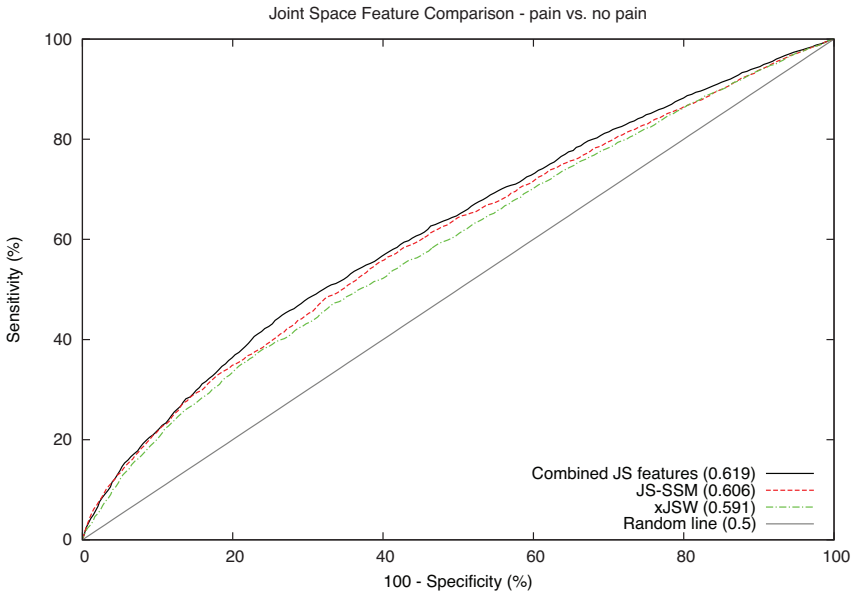

Figure 5.6: Comparing ROCs for the JS features: xJSW, JS-SSM and combined JS.

Table 5.3: Joint Space Pain Detection Results

Analysis Method	AUC (stdev.)
xJSW	0.594 (0.58 - 0.61)
JS-SSM	0.609 (0.59 - 0.62)
Combined JS	**0.621 (0.62 - 0.65)**

The LDA shape modes in Figure 5.7 illustrate some medial JSN and lateral plateau rotation.

Figure 5.7: LDA shape model of the JS-SSM between the painful and non-painful classes (red and blue points).

5.3.2 Future Disease Outcomes

All experiments in this section compare joint space features in predicting later onset outcomes. The images have no reported outcomes and are taken from the baseline visit. The radiographic features are used to detect participants who develop a disease outcome at any point over the four follow-up visits (cases), and those who do not develop the outcome at any point over follow-up (controls).

Later Onset Osteoarthritis

The JS-SSM and xJSW features are compared in predicting later onset OA (KL ≥ 2) from baseline. The JS-SSM achieves a higher AUC with 0.54 (JS-SSM) to 0.526 (xJSW). The combined features do not significantly increase the AUC with 0.543 (CI: 0.51 - 0.57) (combined JS), to 0.54 (CI: 0.51 - 0.57) (JS-SSM). All results are shown in the Figure 5.8 and Table 5.4 below.

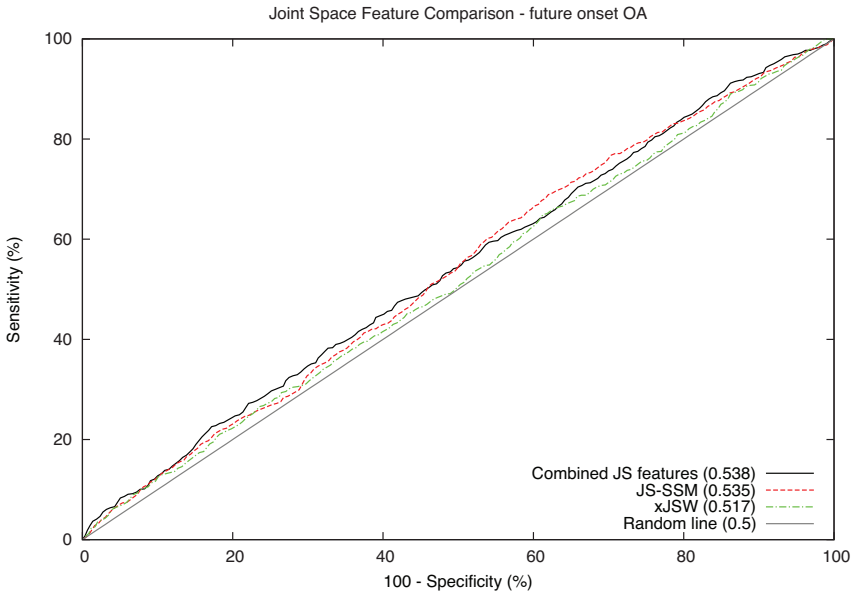

Figure 5.8: Comparing ROCs for the JS features: xJSW, JS-SSM and combined JS.

Table 5.4: Joint Space OA Later Onset Prediction Results

Analysis Method	AUC (stdev.)
xJSW	0.526 (0.5 - 0.56)
JS-SSM	0.54 (0.51 - 0.57)
Combined JS	**0.543 (0.51 - 0.57)**

The LDA shape modes for the JS-SSM in Figure 5.9 show the slight medial plateau shape variation between the knees which do not develop OA (control) and knees which develop OA (case).

Figure 5.9: LDA shape model of the JS-SSM between the case and control classes (red and blue points).

Later Onset Pain

The joint space features are compared in predicting later onset pain (see Figure 5.10 and Table 5.5) The results show that the JS-SSM and xJSW achieve comparable AUC with xJSW - 0.565 (CI: 0.55 - 0.58) to JS-SSM - 0.572 (CI: 0.55 - 0.59). The combined features improve the AUC to 0.583.

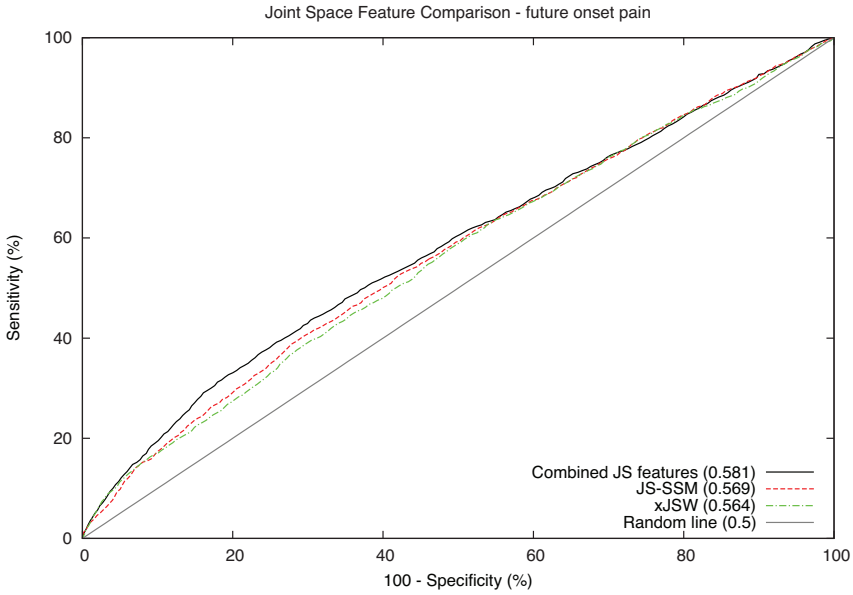

Figure 5.10: Comparing ROCs for the JS features: xJSW, JS-SSM and combined JS.

Table 5.5: Joint Space Pain Later Onset Prediction Results

Analysis Method	AUC (stdev.)
xJSW	0.565 (0.55 - 0.58)
JS-SSM	0.572 (0.55 - 0.59)
Combined JS	**0.583 (0.56 - 0.6)**

The LDA shape modes in Figure 5.11 show minimal shape variation between the knees which do not develop pain (control) and knees which develop pain (case).

Figure 5.11: LDA shape model of the JS-SSM between the case and control classes (red and blue points).

5.4 Discussion

The experiments have shown that the JS-SSM features achieve the best independent accuracy. This is likely to be because the xJSW features are limited to specific measurements at set locations across the joint space, whilst the JS-SSM adds shape and alignment information about the tibia and femur. In the experiments combining both JS-SSM and xJSW achieves the best accuracy, with an increase in AUC between 0.001 - 0.012. The improvement is likely from the xJSW specifically measuring JSW from the front lines of the joint space, whereas the JS-SSM contours have been shown to fit to lines of the rotated tibial plateaus (see Figures 4.53-4.54).

The current pain experiments show a weak association with radiographic features (AUC: 0.621). The results are consistent with findings that joint space measurements are associated with clinical symptoms of pain and stiffness [6] [22]. Further to this, the detection of current OA is stronger than current pain with the AUCs: 0.862 (OA), to 0.621 (pain). Similar findings have been reported in automated radiographic assessments [6][11]. This is because KL grade is based on radiographic features, whilst pain is a clinical measure and is based on subjective participant experience. Weak associations between radiographic features and clinical evaluated OA have been well documented in the past [108] [109]. An improvement to this assessment would be including more radiographic features (osteophytes and overall shape), which have been shown to correlate with current knee pain [20] [30].

The experiments show that the current disease outcomes consistently achieve a higher accuracy than the later onset prediction with: OA - 0.862 (current) to 0.543 (later

onset), and pain - 0.621 (current) to 0.583 (later onset). These findings are consistent with the literature [32] [11], which show that features to predict later onset disease outcomes are much weaker than the current features. The KIDA algorithm [11] assessed JSW and mJSW features from 1002 participants for association with pain and function. The algorithm runs over multiple time points: T0 (current), T2y (2 years prior to incidence) and T5y (5 years prior to incidence). The pain results are reported as Odds Ratios (OR), where 1.0 is not significant. The results show a weak OR with time points over T0, with: T0 - medial JSW - 0.7 (0.49 - 1.0), lateral JSW 1.38 (0.99 - 1.94), T2y - lateral JSW 0.87 (0.76 - 0.99), T5y - mJSW 0.79 (0.7 - 0.9). Galván-Tejada et al. [32] uses xJSW features from the OAI data to evaluate current and later onset pain. The experiments are split into T0, T1y (1 year prior to incidence pain), and T2y. The experiments achieve AUCs of: T0 - 0.695, T1y - 0.623, and T2y - 0.62. The results are higher than the xJSW results achieved in the current experiments (T0 - 0.695 ([32]), 0.619 (combined JS)), this could be caused by better defined joint space measurements in the smaller subset (163 [32] to 2966 combined JS).

The joint space is weakly associated with later onset OA (random AUC = 0.5), with AUCs: 0.526 - 0.543. These results are better than findings from the KIDA algorithm [79], which finds no association when predicting incidence OA (KL $\geq$ 2) over 5 years follow-up from baseline images. The results show a weak OR: medial JSW - 0.67 (0.57 - 0.83), lateral JSW - 1.09 (0.97 - 1.22), and mJSW - 0.75 (0.66 - 0.86). This could be due to limitations in the features used, with the medial and lateral JSW measuring the smallest distance per compartment. The xJSW and JS-SSM both add more information about the shape and widths across the joint space. This is supported by findings in [7], that showed using the multiple xJSW rather than singular mJSW measurements gave better accuracy in assessing longitudinal KL progression.

