FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Vier Texturalgorithmen: Bestimmung erster Anzeichen von Osteoarthrose. Daten von Multicenter Osteoarthritis Study.

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

BSc. Ștefan Ovidiu Oancea, MSc.
Matrikelnummer 01227706

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Dipl.-Ing. Dr.techn. Eduard Gröller
Mitwirkung: MSc. Renata Raidou, Dr.

Wien, 27. Juni 2018

Ștefan Ovidiu Oancea

Eduard Gröller

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Four Texture Algorithms for Recognizing Early Signs of Osteoarthritis. Data from the Multicenter Osteoarthritis Study.

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Medical Informatics

by

BSc. Ștefan Ovidiu Oancea, MSc.
Registration Number 01227706

to the Faculty of Informatics

at the TU Wien

Advisor: Dipl.-Ing. Dr.techn. Eduard Gröller
Assistance: MSc. Renata Raidou, Dr.

Vienna, 27th June, 2018

Ștefan Ovidiu Oancea | Eduard Gröller

Erklärung zur Verfassung der Arbeit

BSc. Ștefan Ovidiu Oancea, MSc.
Vienna

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 27. Juni 2018

Ștefan Ovidiu Oancea

Acknowledgements

Firstly, I would like to thank my family for their unconditional support throughout these academic years and for the motivation they gave me to leave home in the pursuit of a career.

Secondly, I would like to thank all the passionate professors at the Vienna University of Technology for making their subjects so interesting, that they easily convinced me to visit their lectures with pleasure and enthusiasm every time. Among these I would like to name the following: Prof. Andreas Holzinger, Prof. Kaniusas Eugenijus, Prof. Klaus Becker, Prof. Hans Lohninger, Prof. Bernhard Gittenberger, Prof. Eduard Gröller, Prof. Robert Sablatnig, Prof. Werner Purgathofer, Prof. Bruno Schneeweiss, Prof. Michael Reiter and Prof. Georg Langs.

Last but not least, I would like to thank my colleagues at Image Biopsy Lab (`https://imagebiopsylab.com/`) for providing materials, ideas and resources and thus making this work possible.

Multicenter Osteoarthritis Study (MOST) Funding Acknowledgment. MOST is comprised of four cooperative grants (Felson – AG18820; Torner – AG18832, Lewis – AG18947, and Nevitt – AG19069) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by MOST study investigators. This manuscript was prepared using MOST data and does not necessarily reflect the opinions or views of MOST investigators.

Kurzfassung

Diese Masterarbeit zielt ab, einen gründlichen Vergleich zwischen vier Texturalgorithmen zu liefern. Die Kapazität dieser Algorithmen zwischen Patienten mit und ohne Osteoarthrose zu unterscheiden, frühe Anzeichen der Erkrankung zu entdecken und die Entwicklung zu verfolgen anhand von nur 2D Radiographien des trabekulären Gewebes des Knies, wird untersucht. Wegen der fraktalen Eigenschaften des trabekulären Gewebes, werden zwei fraktale Algorithmen (Bone Variance Value (BVV) and Bone Score Value (BSV)) eingeführt. Diese versuchen die intrinsische 3D Komplexität des Knochens zu charakterisieren. Der dritte Algorithmus (Bone Entropy Value (BEV), basiert auf Shannon's Entropie) stammt aus der Informationstheorie und zielt ab, die Knochenstruktur in Hinsicht auf Informationskomplexität zu beschreiben. Der letzte Algorithmus (Bone Coocurrence Value (BCV)) basiert auf der Grauwertematrix eines Bildes und beschreibt die Bildtextur in Hinsicht auf spezifische Haralick Eigenschaften. Wenn diese Versuche erfolgreich wären, würden sie ein riesiges Potential besitzen die Kosten, die mit der Diagnose und Behandlung von Osteoarthrose verbunden sind, zu senken. Das würde durch die komplette Automation der Diagnoseprozedur geschehen. Die früheren Behandlungs- und Risikoverringerungsmaßnahmen sind günstiger als die Maßnahmen, die bei einem fortgeschrittenen Zustand der Erkrankung (Operation, Implantante, usw.), notwendig sind.