The results suggest that the fully automated features (JS-SSM) can replace the semi-automated (and thus time consuming) xJSW features. The remainder of experiments of the project include joint space features from JS-SSM only. The benefit of 0.012 AUC increase from the combined JS features is limited by the restriction of images, which have xJSW measurements available. The xJSW measurements rely on a region to be

placed manually on the knee (see Figure 5.1) to project the line x across the femoral condyles. The restriction of sampling images with xJSW measurements reduced the baseline images from 8847 with pain scores to 5932 (current pain detection set). The JS-SSM is unrestricted and can be run fully automatically across all images without the need for operator input.

The next chapter evaluates the four disease outcomes covered in these experiments (current and later onset OA and pain) using the optimal fully combined features (from Section 4.4.1).

Chapter 6

Osteoarthritis and Pain Experiments

This chapter evaluates the radiographic features from Chapter 4 evaluating current and later onset disease features with all the extracted radiographic features: overall shape, trabecular structure, osteophytes, joint space shape, and tibial spines. The experiments are split into four sections:

1. Current Osteoarthritis (OA) - To split knees into two-class (OA vs. non-OA) and multi-class (KL grades).

2. Current pain - Classifying features to detect pain in the respective knee.

3. Later onset Osteoarthritis - Analysing features from images with no OA (KL ≤ 1) to predict later onset disease (KL ≥ 2)

4. Later onset pain - Analysing features from knees with no pain to predict pain reported in follow-up visits.

The combined feature models are compared to the implicit feature method, WND-CHARM [9], and manual Kellgren and Lawrence (KL) grades taken from the Osteoarthritis Initiative (OAI) data.

6.1 Data

Baseline images from OAI (see Section 3.1) are used to evaluate the extracted radiographic features from Section 4.4.1. The experiments are split into current and later onset disease outcomes. The current OA is revisited to expand on the experiments from Chapter 4. The baseline set includes 9014 radiographs. Knees with artificial joints and implants were removed, leaving 8880 images. Metal implants and screws can cause problems for Raw Pixel Ratio (RPR) trabeculae and the overall shape features. The sets were then split into 8875 images with KL grades, and 8847 images with pain assessments and KL grades. The later onset data contains images with no current outcome (OA or pain) to predict features related to the occurrence (case) or no change (control) over the participant follow-up visits. Pain outcomes are assessed using the binary variable reporting any pain, aching or stiffness over the last 30 days in the respective knee. The stats for each dataset are:

Current features The data is used to evaluate the current outcomes of OA or pain across the baseline images.

- ***Current OA*** 8875 images (4445 participants) with KL grades: KL0 - 3454, KL1 - 1569, KL2 - 2344, KL3 - 1215, KL4 - 293. The two-class (OA vs. non-OA) is: OA - 3852, non-OA - 5023.

- ***Current pain*** 8847 images (4440 participants) taken from baseline visit. The images are split: pain - 2641, no pain - 6206. The KL grades are: KL0 - 3438, KL1 - 1558, KL2 - 2344, KL3 - 1215, KL4 - 292.

Later onset features The data evaluates features to predict later onset outcomes of OA or pain. All baseline images included have no occurrence of the outcome. The features predict knees that later develop OA or pain (cases), and knees that do not change outcome throughout the 6 years follow-up (controls).

- ***Later onset OA*** contains 2060 images (3110 participants) with KL grades ≤ 1: KL0 - 1178, KL1 - 882. The images are split into: 463 (case) and 1597 (control).

- ***Later onset pain*** contains 6206 images (3618 participants) with KL grades:

KL0 - 2724, KL1 - 1151, KL2 - 1559, KL3 - 669, KL4 - 103. The images are split into: 2215 (case) and 3991 (controls).

6.2 Methods

The methods for this chapter are taken from Section 4.4.1. The radiographic features extracted include: overall shape 4.2.2, RPR trabeculae 4.2.3, osteophytes (Statistical Shape Models with Dynamic Programming Search (SSM-DP) 4.2.4 and Haar features 4.2.4), joint space shape models (JS-SSM 4.2.5), and tibial spines 4.2.7. This section covers the parameters for each feature extraction method and Linear Discriminant Analysis (LDA) shape models taken from the new datasets.

6.2.1 Overall Shape

The overall shape extracts features of alignment and Joint Space Narrowing (JSN) from the data. The shape modes are set at 99% variation, and extracted 44-45 modes for all datasets. The LDA shape modes (Figures 6.1-6.3) show the differences between the case/control classes mean shape. The later onset OA and pain LDA shape models (Figures 6.2 and 6.3) have similar tibial spine spiking between the classes. The current pain shape (Figure 6.1) shows tibial spine spiking and medial JSN.

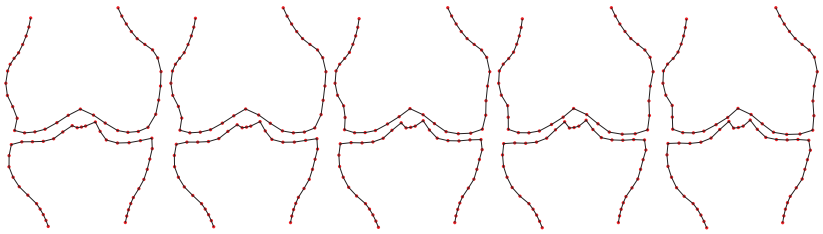

Figure 6.1: LDA interpolated shape model between the two classes: non-painful knees (left), and painful knees (right).

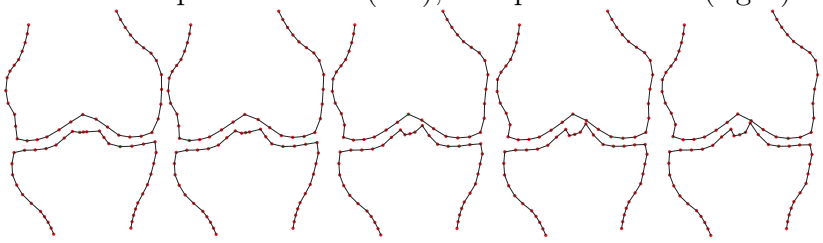

Figure 6.2: Later onset OA LDA shape model with no later onset OA (left) and later onset OA (right).

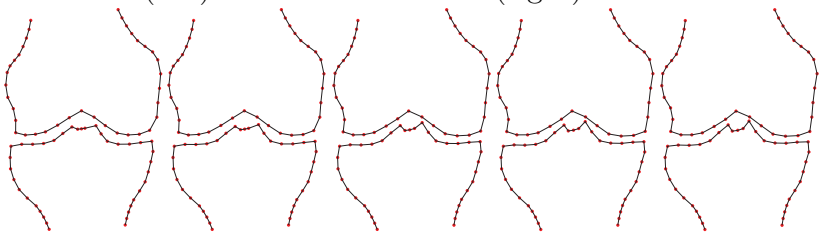

Figure 6.3: Later onset pain LDA shape model between control mean shape (left) and case mean shape (right).

6.2.2 Trabeculae Structure

The trabeculae texture extraction method was altered to handle the larger image sets. The increased number of images meant that the samples per region could be decreased from 670 to 100 (current experiments) and 150 (later onset experiments). This allowed more efficient run-time with minimal decrease in accuracy. The optimal region size $(0.2r \times 0.1r)$ with samples of 32×32 pixels was kept the same for all experiments.

6.2.3 Osteophytes

The osteophyte features are extracted using Haar features and SSM-DP contours placed on the marginal regions of the knee (medial tibia, lateral tibia, medial femur and lateral femur). The SSM-DP features used 85% variation; this equalled 31 shape modes for later onset OA, and 30 shape modes for the remaining experiments. The LDA shape modes (Figures 6.4-6.6) show the differences between the case/control classes mean shape. The later onset OA and pain LDA shape models (Figures 6.5 and 6.6) have similar medial tibia change. The current pain shape (Figure 6.4) shows some medial JSN and knee shape change between current pain and no-pain classes.

Figure 6.4: LDA shape models of non-painful knees (blue points) and painful knees (red points).

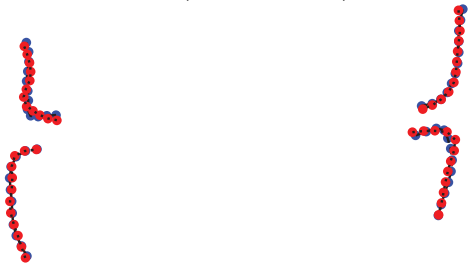

Figure 6.5: Later onset OA LDA shape models of case (blue points) and control (red points).

Figure 6.6: Later onset pain LDA shape models of case (blue points) and control (red points).

6.2.4 Tibial Spines and Intercondylar Notch

The tibial spines and intercondylar notch information is extracted from a region placed over the central area of the joint space (see Figure 4.32). Haar features are used to extract edge and shading information across the region.

6.2.5 Joint Space Shape Model

The joint space features are extracted using shape points placed along the medial and lateral compartment (JS-SSM). The shape models built from these points (see Figure 4.29) explain 99% of the shape variation, which equated to 18-20 shape modes for each of the experiments. The LDA shape models in Section 5.2 show the variation across the different experiments.

6.2.6 Comparative Methods

In the experiments we compare the combined feature model to both KL grade classification and the WND-CHARM algorithm. The WND-CHARM algorithm, explained in Section 3.4.5, reports high accuracies in analysing current [9] and later onset OA [12].

Current manual assessments of pain and later onset OA depend on the KL grades or multiple Joint Space Width (xJSW) measurements. The previous Chapter 5 demonstrated that the JS-SSM features achieve a higher accuracy than xJSW. The experiments compared the combined features against detection using manual KL grades - except in current OA where the KL grade is the outcome. The KL grade is taken from the gold-standard available in the OAI data, detection is assessed by generating a Receiver Operating Characteristic (ROC) curve and Area Under the ROC Curve (AUC) for the KL grade vs. outcome.

6.3 Experiments

The experiments train and test Random Forest (RF) classifiers (Section 3.5.1) on the independent radiographic features. The combined model uses the outputs from all RFs, taking the mean as the image classification. The classifiers are evaluated using 5-fold cross validation (Section 3.5.2) and are compared using the mean AUC and 95% Confidence Intervals (CI) (two-class experiments) and mean accuracy with weighted kappa (kw) (OA multi-class).The results for each experiment are adjusted for the correlation between participants with both knees included in the data following

the statistical analysis explained in Section 3.5.3. The difference in AUC between the independent and correlated knees results in all experiments was < 0.01. The combined feature models are compared against the KL grades and WND-CHARM algorithms (see Section 6.2.6). The WND-CHARM algorithm is a self-contained program; as such we were unable to adjust for the correlation between participant knees. The symbol ** will be used in all the tables to state that the best accuracy is significant compared to the other results reported in the same table.

6.3.1 Current Osteoarthritis

The current OA experiments evaluate the optimal fully combined features (Section 4.4.1) to detect OA (two-class) and KL grades (multi-class). The results in Table 6.1 show that the best independent feature is osteophytes (OS) (see Figure 6.7), and combining all features increases the AUC from 0.887 (OS Haar features) to 0.903 (fully combined). The WND-CHARM algorithm achieves a lower AUC (see Fig. 6.8) with 0.704 (WND-CHARM) to 0.903 (fully combined). The model was also compared with the previous results in Section 4.3.6. The model run on 747 images in the previous Chapter 4 achieves a higher AUC than the 8875 images: 0.903 (8875 images), 0.939 (747 images). This may be caused by the inclusion of lateral OA images or poor image quality in the baseline set.

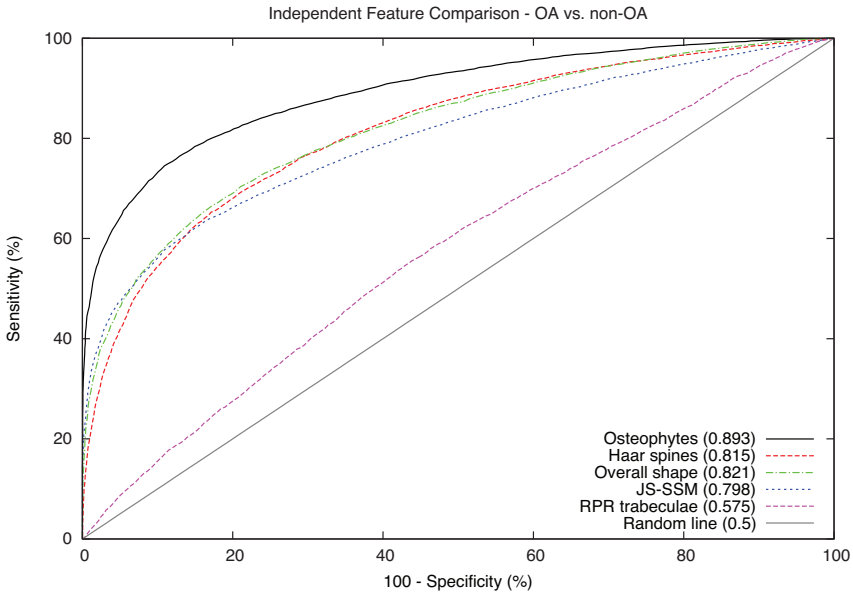

Figure 6.7: ROC curves of all independent features: osteophytes (SSM-DP + Haar), Join space shape models (JS-SSM), overall shape SSM, tibial spines (Haar features), trabeculae (RPR) in detecting OA vs. non-OA images.

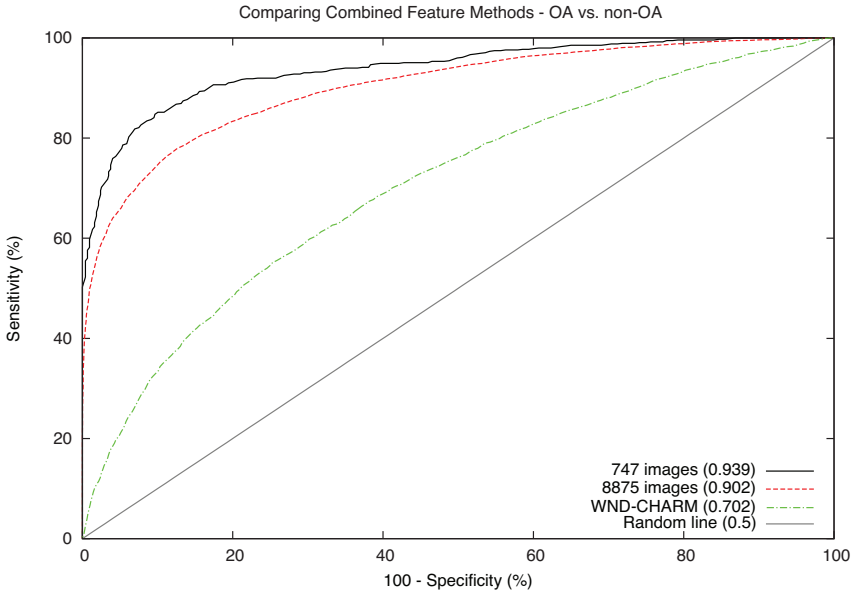

Figure 6.8: Comparing the OA detection of the previous experiments (747 images), current experiments (8875 images) and the WND-CHARM algorithm (8875).

The optimal multi-class features are taken from the experiments on the smaller dataset (Section 4.3.7), and contain: osteophytes (Haar and SSM-DP features), tibial spines and overall shape features.

The results show only a marginal difference between the combined osteophyte features and the optimal combined features (see Table 6.2), with the mean accuracies: 54.2% ±0.2 (osteophyte Haar features) and 55.6% ±0.2 (combined model). The optimal combined features achieve a higher accuracy than the WND-CHARM algorithm, with 36.5% (WND-CHARM) and 55.6% (combined model).

Table 6.1: Current OA Two-class Detection Results

Analysis Method	AUC (CI 95%)
Overall shape	0.824 (0.82 - 0.83)
RPR trabeculae	0.576 (0.56 - 0.59)
OS Haar	0.887 (0.88 - 0.89)
OS SSM-DP	0.812 (0.8 - 0.82)
Spines	0.818 (0.81 - 0.83)
JS-SSM	0.76 (0.75 - 0.77)
WND-CHARM	0.704 (0.69 - 0.71)
Combined features	
Combined features	**0.904 (0.9 - 0.91)****

Table 6.2: Current OA Multi-class Classification Results

	Accuracy (%)						
Analysis Method	KL 0	KL 1	KL 2	KL 3	KL 4	Overall(stdev.)	kw(CI 95%)
Overall shape	68.6	15.0	27.1	27.8	26.3	41.1 (0.0)	0.28(0.26-0.3)
RPR trabeculae	60.4	13.7	21.3	5.7	0.5	32.4 (0.1)	0.01(0-0.02)
OS Haar	86.0	6.6	**45.1**	45.8	**42.9**	54.2 (0.4)	0.47(0.45-0.48)
OS SSM-DP	70.0	14.8	27.7	45.0	34.3	44.4 (0.5)	0.34(0.33-0.36)
Spines	72.7	**16.1**	30.6	24.5	14.2	43 (0.7)	0.3(0.3-33)
JS-SSM	68.0	13.3	28.1	30.0	31.2	41.3 (0.3)	0.29(0.27-0.3)
WND-CHARM	38.6	30.0	27.6	29.5	57.0	36.5 (2.2)	-
Combined features							
Combined features	**91.2**	3.6	43.6	**49.9**	39.2	**55.6 (0.2)****	**0.49(0.48-0.5)**

The baseline images contain both medial and lateral OA knees, with 3264 medial OA and 588 lateral OA. Lateral OA was defined as OA positive images with a higher

lateral compartment JSN grade. The data was split into medial and lateral OA, with all 5023 non-OA used in each. The AUCs were generated using subsets of the RF outputs from the combined medial and lateral experiments above. The results show a much higher AUC for JS-SSM features in the lateral OA detection (see Table 6.3), with AUCs: 0.765 (all OA), 0.778 (medial OA) and 0.929 (lateral OA). The combined features show similar results, with: 0.903 (all OA), 0.891 (medial OA) and 0.971 (lateral OA). Multi-class accuracies show an increase in the lateral set for the optimal combined model, with: 55.6% ±0.2 (all OA), 56.5% ±0.1 (medial), and 59.5% ±0.6 (lateral). The kw is higher in the medial subset: medial - 0.48 (CI: 0.47 - 0.5), lateral - 0.36 (CI: 0.33 - 0.38).

Table 6.3: Current OA - Medial and Lateral OA Detection Results

	All OA	Medial OA	Lateral OA
Analysis Method	AUC (CI 95%)	AUC (CI 95%)	AUC (CI 95%)
Overall shape	0.824 (0.82 - 0.83)	0.808 (0.8 - 0.82)	0.909 (0.9 - 0.92)
RPR trabeculae	0.576 (0.56 - 0.59)	0.578 (0.57 - 0.59)	0.568 (0.54 - 0.59)
OS Haar	0.887 (0.88 - 0.89)	0.877 (0.87 - 0.89)	0.944 (0.93 - 0.95)
OS SSM-DP	0.812 (0.8 - 0.82)	0.795 (0.78 - 0.8)	0.914 (0.9 - 0.93)
Spines	0.818 (0.0.81 - 0.83)	0.8 (0.79 - 0.81)	0.9 (0.89 - 0.91)
JS-SSM	0.765 (0.75 - 0.77)	0.778 (0.77 - 0.79)	0.929 (0.92 - 0.94)
Combined Features			
Combined features	**0.903 (0.9 - 0.91)****	**0.891 (0.88 - 0.9)**	**0.971 (0.96 - 0.98)****

6.3.2 Current Pain

The extracted features are used to detect current pain across the 8847 images from baseline. We compare our combined model against KL grade and WND-CHARM detection on the same images (see Figure 6.10). The osteophyte features (Haar features and SSM-DP) achieve the highest AUC with 0.634. The fully combined model achieves higher detection accuracy than KL grade and WND-CHARM, with AUCs: 0.574 (WND-CHARM), 0.629 (KL grades) and 0.661 (combined model). The KL grade was added to the combined model output to form a Composite result, this achieves a similar AUC of: composite - 0.666 (CI: 0.65 - 0.68), combined - 0.663 (CI: 0.65 - 0.68).

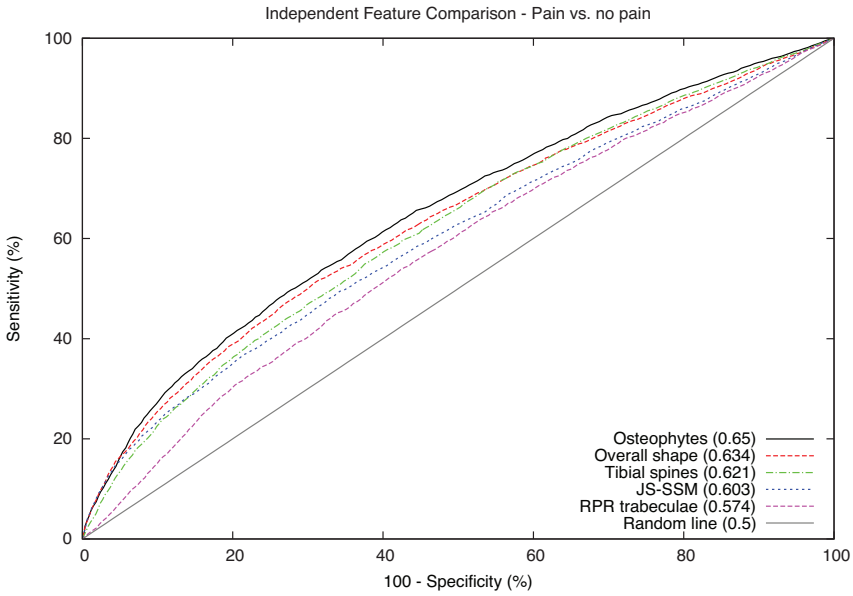

Figure 6.9: Detection of current pain using independent features: osteophytes (SSM-DP and Haar features), joint space (JS-SSM), overall shape, tibial spines (Haar features), and trabeculae (RPR).