Zunächst wird eine Motivation zur Früherkennung von Osteoarthrose gegeben. Zweitens werden eine detaillierte Beschreibung und ein mathematischer Hintergrund der Algorithmen präsentiert und anhand von künstlichen Daten validiert. Drittens werden die für Klassifikationstests verwendeten Datensätze eingeführt. Viertens werden die verwendeten statistischen Methoden und neuronalen Netzwerkmodelle vorgestellt und diskutiert. Fünftens werden die von jedem Algorithmus erzeugten Eigenschafen (features) diskutiert und ihre unabhängige und kombinierte Fähigkeit, zwischen Knochen mit frühen Anzeichen von Osteoarthrose und gesunden Knochen zu unterscheiden, untersucht. Auch die Fähigkeit der Verfolgung der Entwicklung über die Jahre hinweg wird durch statistische Tests quantifiziert. Auch in diesem Teil präsentieren wir die besten Klassifizierungswerte (classification scores), die von den optimalen neuronalen Netzen für jeden Anwendungsfall berechnet werden. Schließlich werden Gedanken zu zukünftigen Verbesserungen und die Anwendbarkeit der Algorithmen bei anderen anatomischen Kontexten, bei anderen Krankheiten oder in anderen Bereichen, wie Histologie und Mammographie, gemacht.

Mit dieser Arbeit zeigen wir, dass der Stand der Technik in Hinsicht auf Osteoarthrosevorhersage durch die Verwendung von Modellen, die nur auf reinen Textureigenschaften basieren, übertroffen werden kann. Unsere geschlechtsstratifizierte Analyse ergibt einen Vorhersagewert von 83% für Männer und 81% für Frauen in Bezug auf Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Abstract

This master thesis aims to provide an in-depth comparison of four texture algorithms in their capacity of discriminating patients with osteoarthritis (OA) from the ones without, recognizing early signs of Osteoarthritis and tracking disease progression from 2D radiographs of the knee trabecular bone (TB). Given the fractal properties of the trabecular bone (TB), two fractal-based algorithms (Bone Variance Value (BVV) and Bone Score Value (BSV)) that try to characterize the complexity of the underlying 3D structure of the bone are presented. The third algorithm (Bone Entropy Value (BEV), based on Shannon's Entropy) stems from the information theory and aims to describe the bone structure in terms of information complexity. The last algorithm (Bone Coocurrence Value (BCV)) is based on the co-occurrence matrix of an image and describes the image texture in terms of certain Haralick features. If successful, such algorithms posses a great potential to lower the costs (financial, time) associated with the diagnosis of osteoarthritis (OA) through automation of the procedure, and with the treatment. The earlier treatments and risk reduction measures are less costly than the procedures involved due to a more advanced stage of the disease (surgery, implants, etc.).

First, a motivation for the detection of early osteoarthritis (OA) is given. Second, a detailed description and mathematical background of the algorithms are presented and validated on sample, artificial data. Third, the employed data sets used for classification tests are introduced. Fourth, the statistical methods and neural network models employed are presented and discussed. Fifth, the features produced by each algorithm are discussed and their independent and combined capacity of discriminating between bones with early signs of OA and healthy bones. Also the capacity of tracking OA progression through the years is quantified by statistical tests. Also in this part we present the best classification scores obtained from the most optimal neural networks for each use case. Finally, thoughts on future improvements and the generalization of the algorithms in other anatomical contexts, for other diseases or in other fields, like histology and mammography, are made.

In this work we show that the state-of-the-art in OA prediction can be surpassed by utilizing only models based on texture features alone. Our gender-stratified analysis produces a prediction score of 83% for males and 81% for females in terms of Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Contents

CHAPTER 1

Introduction

Arthritis is the leading cause of morbidity and disability in developing and developed countries [1]. Arthritis denotes acute or chronic joint inflammation possibly due to genetics, infections, abnormal mineral crystals deposition, and injuries [2]. Treating different forms of arthritis involves high socioeconomic costs. For example, in 2014 arthritis costed the United States $802 billion compared to $182 in 1996, which represents almost 5% of their GDP (c.f. Figure 1.1). osteoarthritis (OA) is by far the most common form of arthritis which causes disabilty and reduction of quality of life [3].