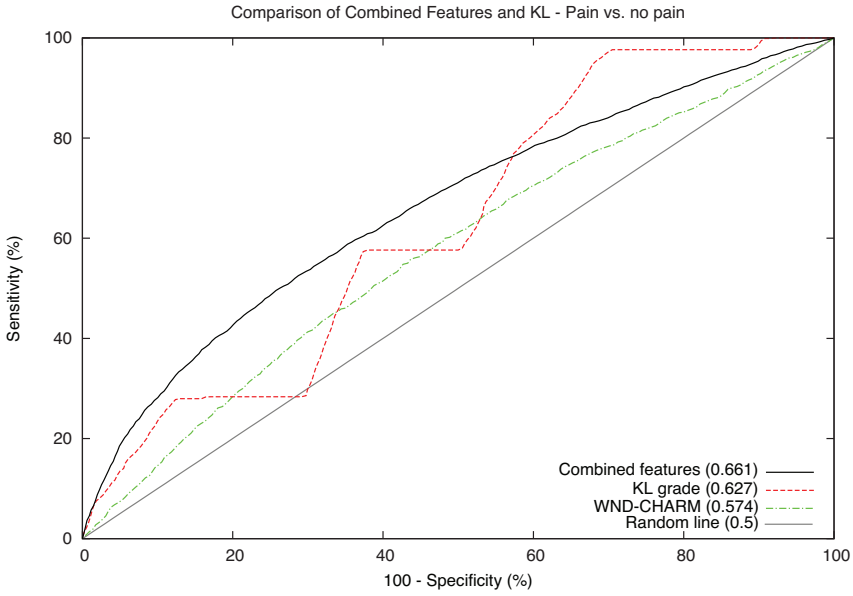

Figure 6.10: Comparing the combined features, WND-CHARM and KL grade in detecting current pain.

Table 6.4: Current Pain Detection Results

Analysis Method	AUC (CI 95%)
Overall shape	0.636 (0.62 - 0.65)
RPR trabeculae	0.572 (0.56 - 0.58)
OS Haar	0.645 (0.63 - 0.66)
OS SSM-DP	0.62 (0.61 - 0.63)
Spines	0.625 (0.61 - 0.64)
JS-SSM	0.605 (0.59 - 0.62)
WND-CHARM	0.57 (0.56 - 0.58)
KL	0.629 (0.62 - 0.64)
Combined Features	
Combined features	0.663 (0.65 - 0.68)
Composite (KL + combined)	**0.666 (0.65 - 0.68)**

6.3.3 Later Onset Osteoarthritis

The later onset OA experiments use the radiographic features to split baseline images with no OA (KL $\leq$ 1) into two classes: developing OA during the follow-up visits (cases), and not developing OA over follow-up (controls). The osteophytes achieve the best independent AUC with 0.596 (see Fig. 6.11). The combined model is compared to KL grade and WND-CHARM prediction features. The composite features (combined model + KL grade) achieves a higher AUC than the combined model and the WND-CHARM algorithm (see Fig. 6.12), with: 0.614 (combined), 0.583 (WND-CHARM) and 0.746 (composite features).

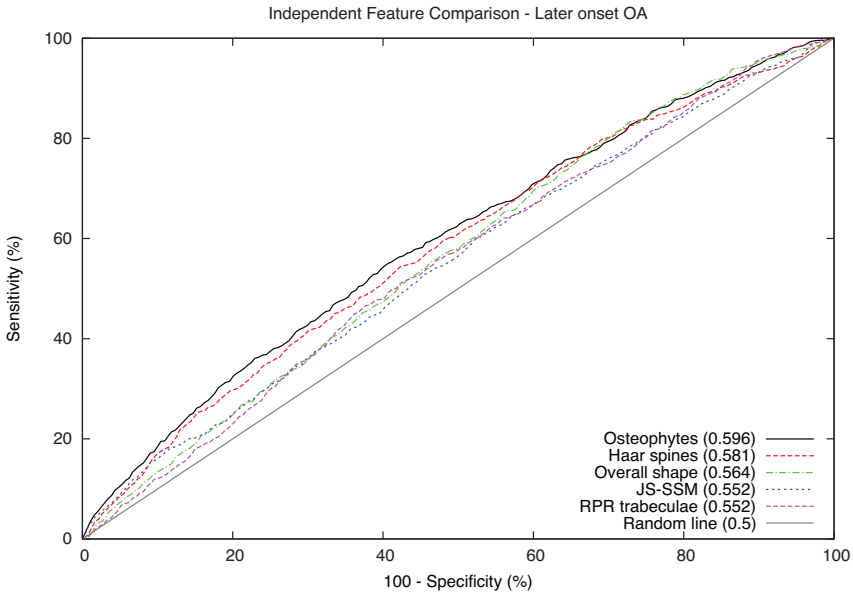

Figure 6.11: Prediction of later onset OA using: osteophytes (SSM-DP and Haar features), joint space (JS-SSM), overall shape, tibial spines (Haar features), and trabeculae (RPR).

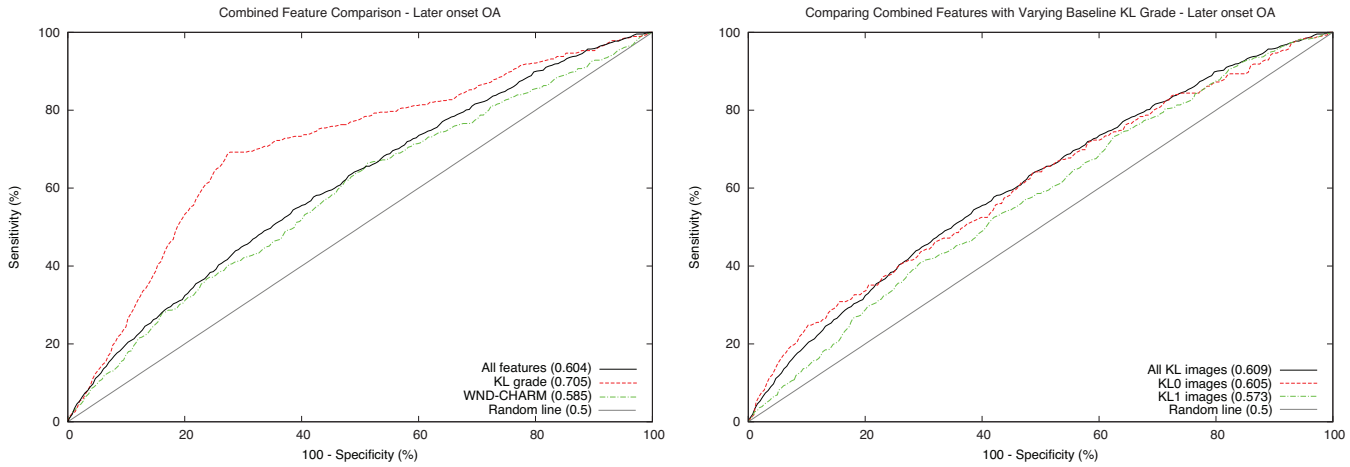

Figure 6.12: ROC curves comparing: combined features, WND-CHARM and KL grades (left). The figure to the right shows the combined feature accuracy when the data is split between the KL grades in the set.

The data was split into KL 0 and KL 1 images (see Table 6.5) to determine if either subset increased the prediction accuracy. The results found the KL0 set achieved a higher AUC: 0.619 (KL0), 0.563 (KL1).

Table 6.5: Later Onset OA Prediction Results

	AUC (CI 95%)		
Analysis Method	Both KL	KL 0	KL 1
Overall shape	0.56 (0.53 - 0.59)	0.592 (0.54 - 0.64)	0.517 (0.48 - 0.55)
RPR trabeculae	0.545 (0.53 - 0.59)	0.548 (0.5 - 0.6)	0.541 (0.51 - 0.58)
OS Haar	0.593 (0.57 - 0.62)	0.581 (0.53 - 0.63)	**0.566 (0.53 - 0.6)**
OS SSM-DP	0.543 (0.52 - 0.57)	0.563 (0.51 - 0.61)	0.513 (0.48 - 0.55)
Spines	0.584 (0.56 - 0.61)	0.61 (0.56 - 0.66)	0.53 (0.49 - 0.56)
JS-SSM	0.554 (0.53 - 0.58)	0.585 (0.53 - 0.63)	0.519 (0.48 - 0.55)
WND-CHARM	0.583 (0.55 - 0.61)	0.542 (0.49 - 0.59)	0.551 (0.52 - 0.59)
KL	0.709 (0.69 - 0.73)	0.463	0.506
Combined Features			
Combined features	0.614 (0.59 - 0.64)	**0.619 (0.57 - 0.67)**	0.563 (0.53 - 0.6)
Composite features	**0.746 (0.72 - 0.77)**	**0.619 (0.57 - 0.67)**	0.563 (0.53 - 0.6)

6.3.4 Later Onset Pain

The later onset pain experiments predict baseline non-painful knees as either, developing pain during follow-up visits (case), or not developing pain over follow-up (control). The radiographic features are applied to the 6206 images from baseline (see Table 6.6). The best independent feature to predict the data is the overall shape, with an AUC of 0.603 (see Figure 6.13). Combining all features increases the AUC to 0.609; the best AUC is from the composite model with 0.617. The RPR trabeculae features were removed from the combined model because of the low accuracy (AUC: 0.527). This improved the combined model from 0.604 (with RPR), to 0.609 (without RPR). The combined model is compared against KL grade and WND-CHARM models (see Figure 6.14), and achieves the highest accuracy, with: 0.531 (WND-CHARM), 0.6 (KL grade), and 0.608 (combined model). All results reported are adjusted for the contralateral knees as the differences in AUC is < 0.004 for all features.