1.1 Problem Statement

While the exact causes and the initiation and evolution of biological processes of the disease are still a matter of debate ([4]), it is known that the increasing obesity and age in a population, unequal feet length and jobs that involve high levels of joint stress, are directly linked to a massive rise in OA-related costs and morbidity [1, 5]. This leads to the conclusion that generally OA begins with cartilage loss due to abnormal loads. OA can generally affect any joint, but the incidence at the weight-bearing joints (the hips and the knees) is most common. Even though the prevalence of OA is higher in the 21st centry due to older population and increasing obesity, the disease is not a modern burden of the society. Given that the bone tissue is not easily degradable over time, OA is one of the best documented diseases. Signs of OA in animal bones have been traced to as early as 100 million years in the past in the fossils of two dinosaurs. The researchers have concluded that the general characteristics of the disease have consequently remained unchanged even though the hosts have evolved [6]. There are studies that argue that OA *'appears to be a solid immutable part of life which is oblivious to all evolution'* [7, p. 173].

OA is characterized by cartilage loss, subchondral bone changes, synovial inflammation and meniscus degeneration [8]. The measures taken against end-stage OA mostly involve full joint replacement. Non-surgical treatments only show limited success due to their

delayed start when the joint structure degeneration is advanced. OA patients only begin to experience symptoms like i.e. joint pain and stiffness at a late stage of OA when curative or palliative treatments are inefficient. This is due to the fact that the cartilage tissue does not have any nerves that could sense mechanic modifications as opposed to the bone that reacts to damage. The patient experiences symptoms only when the damage extends deeper into the bone. At this stage, a high volume of cartilage had already degenerated. OA typically becomes clinical/radiographic OA (bone cysts formation, osteophyte formation, nonuniform joint space loss, and subchondral sclerosis) only many years after its onset. It is thought that this long window of time could be used to alter its course if early changes could be successfully detected [1]. Thus, common palliative treatments of the late disease that are used today could be replaced by prevention treatments that reduce OA risk overall.

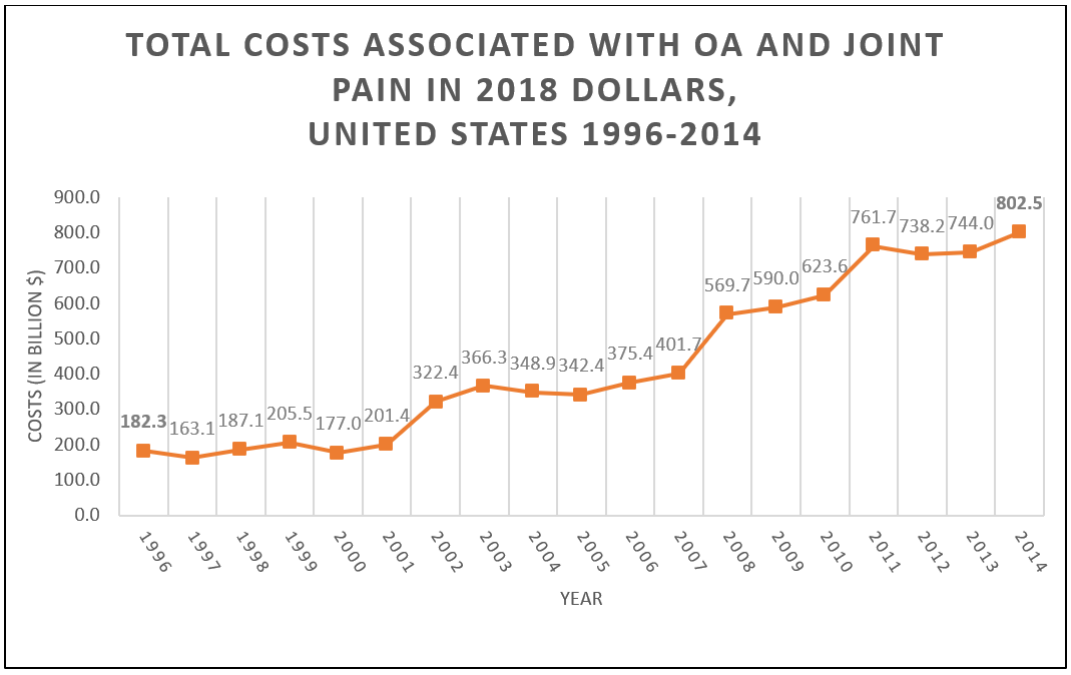


Figure 1.1: Arthritis costs in the U.S. between 1996-2014. Source: [9].