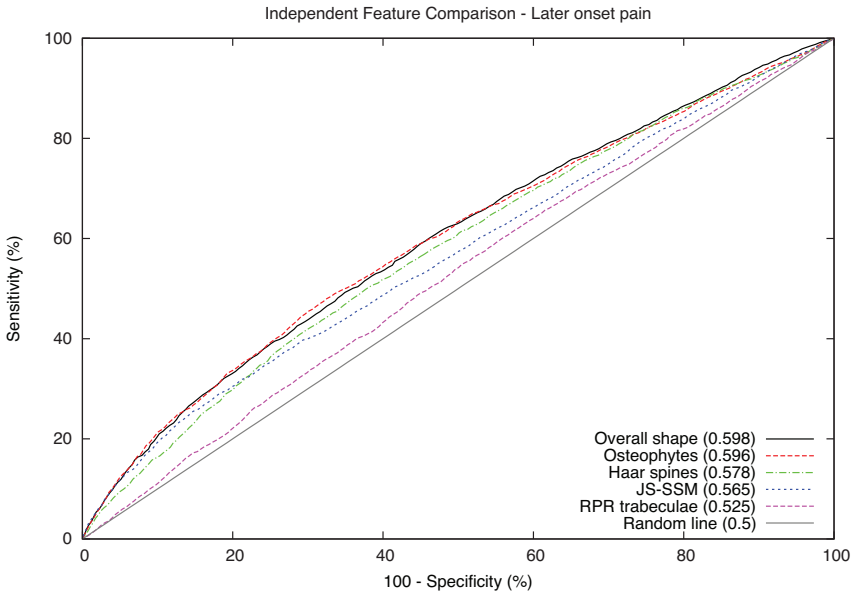

Figure 6.13: Prediction of later onset pain using: osteophytes (SSM-DP and Haar features), joint space (JS-SSM), overall shape, tibial spines (Haar features), and trabeculae (RPR).

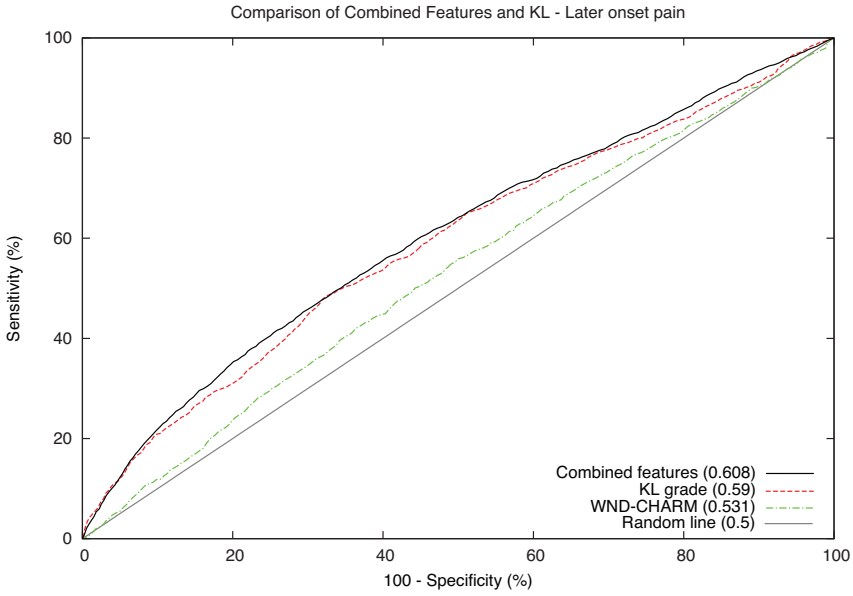

Figure 6.14: Comparison of combined features with WND-CHARM and KL grade prediction.

Table 6.6: Later Onset Pain Prediction Results

Analysis Method	AUC (CI 95%)
Overall shape	0.603 (0.59 - 0.62)
RPR trabeculae	0.527 (0.51 - 0.54)
OS Haar	0.59 (0.57 - 0.6)
OS SSM-DP	0.585 (0.57 - 0.6)
Spines	0.584 (0.57 - 0.6)
JS-SSM	0.57 (0.56 - 0.59)
WND-CHARM	0.528 (0.51 - 0.54)
KL	0.6 (0.59 -0.62)
Combined Features	
Combined features	0.604 (0.58 - 0.61)
Combined without RPR	0.609 (0.59 - 0.62)
Composite features	**0.617 (0.6 - 0.63)**

6.4 Discussion

The results have shown that combining multiple radiographic features improves prediction accuracy for all four experiments (current OA, current pain, later onset OA and later onset pain). Our combined explicit and implicit radiographic features achieve a higher accuracy than the WND-CHARM algorithm in each experiment. This shows that implicitly capturing radiographic OA features is weaker than combining multiple explicit and implicit features from the same knee.

The images from each set were restricted to knees with disease outcomes available, and no reported replacement surgery. No adjustments for image quality were made, meaning some pixelated and over/under exposed images were included in the experiments (see Figures 6.15).

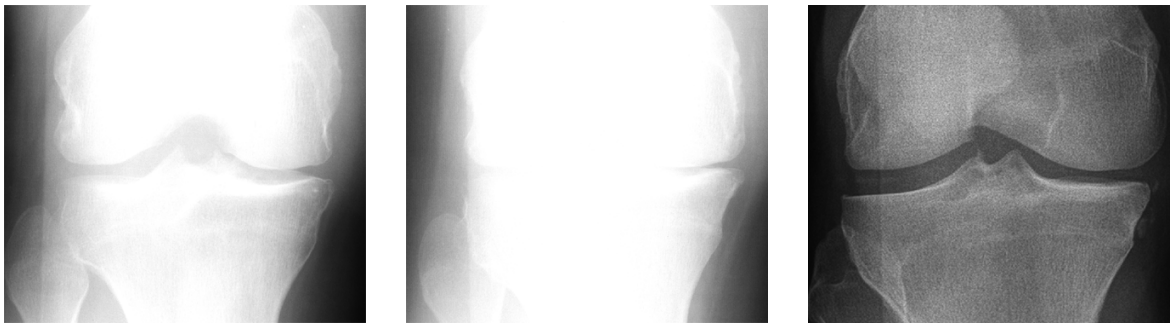

Figure 6.15: Examples of poor quality images found in the baseline datasets.

Current OA The results show that the best independent feature for current OA detection is the marginal osteophytes with an AUC of 0.893. This is improved with the fully combined radiographic features to an AUC of 0.902 and an overall multi-class accuracy of 55.6%. The kw when comparing the results to the gold standard manual KL grades is within the range of manual grading, with combined features - 0.49 (CI: 0.48-0.51), and manual grading - 0.58 (0.36 - 0.8). The OAI conducted a reliability test on 150 participant images from baseline, achieving a kw of 0.70 (CI: 0.65-0.76). The manual reliability could achieve higher because of the smaller subsample of images compared to the larger set of baseline images (8875 images) used in our optimal combined algorithm.

The combined feature AUC is lower in the larger OA experiments, with: 0.903 (8875 images) to 0.939 (747 images). This drop in accuracy is likely to come from the poor images included in the data (Figures 6.15). The overall multi-class accuracy is higher for the larger data, with: 55.6% (8875 images) and 51.8% (747 images). This is from the number of KL0 images shifting the overall mean higher. The proportion of KL0 images are larger in the 8875 image set (8875 - 38.9% KL0, 747 - 22.7% KL0) and achieves a multi-class accuracy of 91.2%, the remaining grades have significantly lower accuracies (3.6% - 49.9%).

The current OA images include both medial OA (3264) and lateral OA (588) images. We split the data to determine the accuracy of the two compartments, with a fully combined AUC of 0.903 (all) to 0.891 (medial) and 0.971 (lateral). This increase in AUC is especially seen in the JS-SSM (medial: 0.778, lateral: 0.929), shape (medial: 0.808, lateral: 0.909) and tibial spine (medial: 0.8, lateral: 0.9) features. Clearer

features defining the disease in the lateral compartment may potentially explain the increase in lateral joint space accuracy. The distribution of grades in the subsets supports this, with: KL2 - 63.9% (medial), 44% (lateral), KL3 - 29.9% (medial), 40.6% (lateral), KL4 - 6.2% (medial), 15.3% (lateral). The medial images contain a higher proportion of KL2 and a lower proportion of KL4 images than the lateral OA set. KL2 images are often mistaken for non-OA images with both medial and lateral OA sets classifying the majority of KL2 images as KL0 and KL1: medial - 50.5% (KL0), 14.0% (KL1), 28.1% (KL2), lateral - 55.4% (KL0), 14.9% (KL1), 19.5% (KL2). The higher proportion of KL4 images also makes the lateral set easier to distinguish in the two-class experiments (OA vs. non-OA) as the features are more pronounced than the less severe grades.

The trabeculae features have dropped the most in accuracy from the previous experiments (Section 4.3.7), with an AUC of 0.703 (747 images) to 0.576 (8875 images). Trabeculae texture is restricted to the medial subchondral bone, however, splitting the data into medial and lateral compartments only marginally improved the accuracy, with: 0.576 (all OA), 0.578 (medial OA), and 0.568 (lateral OA). Poor image quality is likely to be the cause of the lower accuracy as the RPR features are based on unadjusted intensities taken from raw image data. This means that any images with over and under exposed regions like in Figure 6.15, would result in trabeculae features with minimal structure information.

Later Onset Osteoarthritis The experiments show that prediction of future disease is higher using KL grade features to split the data, with KL1 images more likely to develop later onset disease than KL0 images. The KL grade improves the accuracy of the combined feature model, with: 0.614 (combined features), 0.705 (KL grade), to 0.746 (composite features). The images were split by KL grade to determine if the combined feature analysis performed better in the separate subsets. The results show that the KL1 group achieved a worse AUC: 0.614 (all KL), 0.619 (KL0), and 0.563 (KL1). This outcome could be caused by the smaller number of images and disproportionate amount of case images to controls in the two sets, with: KL0 - 141 case and 3311 controls (3452 images), and KL1 - 319 case and 1250 controls (1569 images).

The proportion of images in the KL0 set means that if the model predicts most of the images as not developing disease, then it will automatically get a higher AUC than the KL1 set which has a higher proportion of case examples.

Prediction of later onset OA features is fairly poor, with a combined AUC of 0.614. The combined features improve the accuracy from the JS-SSM (0.54) and the combined joint space features (0.543) from Section 5.3.2. The optimal AUC is also higher than the WND-CHARM algorithm (0.583). The WND-CHARM algorithm in [12] reports a higher overall accuracy of 62%-72% using a different dataset. The experiments used KL0 baseline images and predicted later onset disease 20 years after baseline. The current experiments analyse change over 6 years follow-up. The OAI dataset has taken follow-up scans later than 6 years after baseline, these visits have KL and JSN grades provided by the separate centres, in the future similar experiments will be conducted to analyse OA development and compare the automated radiographic feature methods over the larger time frame.

The KIDA algorithm also reports results on later onset OA prediction in [79], the AUC achieved is higher than the combined features, with 0.614 (combined) and 0.74 (KIDA). This is likely because of the clinical information included in the KIDA features, which uses BMI and gender to split the data. The purpose of the combined radiographic feature model was to predict disease outcomes without any manual or clinical information to determine the accuracy of a fully automated system. Later experiments will be run including similar participant information to improve the accuracy.

Pain Experiments Current pain experiments show that osteophyte Haar features are the best independent feature at detecting current pain, with an AUC of 0.645. The results support findings in [11] [30] and [110] that found osteophyte grade and area better detectors for pain than other OA features. This is improved to 0.663 with the combined radiographic features. The composite features achieve the best AUC with 0.666, compared to the WND-CHARM algorithm (0.57), KL grade features (0.627) and combined features (0.663).

The tibial spines and intercondylar notch show a relatively high AUC in detecting current pain, with an AUC of 0.625. This is supported by the KIDA algorithm [11] which found the height of the tibial spines correlated with current pain in a set of 1002 participants from the CHECK cohort.