Early OA detection is thus important and no accurate, reliable solution is known. There is no widely-accepted definition for early OA. Histologically, early OA is defined by the Osteoarthritis Research Society International (OARSI) scoring system as having a grade between 1.0 - 3.0. This represents the depth of degradation into articular cartilage [10]. However, studies have shown that early signs of OA can be seen in the trabecular bone (TB) of the knee, years before any cartilage degradation is detectable [11, 12]. In fact, the subchondral bone as a whole appears to play an important role in OA development, especially in the initiation phase. Bone remodelling is found to occur at this site particularly in early OA [13]. This change immediately leads to a reduced capability

of the subchondral plate to absorb and dissipate energy [14], which in turn leads to increased forces that act on the remaining elements of the joint: articular cartilage, tendons, ligaments, menisci, and bursae. In other words, the other joint structures must compensate for the loss of resistance of the bone.

OA is routinely assessed by imaging tests possibly in combination with lab tests [15]. X-ray images show a reduction in joint space, which translates to loss of cartilage as this softer tissue is not directly visible by this modality. Bone spurs at the bone extremities are also discovered in radiographs. Even at this stage, patients do not always experience characteristic symptoms like pain, joint stiffness, swelling, redness, and reduction of motion range. MRI is used to also image the cartilage and other softer tissues of the joint in more complex cases. Some earlier abnormalities of OA can also be detected by this modality. Joint fluid aspiration is sometimes performed at the presumably affected joint by which the doctor can determine if there is an inflammation and if the discomfort is caused by gout or an infection. Blood tests can be employed in addition to the methods described above. Even though there exist no blood test specifically engineered to detect OA, certain diseases with specific blood markers can be excluded from the doctors' hypotheses.

1.2 Aim of the Work

Since the currently standardized methods described above for detecting OA are not able to assess early changes in the subchondral bone, and given the reduced costs and popularity of 2D radiographs, we attempt to implement and test four texture algorithms that could be suitable for a detection of early OA signs in the knee TB using only X-ray images. Given the fractal properties of the TB [16], two fractal-based algorithms that attempt to characterize the complexity of the underlying 3D structure of the bone are presented [17, 18]. The third algorithm, based on Shannon's Entropy, stems from information theory and aims to describe the bone structure in terms of information complexity. The last algorithm is based on the co-occurrence matrix of an image and describes the texture in terms of Haralick features [19].

The algorithms will be tested on two data sets with different acquisition parameters (pixel spacing, exposure, machine manufacturer) to find their independent capacity of discriminating between OA and non-OA patients. One of the two datasets presents a longitudinal study, which enables us to investigate whether the algorithms are also able to predict OA. The influence of the said imaging parameters and also of other confounding variables such as sex, age, and body mass index (BMI) is also tested at this step and eliminated from the features if it is too large. We want to produce features that only characterize the texture and do not hold any other intrinsic information. Each algorithm computes different texture features that will be investigated regarding their discrimination and prediction power. We want to build Artificial Intelligence (AI) models based on the best-scoring features to show that simple models can be obtained to deal with future data. Moreover, the possibility to track subchondral bone remodeling over

the years is also investigated within the longitudinal study. Provided that this is possible, the result will be a proof that the detection of OA in its early stages may be possible.

1.3 State of the Art

An attempt from literature to predict the onset of knee OA was made by Janvier et al. [20]. Using a logistic regression model with texture parameters and BMI, age and gender combined, they are able to predict the Kellgren-Lawrence grade (KL) scores (i.e. the OA incidence) at a 48-month follow up with an accuracy of 69% in terms of Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

In another attempt, Kraus et al. obtained a classification score of 79% in terms of Area Under the Receiver Operating Characteristic Curve (ROC-AUC) [21]. In their model they combined fractal signature features (texture features) at baseline, knee alignment, traditional covariates, and bone mineral content to predict the Joint Space Narrowing (JSN) grade over a three-year period.

Woloszynski et al. use roughness, degree of anisotropy, and direction of anisotropy based on a signature dissimilarity measure method [22]. They manage to predict the JSN grade with a score of 75% in terms of ROC-AUC over a 3-year period.

In this work we attempt to surpass the scores obtained in previous publications. To achieve this, we enhance our models by taking into consideration a longitudinal study with three visits.