The combined features achieve a higher AUC than the results in [32] which analyse semi-quantitative features and JSW separately. The semi-quantitative features combine grades for: osteophytes, sclerosis, cysts, attrition, JSN and Chondrocalcinosis. The current pain AUCs are 0.695 JSW and 0.62 semi-automated. The combined features likely achieve a higher AUC from the combination of JSW and the other radiographic features. The later onset results in the same paper show a better detection AUC with JSW and semi-quantitative features over the time points, JSW - 0.623(T0) - 0.62(T2y), and semi-quantitative - 0.62(T0) - 0.61(T2y). This could be higher because of the smaller datasets the features were taken from (163 JSW images and 123 semi-quantitative images), or the inclusion of extra features not specifically measured by our combined model (cysts, sclerosis, and Chondrocalcinosis).

The later onset pain experiments show the overall shape features achieve the best independent AUC (0.603). The optimal combined features achieve an AUC of 0.609. The optimal method contains all radiographic features except the RPR trabeculae, which lowered the AUC to 0.604. This is likely because the trabecular structure has a weak accuracy in predicting later onset pain (AUC: 0.527) and the increased features (150 examples per image) shifted the mean taken across all RF classifier outputs.

The results find a weak correlation between clinical pain and radiographic features, similar to findings in the KIDA [11] and KOACAD [6] algorithms. This lack of strong association is well documented in previous literature [108] [109], which have found that clinical symptoms of pain and function are highly variable in participants with radiographic knee OA. The findings have also shown that including more radiographic views creates a better assessment of OA and the features present in the knee. This allows a better correlation to be found between the features of OA across multiple views per participant and the presence of pain in the respective knee. Future development

will include lateral view radiographs to determine if the assessment on current and later onset pain can be improved.

The table below (Table 6.8) shows a comparison between our algorithm and the various automated OA methods from the literature.

Table 6.7: Comparison of Automated OA and Pain Methods

				Features Evaluated (AUC, or accuracy (%))					
Outcome	Method	Dataset	No.	JS	OS	Tr	Im	TS	Combined
Current OA	SDM [8]	N/A	41			**85.4%**			
	KOACAD [68]	ROAD	3040 (p)	0.622	0.645			0.645	N/A
	CLAHE [73]	N/A	308				53.3%		
	WND-CHARM [9]	N/A	350 (k)				**91.5%**		
	WND-CHARM	OAI	8875 (k)				0.704 (36.5%)		
	Our algorithm	OAI	8875 (k)	**0.76**	**0.85**	0.576 (32.4%)	0.824 (41.1%)	**0.818**	**0.903**
Current Pain	Multivariate model [32]	OAI	163	0.534	0.577				**0.695**
	KIDA [11]	CHECK	1002	N/A	N/A			N/A	0.6
	Manual KL grade [111]	OAI	8847 (k)						0.629
	WND-CHARM	OAI	8847 (k)				0.57		
	Our algorithm	OAI	8847 (k)	**0.605**	**0.633**	**0.572**	**0.636**	**0.625**	0.663
Later onset pain	Multivariate model [32]	OAI	163	**0.585**	0.56				**0.62**
	Manual KL grade [111]	OAI	5023 (k)						0.6
	WND-CHARM	OAI	8847 (k)				0.528		
	Our algorithm	OAI	8847 (k)	0.57	**0.588**	**0.527**	**0.603**	**0.584**	0.609
Longitudinal OA	WND-CHARM [12]	N/A	246 (k)				67%		
	KIDA [79]	CHECK	1002	N/A	N/A			N/A	**0.74**
	Manual KL grade [111]	OAI	6206 (k)						0.709
	WND-CHARM	OAI	6206 (k)				**0.583**		
	our algorithm	OAI	6206 (k)	**0.554**	**0.568**	**0.545**	0.56	**0.584**	0.614

Table 6.8: JS = Joint Space, OS = Osteophytes, Tr = Trabeculae, Im = Implicit OA features, TS = Tibial spines, Combined = optimal combined features. Numbers = number of participants (p) or knee images (k) used in the studies. N/A = no reference to the values, or representative values that can be compared, found in the paper. CHECK = Cohort Hip and Cohort Knee. ROAD = Research on Osteoarthritis Against Disability.

Chapter 7

Discussion and Future Work

The project has shown that combining explicit and implicit radiographic features improves Osteoarthritis (OA) and related pain detections. Our combined feature model achieved better results than other state-of-the-art algorithms tested on the same data (WND-CHARM), and on different datasets (KIDA and KOACAD). The model was tested using large sets of radiographs, which were not filtered for image quality. This shows that the algorithm can predict the level of knee OA on data representative of real-world clinical radiographs.

The fully automated methods improve on manual grading methods: removing the need for operator input, providing consistent detection accuracies on large datasets, and evaluating features across the whole knee (including trabeculae structure and tibial spines). The current OA classification results achieve a similar inter-observer reliability to manual KL grading when comparing the automated results to the gold standard manual KL grades, with weighted kappa (kw): combined - 0.58 (CI: 0.54-0.62), and manual - 0.58 (0.36 - 0.8). Combining manual Kellgren and Lawrence (KL) grades improves the accuracy of the combined features in all experiments (composite features), meaning the automated method still needs improvement to capture all information seen by observers in the radiographs.

The project demonstrates the first fully automated analysis of marginal osteophytes, analysing both shape and texture information over the four regions. Classifiers based on osteophytes achieve the highest independent accuracy in all experiments. The

classification of OARSI osteophyte grades achieved high two-class osteophyte detection accuracy, but a weaker multi-class accuracy when compared to the gold standard manual OARSI grades.

The trabecular structure analysis compared various fractal signature and texture analysis methods, finding Raw Pixel Ratios achieved the highest current OA detection. The trabeculae achieved low accuracies in the pain and later onset pain and OA experiments; this is attributed to the features having minimal association with the outcomes and poor image quality.

The joint space shape models achieve a higher accuracy than multiple Joint Space Width (xJSW) measurements from the Osteoarthritis Initiative (OAI) data, and achieve a comparable kw accuracy when comparing the results against the gold standard manual Joint Space Narrowing (JSN) grading. The combination of both joint space features increases the accuracy in all two-class experiments; later work will improve on the Joint Space Statistical Shape Model (JS-SSM) features by capturing the accurate quantitative measurements.

The pain and later onset experiments show that there is some correlation between the outcomes and radiographic features, but these results are still weak. Further analysis will be done in the future to improve results with features from other radiographic views (i.e. skyline and lateral).

7.1 Future Work

This section suggests ways in which the work could be extended.

Pain Assessments The pain assessment variable used in the project is a good measure for recent pain (over the last 30 days), however, there are many pain assessments available in the data. A comprehensive analysis of pain should be run using WOMAC pain scores and any pain experienced over a longer period of time, i.e. over the last 12 months. This will add extra detail to the type of pain (WOMAC) and remove any

baseline knees that experienced less frequent pain, longer than 30 days prior to the baseline visit.

Combining Radiographic Views Another aspect of the project to be evaluated is the combination of features in other radiographic views. This will include shape, texture, osteophytes and alignment features from lateral and eventually skyline views of the same knee. The extra information will give a better analysis on the amount of OA present and development of features missed in Posterior-Anterior (PA) view knees. This will lead to better current OA and pain assessments, and potentially add more features to predict later onset disease and pain. Current work is being conducted on the combination of lateral and PA knee views.

Clinical Features Some current assessments of later onset disease include clinical features to increase the prediction accuracy, such as BMI and gender of the participant. The inclusion of these features would be useful in determining correlations between radiographic and clinical features, and improving detection and prediction accuracy.

Trabeculae Features The trabeculae features are dependent on image quality. More experiments will be conducted in the future with poor quality images filtered out of the datasets. The re-run experiments will then compare the trabecular structure with the current and later onset outcomes. We will also evaluate Fractal Signature (FS) features, reported to be robust to poor image quality, on the pain and later onset OA experiments.

Longitudinal Data The experiments in this project all use cross-sectional data. The future work of this project will analyse the longitudinal change of the features across a single participant. This analysis will evaluate how the features change across the follow-up visits and create a stronger understanding of the progression of the disease. This can be analysed across many participants to develop a threshold of the change in features that determines OA progression and the development of early Osteoarthritic features. The longitudinal data will include the later OAI images > 6 years follow-up to evaluate development of OA and pain over a longer time frame.

Comparison to 3D Features This project has focused on manually assessed outcomes that are widely documented as reliant on subjective views of the observer. To quantitatively evaluate the radiographic features grading of 3D MRI features could be used. The features would each be compared to the 3D counterparts: osteophytes compared with osteophyte grades in the same regions, JSN compared with cartilage scores, and trabeculae correlated with Bone Marrow Lesions (BMLs) scored in the region. The comparison would provide an insight into the extent the radiographic features describe the 3D features of the knee.

System Optimisation The features used in the project are optimised using preliminary two-class current OA data. Optimisation to multi-class and later onset outcomes will be run in the future. The JS-SSM and osteophyte features in particular require optimisation to capture all quantified joint space measurements, and improve osteophyte grade classification. The osteophyte features could be improved by adjusting the image contrast to make edges clearer before applying Statistical Shape Model with Dynamic Programming search (SSM-DP) contours, or expanding the Haar feature texture region to cover more of the surrounding areas. Altering the curve restrictions and combining multiple JS-SSM models per image with different curve constraints could improve the joint space features. The OARSI experiments used in Chapter 3 were taken from the subset of 747 images and were found sufficient for comparing the feature specific methods (osteophytes and joint space narrowing), however it is recognised that there is class unbalance in the lower proportion of OARSI grades 2 and 3 compared to the less severe grades (OARSI 0 and 1). This is a limitation of the work and later experiments with even OARSI grades can be run to fully evaluate how well the methods classify and detect osteophytes and joint space narrowing. Optimising the Random Forest Constrained Local Model (RFCLM) and manual annotations to improve detection accuracy on the bad images found in the experiments will also improve shape feature accuracies. The osteophyte and joint space features could be improved by using a multiresolution shape model, such as the ones used in [112], instead of the Dynamic Programming edge detection. The shape models could specifically be trained on osteophyte and joint space areas of the knee to improve detection accuracy.

New Datasets The OAI dataset contains images from four sites across the US. This data could be expanded using alternative OA and general knee datasets to determine how well the model performs using data with different angle and knee positioning specifications. Subchondral bone density (sclerosis) could also be measured in the future by analysing radiographs with an aluminium step wedge placed next to the knees. The image intensity of the subchondral bone is compared to the gradients of the wedge to determine the thickness of the bone. The use of a wedge can also be used to correct for any image contrast errors caused during the x-ray scanning procedure. The sclerosis analysis would use a region fitted to the area beneath the medial and lateral tibial plateaus, comparing the mean gradient to the varying thickness of the step wedge.

Bibliography

[1] December 2014. URL `www.arthritisresearchuk.org/mskcalculator`.

[2] J. H. Kellgren and J. S. Lawrence. Radiological assessment of osteoarthrosis. *Annals of the Rheumatic Diseases*, 16:494–502, 1957.

[3] 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*, 59:587–595, 2000.

[4] L. Sharma, J. Song, D. D. Dunlop, D. T. Felson, C. E. Lewis, N. A. Segal, J. Torner, T. D. V. Cooke, J. Hietpas, J. Lynch, and M. C. Nevitt. Varus and valgus alignment and incident and progressive knee osteoarthritis. *Annals of the Rheumatic Diseases*, 69(11):1940–1945, 2010.

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

[6] 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*, 16:1300–1306, 2008.

[7] G. Neumann, D. Hunter, M. Nevitt, L. B. Chibnik, K. Kwoh, H. Chen, T. Harris, S. Satterfield, and J. Duryea. Location specific radiograph joint space width for osteoarthritis progression. *Osteoarthritis and Cartilage*, 17(6):761–765, June 2009.

[8] T. Woloszynski, P. Podsiadlo, and G. W. Stachowiak. A signature dissimilarity measure for trabecular bone texture in knee radiographs. *Medical Physics*, 37 (5):2030–2042, May 2010.

[9] L. Shamir, S. M. Ling, W. W. Scott, A. Bos, N. Orlov, T. J. Macura, 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, February 2009.

[10] R. J. Barr, J. S. Gregory, K. Yoshida, S. Alesci, D. M. Reid, and R. M. Aspden. Knee joint shape assessed by active shape modelling of plain radiographs is related to osteoarthritis severity. In *ABSTRACTS OF THE 2009 WORLD CONGRESS ON OSTEOARTHRITIS*, volume 17, pages S200–S201. OsteoArthritis Society International, 2009.

[11] M. B. Kinds, A. C. A. Marijnissen, J. W. J. Bijlsma, M. Boers, F. P. J. G. Lafeber, and P. M. J. Welsing. Quantitative radiographic features of early knee osteoarthritis: Development over 5 years and relationship with symptoms in the check cohort. *The Journal of Rheumatology*, 40(1):58–65, 2013.

[12] L. Shamir, S. M. Ling, W. 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, October 2009.

[13] WebMD. Knee pain health centre, 2010. URL `http://www.webmd.com/pain-management/knee-pain/picture-of-the-knee`.

[14] Encyclopedia Britannica. Cancellous bone, 2013. URL `http://www.britannica.com/EBchecked/topic/92222/cancellous-bone`.

[15] D. B. Burr and M. A. Gallant. Bone remodelling in osteoarthritis. *Nature Reviews Rheumatology*, 8(11):665–673, November 2012.

[16] J-H. Chen, C. Liu, L. You, and C. A. Simmons. Boning up on wolff's law: Mechanical regulation of the cells that make and maintain bone. *Journal of Biomechanics*, 43:108–118, 2010.

[17] T. Neogi. Clinical significance of bone changes in osteoarthritis. *Theraputic Advances in Musculoskeletal Disease*, 4(4):259–267, 2012.

[18] J. H. Waarsing, S. M. A. Bierma-Zeinstra, and H. Weinans. Distinct subtypes of knee osteoarthritis: Data from the osteoarthritis initiative. *Rheumatology*, 54: 1650–1658, 2015.

[19] I. Sulzbacher. Osteoarthritis: Histology and pathogenesis. *Weiner Medizinische Wochenschrift*, 163:212–219, 2013.

[20] C. S. Shin and J. H. Lee. Arthroscopic treatment for osteoarthritic knee. *Knee Surgery and Related Research*, 24(4):187–192, December 2012.

[21] S. Amin, M. P. LaValley, A. Guermazi, M. Grigoryan, D. J. Hunter, M. Clancy, J. Niu, D. R. Gale, and D. T. Felson. The relationship between cartilage loss on magnetic resonance imaging and radiographic progression in men and women with knee osteoarthritis. *Arthritis and Rheumatism*, 52(10):3152–3159, October 2005.

[22] W. P. Chan, G-S. Huang, S-M. Hsu, Y-C. Chang, and W-P. Ho. Radiographic joint space narrowing in osteoarthritis of the knee: Relationship to meniscal tears and duration of pain. *Skeletal Radiology*, 37(10):917–922, October 2008.

[23] T. Derek V. Cooke, D. Siu, and B. Fisher. *Recent Developments in Orthopaedic Surgery*, chapter The use of standardized radiographs to identify the deformities associated with osteoarthritis, pages 264–273. Number ISBN0719025427. Manchester University Press, 1987.

[24] T. Derek V. Cooke, E. A. Sled, and R. Allan Scudamore. Frontal plane knee alignment: A call for standardized measurement. *Journal of Rheumatology*, 34: 1796–1801, 2007.

[25] T. M. Keaveny, E. F. Morgan, G. L. Niebur, and O. C. Yeh. Biomechanics of trabecular bone. *Annual Review of Biomedical Engineering*, 3:307–333, 2001.

[26] J. A. Lynch, D. J. Hawkes, and J. C. Buckland-Wright. Analysis of texture in macroradiographs of osteoarthritic knees using fractal signature. *Physics in Medicine and Biology*, 36(6):709–722, 1991.

[27] J. C. Buckland-Wright, J. A. Lynch, and D. G. Macfarlane. Fractal signature analysis measures cancellous bone organisation in macroradiographs of patients with knee osteoarthritis. *Annals of Rheumatic Diseases*, 55:749–755, 1996.

[28] L. Kamibayashi, U. P. Wyss, T. D. V. Cooke, and B. Zee. Changes in mean trabecular orientation in the medial condyle of the proximal tibia in osteoarthritis. *Orthopedic Surgical Forum*, 57(1):69–73, July 1995.

[29] L. Sharma, J. Song, D. T. Felson, S. Cahue, E. Shamiyeh, and D. D. Dunlop. The role of knee alignment in disease progression and functional decline in knee osteoarthritis. *The Journal of the American Medical Association*, 286(2):188 – 195, July 2001.

[30] F. M. Cicuttini, J. B. Deborah, J. Hart, and T. D. Spector. Association of pain with radiological changes in different compartments and views of the knee joint. *Osteoarthritis and Cartilage*, 4(2):142–147, 1996.

[31] S. Donnelly, D. J. Hart, D. V. Doyle, and T. D. Spector. Spiking of the tibial tubercles - a radiological feature of osteoarthritis? *Annals of Rheumatic Diseases*, 55:105–108, 1996.

[32] J. I. Galván-Tejada, J. M. Celaya-Padilla, V. Treviño, and J. G. Tamez-Peña. Multivariate radiological-based models for the prediction of future knee pain: Data from oai. *Computational and Mathematical Methods in Medicine*, 2015: 1–10, October 2015.

[33] R. D. Altman, M. C. Hochberg, W. A. Murphy, and F. Wolfe. Atlas of individual radiographic features in osteoarthritis. *Osteoarthritis and Cartilage*, 57:595–601, 1995.

[34] S. Ahlbäck. Osteoarthritis of the knee: A radiographic investigation. *Acta Radiologica Stockholme*, Suppl 277:7–72, 1968.

[35] D. I. Riddle, W. A. Jiranek, and J. R. Hull. Validity and reliability of radiographic knee osteoarthritis measures by arthroplasty surgeons. *Orthopedics*, 36 (1):e25–e32, 2013.

[36] D. Schiphofand B. M. de Klerk, H. J. Kerkhof, A. Hofman, B. W. Koes, M. Boers, and S. M. Bierma-Zeinstra. Impact of different descriptions of the kellgren and lawrence classification criteria on the diagnosis of knee osteoarthritis. *Annals of the Rheumatic Diseases*, 70:1422 – 1427, 2011.

[37] D. T. Felson, J. Niu, A. Guermazi, B. Sack, and P. Alibadi. Defining radiographic incidence and progression of knee osteoarthritis: Suggested modifications of the kellgren and lawrence scale. *Annals of the Rheumatic Diseases*, 70:1884–1886, 2011.

[38] K. Brandt, R. Fife, E. Braunstein, and B. Katz. Radiographic grading of the severity of knee osteoarthritis: Relation of the kellgren and lawrence grade to a grade based on joint space narrowing and correlation with arthroscopic evidence of articular cartilage degeneration. *Arthritis and Rheumatism*, 32:1584–1591, 1989.

[39] C. Wilkinson, A. Carr, and M. Doherty. Does increasing the grades of the knee osteoarthritis line drawing atlas alter its clinimetric properties? *Annals of the Rheumatic Diseases*, 64(10):1467 – 1473, October 2005.

[40] D. Hayashi, J. Gusenburg, F. W. Roemer, D. J. Hunter, L. Li, and A. Guermazi. Reliability of semiquantitative assessment of osteophytes and subchondral cysts on tomosynthesis images by radiologists with different levels of expertise. *Diagnostic and Interventional Radiology*, 20:353–359, 2014.

[41] L. Sheehy, E. Culham, L. McLean, J. Niu, J. A. Lynch, N. A. Segal, J. A. Singh, M. C. Nevitt, and T. D. V. Cooke. Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the multicenter osteoarthritis study (most). *Osteoarthritis and Cartilage*, 23(9): 1491–1498, September 2015.

[42] M. Galli, V. De Santis, and L. Tafuro. Reliability of the ahlback classification of knee osteoarthritis. *Osteoarthritis and Cartilage*, 11:580 – 584, 2003.

[43] N. E. Lane, M. C. Nevitt, H. K. Genant, and M. C. Hochberg. Reliability of

new indicies of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration. *Journal of Rheumatology*, 20:2307–2319, 1993.

[44] R. D. Altman, J. F. Fries, and D. A. Bloch. Radiographic assessment of progression in osteoarthritis. *Arthritis and Rheumatism*, 30:1214–1225, 1987.

[45] T. Conrozier and E. Vignon. Quantitative radiography in osteoarthritis: Computerized measurements of radiographic knee and hip joint space. *Bailliére's Clinical Rheumatology*, 10(3):429–433, August 1996.

[46] J. C. Buckland, I. Carmichael, and S. R. Walker. Quantitative microfocal radiography accurately detects joint changes in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 45:379–383, 1986.

[47] J. C. Buckland-Wright. Quantitative radiography of osteoarthritis. *Annals of Rheumatic Diseases*, 53:268–275, 1994.

[48] J. E. Dacre, J. S. Coppock, K. E. Herbert, and D. Perrett. Development of a new radiographic scoring system using digital image analysis. *Annals of the Rheumatic Diseases*, 48:194–200, 1989.

[49] M. Wolski, P. Podsiadlo, and G. W. Stachowiak. Directional fractal signature analysis of trabecular bone: Evaluation of different methods to detect early osteoarthritis in knee radiographs. In *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, volume 223, pages 211–236, 2009.

[50] S. D. Rockoff. Radiographic trabecular quantitation of human lumbar vertebrae in situ 1: theory and method for the study of osteoporosis. *Investigative Radiology*, 2(4):272–289, August 1967.

[51] P. Christensen, J. Kjaer, F. Melsen, H. E. Nielsen, O. Sneppen, and P-S. Vang. The subchondral bone of the proximal tibial epiphysis in osteoarthritis of the knee. *Acta Orthopaedica Scandinavica*, 53:889–895, 1982.