1.4 Methodological Approach

First, background information about OA progression (from a biological point of view) and diagnosis is gathered that serves a better understanding of the phenomenon and implicitly of the need for early detection. A suitable programming language is also chosen at the beginning depending on the available libraries that aid the implementation of the algorithms and libraries. They should offer proper validation and result quantification tools for unsupervised/supervised learning and model building, feature selection and statistical tests.

Second, suitable data sets are chosen. Sets of images with different imaging parameters and balanced instances are selected.

Third, the algorithms that are not already available (BEV, BCV) are investigated, implemented and validated using test images with known theoretical parameters. The BVV algorithm has been previously introduced and presented [23] and BSV is an algorithm currently used in our company (Image Biopsy Lab G.m.b.H. (IB Lab)) in different research areas. The last two approaches are also validated using artificially generated fractal images (isotropic and an-isotropic) with known fractal dimensions.

Fourth, features for regions of interest (ROIs) of the selected data sets are extracted using the proposed algorithms and a feature selection is employed to detect the significance of each feature for each data set. These features will then be used not only to test the discrimination power of the algorithms between OA and non-OA subjects with statistical tests and neural models, but also to inspect the capability of tracking the subchondral bone remodeling that takes place during OA progression.

1.5 Structure of the Work

The following Chapter 2 discusses related approaches that generally attempt to assess early OA. Not only related texture algorithms are presented, but completely novel ideas like for example blood serum tests are also presented and shortly discussed.

Chapter 3 introduces the reader to the anatomical basics of the human (long) bones by analyzing their whole complex microarchitecture. Also possible causes and general pathophysiology of OA are presented.

In Chapter 4 the chosen data sets and their properties are introduced. Depending on the data available from each set, the experiment tasks with their corresponding research hypotheses are defined. Based on these tasks, selection criteria are introduced to build groups of patients suitable for the testing of the defined hypotheses.

Chapter 5 describes all the used algorithms in detail from a mathematical point of view. The discussed algorithms will also be validated in Chapter 6 by using *synthetic* surfaces with known theoretical parameters. The calculated parameters by the algorithms will be compared with the theoretical ones to investigate how well the algorithms approximate the theoretical parameters of the artificial surfaces.

In Chapter 7 all the statistical models that are used for feature interpretation and model building are described in detail. From statistical tests that assess differences between samples of data to Random Forests to simple linear Support Vector Machines (SVMs), all the used methods that are employed to detect the most significant features capable of separating OA from non-OA, of predicting OA, or of tracking the changes due to the disease, are presented and analyzed in detail.

In Chapter 9 the results obtained are presented and the influence of the machine parameters are investigated and if needed removed. The most significant features per task will be discussed and the classification scores interpreted.

Chapter 10 presents a summary of the presented work. The most important findings and their meaning are reiterated.

The last Chapter 11 describes ideas of how the present work can be not only improved but also extended. In other words, we shortly discuss the applicability of the investigated algorithms to other use cases.

CHAPTER 2

Related Work

Before we describe our attempts at detecting OA we first take a look at other approaches in the literature that pursue the same goal. The following methods cover a wide spectrum of different approaches that tackle the problem of early detection of OA. We shortly introduce works that employ other texture-based methods to achieve the same result. We also present completely other approaches based on different imaging modalities, such as for example MRI, that are used to find early micro lesions in the knee joint. Also, we show how plasma and synovial fluid analysis is used for OA-specific marker detection. In other words, the scientists employ a variety of methods and techniques that at some point could become a standard in the efficient prevention of OA.

2.1 Lesion Detection from MRI Data

Sharma et al. attempt to isolate different bone lesions, such as cartilage damage, bone marrow lesions, and meniscal damage, in patients that at the point of the experiment had not shown classic signs and symptomps of radiographic OA, i.e. a KL of 0 in both knees [24]. However, only samples of subjects at high risk of developing OA were selected for this study. The lesions were assessed using the MRI OA knee score (MOAKS) by experts on MRI volumes [25]. Prevalent frequent knee symptoms, incident persistent symptoms, and incident cartilage damage were also assessed from 12-month, 48-month, and 60-month follow-ups. The detected lesions were then found to be strongly correlated (p-values < 0.005) with the symptomatic outcomes. This leads to the conclusion that the lesions are not 'incidental and may represent early disease in persons at increased risk of knee OA' [24, p. 1811].