[52] L. Pothuaud, C. L. Benhamou, P. Porion, E. Lespessailles, R. Harba, and P. Levitz. Fractal dimension of trabecular bone projection texture is related

to three-dimensional microarchitecture. *Journal of Bone and Mineral Research*, 15(4):691–699, 2000.

[53] A. P. Pentland. Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):661–674, 1984.

[54] S. Peleg, J. Naor, R. Hartley, and D. Avnir. Multiple resolution texture analysis and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(4):518–523, July 1984.

[55] V. B. Kraus, S. Feng, S-C. Wang, S. White, M. Ainslie, A. Brett, A. Holmes, and H. C. Charles. Trabecular morphometry by fractal signal analysis is a novel marker of osteoarthritis progression. *Arthritis and Rheumatism*, 60(12):3711–3722, December 2009.

[56] P. Podsiadlo and G. W. Stachowiak. The development of the modified hurst orientation transform for the characterization of surface topography of wear particles. *Tribology Letters*, 4:215–229, 1998.

[57] P. Podsiadlo and G. W. Stachowiak. Analysis of trabeculae bone texture by modified hurst orientation transform method. *Medical Physics*, 29(4):460–474, April 2002.

[58] M. Wolski, P. Podsiadlo, and G. W. Stachowiak. Directional fractal signature methods for trabecular bone texture in hand radiographs: Data from the osteoarthritis initiative. *Medical Physics*, 41(8):081914–1–081914–17, August 2014.

[59] M. Wolski, P. Podsiadlo, G. W. Stachowiak, L. S. 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*, 18:684–690, 2010.

[60] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution grey-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.

[61] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

[62] J. A. Lynch, J. C. Buckland-Wright, and D. G. Macfarlane. Precision of joint space width measurement in knee osteoarthritis from digital image analysis of high definition macroradiographs. *Osteoarthritis and Cartilage*, 1:209–218, 1993.

[63] J. Duryea, J. Li, C. G. Peterfy, C. Gordon, and H. K. Genant. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. *Journal of Medical Physics*, 27:580–591, 2000.

[64] H. Tariq and S. M. A. Burney. Contour extraction of femur and tibia condyles on plain aneroposterior (ap) radiographs. *International Journal of Computer Applications*, 52(15):26–30, August 2012.

[65] S. J. Grochowski, K. K. Amrami, and K. Kaufman. Semi-automated digital image analysis of patellofemoral joint space width from lateral knee radiographs. *Skeletal Radiology*, 34:644–648, July 2005.

[66] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, November 1986.

[67] L. G. Roberts. *Machine Perception of Three-Dimensional Solids*. PhD thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering, 1963.

[68] H. Oka, S. Muraki, T. Akune, K. Nakamura, H. Kawaguchi, and N. Yoshimura. Normal and threshold values of radiographic parameters for knee osteoarthritis using a computer-assisted measuring system (koacad): The road study. *Journal of Orthopaedic Science*, 15:781–789, 2010.

[69] T. Iranpour-Boroujeni, J. Li, J. A. Lynch, M. Nevitt, and J. Duryea. A new method to measure anatomic knee alignment for large studies of oa: Data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 22:1668–1674, 2014.

[70] J. C. Buckland-Wright, D. G. Macfarlane, M. K. Jasani, and J. A. Lynch. Quantitative microfocal radiographic assessment of osteoarthritis of the knee from weight bearing tunnel and semiflexed standing views. *Journal of Rheumatology*, 21(9):1734–1741, September 1994.

[71] M. R. Hayeri, M. Shiehmorteza, D. J. Trudell, T. Hefflin, and D. Resnick. Proximal tibial osteophytes and their relationship with the height of tibial spines of the intercondylar eminence: Paleopathological study. *Skeletal Radiology*, 39: 877–881, 2010.

[72] C. M. Bishop. *Pattern Recognition and Machine Learning*. Number ISBN 978-0-387-31073-2. Springer-Verlag, New York, 2006.

[73] L. Anifah, K. E. Purnama, M. Hariadi, and M. Purnomo. Osteoarthritis classification using self organising map based on gabor kernel and contrast-limited adaptive histogram equalisation. *The Open Biomedical Engineering Journal*, 7: 18–28, 2013.

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

[75] 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, January 1995.

[76] J. L. Astephen Wilson, K. J. Deluzio, M. J. Dunbar, G. E. Caldwell, and C. L. Hubley-Kozey. The association between knee joint biomechanics and neuromuscular control and moderate knee osteoarthritis radiographic and pain severity. *Osteoarthritis and Cartilage*, 19(2):186–193, February 2011.

[77] T. Neogi, M. A. Bowes, J. Niu, K. M. De Souza, G. R. Vincent, J. Goggins, Y. Zhang, and D. T. Felson. Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis. *Arthritis and Rheumatism*, 8(65):2048–2058, August 2013.

[78] M. A. Bowes, G. R. Vincent, C. B. Wolstenholme, and P. G. Conghan. A novel method for bone area measurement provides new insights into osteoarthritis and its progression. *Annals of the Rheumatic Diseases*, 0:1–7, December 2013.

[79] M. B. Kinds, A. C. A. Marijnissen, K. L. Vincken, M. A. Viergever, K. W. Drossaers-Bakker, J. W. J. Bijlsma, S. M. A. Bierma-Zeinstra, P. M. J. Welsing, and F. P. J. G. Lafber. 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, June 2012.

[80] G. E. Hinton, C. K. I. Williams, and M. D. Revow. Adaptive elastic models for hand-printed character recognition. In *Advances in Neural Information Processing Systems 2*, volume 2, 1992.

[81] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proceedings of the First International Conference on Computer Vision*. IEEE, 1987.

[82] C. Lindner, J. Thomson, and T. F. Cootes. Learning-based shape model matching: Training accurate models with minimal manual input. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *The 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2015, Part III)*, volume 9351 of *Lecture Notes in Computer Science*, pages 580–587. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015.

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

[84] T. F. Cootes, M. C. Ionita, C. Linder, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *12th European Conference on Computer Vision - ECCV 2012*, volume 7578, pages 278–291, 2012.

[85] P. A. Bromiley, J. E. Adams, and T. F. Cootes. Localisation of vertebrae on dxa images using constrained local models with random forest regression voting. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lecture Notes in Computational Vision and Biomechanics*, pages 159–171, 2015.

[86] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(4):137–154, May 2004.

[87] P. Podsiadlo, M. Wolski, G. W. Stachowiak, J. Lynch, I. Tolstykh, D. T. Felson, M. C. Nevitt, N. A. Segal, C. E. Lewis, and M. Englund. Directional fractal signature analysis of trabecular bone textue and the risk of incident radiographic knee osteoarthritis: the most study. *Osteoarthritis and Cartilage*, 21:S57–S58, April 2013.

[88] P. Podsiadlo and G. W. Stachowiak. Applications of hurst orientation transform to the characterization of surface anisotropy. *Tribology International*, 32(7):387 – 392, July 1999.

[89] The osteoarthritis initiative, 2013. URL `https://oai.epi-ucsf.org/datarelease/default.asp`.

[90] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, and T. F. Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE Transactions on Biomedical Engineering*, 32(8):1462 – 1472, August 2013.

[91] F. M. Sukno, S. Ordás, C. Butakoff, S. Cruz, and A. F. Frangi. Active shape models with invariant optimal features: Application to facial analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1105 – 1117, 2007.

[92] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape models segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924 – 933, 2002.

[93] C. Lindner, S. Thiagarajah, J. M. Wilkinson, arcOGEN Consortium, G. A. Wallis, and T. F. Cootes. Accurate bone segmentation in 2d radiographs using fully automated shape model matching based on regression-voting. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *MICCAI 2013*, volume 8150 of *LNCS*, pages 181–189. Springer, Heidelberg, 2013.

[94] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

[95] W. Boehm and A. Müller. On de casteljau's algorithm. *Computer Aided Geometric Design*, 16(7):587–605, August 1999.

[96] OpenCV. Face detection using haar cascades, December 2015.

[97] Goldberg group at NIH/NIA. wnd-charm. URL `https://github.com/wnd-charm/wnd-charm`.

[98] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70(8):920–930, August 1980.

[99] H. Tamura, S. Mori, and T. Yamavaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-8(6): 460–472, June 1978.

[100] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3 (6):269–285, November 1973.

[101] I. Gradshtein and I. Ryzhik. *Table of Integrals, Series and Products*. Elsevier: Academic Press, 7th edition, 1994.

[102] M. McMahon and J. A. Block. The risk of contralateral total knee arthroplasty after knee replacement for osteoarthritis. *Journal of Rheumatology*, 30(8):1822 – 1824, August 2003.

[103] J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213 – 220, 1968.

[104] F. Eckstein, W. Wirth, and M. C. Nevitt. Recent advances in osteoarthritis imaging - the osteoarthritis initiative. *Nature Reviews Rheumatology*, 8:622–630, October 2012.

[105] S. Muraki, H. Oka, T. Akune, A. Mabuchi, Y. En-yo, M. Yoshida, A. Saika, T. Suzuki, H. Yoshida, H. Ishibashi, S. Yamamoto, K. Nakamura, H. Kawaguchi,

and N. Yoshimura. Prevalence of radiographic knee osteoarthritis and its association with knee pain in the elderly of japanese population-based cohorts: The road study. *Osteoarthritis and Cartilage*, 17(9):1137–1143, September 2009.

[106] Osteoarthritis Initiative. Enrolment visit workbook, 2008. URL `https://oai.epi-ucsf.org/datarelease/form_workbooks/EnrollmentVisitWorkbook.pdf`.

[107] E. M. Roos and L. S. Lohmander. The knee injury and osteoarthritis outcome score (koos): From joint injury to osteoarthritis. *Health and Quality of Life Outcomes*, 1(64):1 – 8, November 2003.

[108] M. B. Kinds, P. M. J. Welsing, E. P. Vignon, J. W. J. Bijlsma, M. A. Viergever, A. C. A. Marijnissen, and F. P. J. G. Lafber. A systematic review of the association between radiographic and clinical osteoarthritis of hip and knee. *Osteoarthritis and Cartilage*, 19(7):768–778, July 2011.

[109] 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(116):1–11, 2008.

[110] M. A. Davis, W. H. Ettinger, J. M Neuhaus, and M. R. Segal. Correlates of knee pain among us adults with and without radiographic knee osteoarthritis. *Journal of Rheumatology*, 19(12):1943–1949, 1992.

[111] Osteoarthritis Initiative. Project 15 test-retest reliability of semi-quantitative readings from knee radiographs. URL `https://oai.epi-ucsf.org/datarelease/SASDocs/kXR_SQ_Rel_BU_Descrip.pdf`.

[112] J. J. Cerrolaza, A. Villanueva, F. M. Sukno, C. Butakoff, A. F. Frangi, and R. Cabeza. Full multiresolution active shape models. *Journal of Mathematical Imaging and Vision*, 44(3):463 – 479, November 2012.