2.2 Synovial Fluid and Blood Serum Tests

In a completely other kind of approach Ahmed et al. attempt to detect early signs of OA by analysing plasma and the synovial fluid of patients [26]. They detected specific oxidized, nitrated, and glycated proteins and aminoacids that form due to the damaged articular tissues using mass spectrography. Subsequently, two algorithms were used in turn on the obtained experimental data. First, a discrimination between healthy and diseased subjects was made and second, the ill subjects were classified as OA, rheumathoid arthritis (RA), or non-RA. The early-stage OA sensitivity/specificity of the detection was 0.92/0.90, which increased even further in severe and advanced OA and RA cases.

2.3 Whole-Joint Analysis from 2D Radiographs

In [27] Shamir et al. use plain 2D radiographs to extract ROIs automatically. They achieve this by scaling down 20 pre-marked knee ROIs by a factor of 10. In a new image, the knee is detected also by first scaling down the whole image by a factor of 10 and then comparing 15 x 15 shifted windows in turn to each of the 20 initial ROIs in terms of Euclidean distance. Finally the shortest distance is picked and the corresponding downscaled window indicates the position of the knee in the initial, unscaled space. The detected ROIs are then transformed into six other spaces: Wavelet, Fourier, Chebyshev transforms alone, and combinations thereof, such as i.e. Wavelet after Fourier and so on. Additionally, Zernike features, multi-scale histograms, moments of mean, skewness and kurtosis, Tamura texture features, Haralick features, and Chebyshev statistics are computed. In total 1470 image descriptors were obtained. Using this method, moderate OA can be differentiated from controls with 91.5% accuracy.

2.4 Gabor Filters and GLRLM Features

Boniatis et al. employ Gabor filters in combination with grey level run length matrix (GLRLM) features to detect organized structures in the ROIs extracted from hip bones [28]. They propose a two-step classification: first, OA and non-OA hips are separated and second, the OA hips are classified regarding the severity of the disease. They reach a discrimination accuracy between mild/moderate and severe osteoarthritic hips of 95.7%, which suggests that their methods could be sensitive even to smaller structural changes.

CHAPTER 3

Osteoarthritis (OA) Background

We motivate our determination to pursue the development of an automated decision support system for OA prediction and detection by looking at the anatomy of human long bones and by analysing the pathophysiology of OA. Understanding the micro architecture of bones is crucial to discover causes for bone failure and degeneration. At the same time, the understanding of the gradual progression of OA is key to identifying useful features that can be extracted from the radiographs and used for training of an AI model that could serve as an automated early prediction system.

3.1 Human Long Bones

Humans are born with around 305 bones in the body that confer six major functions: posture, mobility, protection, blood cells production, storage of minerals, and regulation of some endocrine systems. This number is later reduced to around 205 in adult humans due to the fusion of some parts during growth [29]. The peak bone density is reached at around the age of 21 years in a healthy individual. One can differentiate five types of bones in the human skeleton: long, such as femur and tibia, short, such as the finger bones, flat such as the scapula and the sternum, irregular, such as the vertebral bones and sesamoid bones that act as interfaces for tendons to transfer muscular forces more efficiently [30]).

While any joint in the body can can be affected by OA, the focus of this thesis are the long bones that form the knee articulation and that support the highest loads and thus are more prone to failure. Bones are extremely complex organs whose mechanical properties can not be fully understood if one does not consider their whole hierarchical structure. Up to seven architecture levels can be distinguished in human bones generally. We will now shortly consider each level [31].

3.1.1 **Level 1 -** Whole Bone

In the meter range, the long bones consist of two epiphyses (ends or 'heads') and a diaphysis that connects them (as seen in Figure 3.1). The epiphyses are coated with so-called hyaline (articular) cartilage. This cartilage sits in synovial fluid secreted by the synovial membrane which, serves for lubrication and as a source of nutrients for the joint elements. The role of the cartilaginous space is to transmit joint loads from one bone to another one efficiently, with a very low friction coefficient of around 0.001 [32]. In comparison, in case of a hip replacement for example, a metal-plastic combination achieves a friction coefficient of only 0.04 in best case scenarios [33].

Human bones are vascularized and innervated. The arterioles bring nutrients into the tissue, while nerves sense damage and are the pathways for the initiation of bone remodelling by communicating with the brain. The human bones are metabolically active tissues. They have the capacities to adapt their structure to loads over time and to repair themselves in case of aging or damage [34].

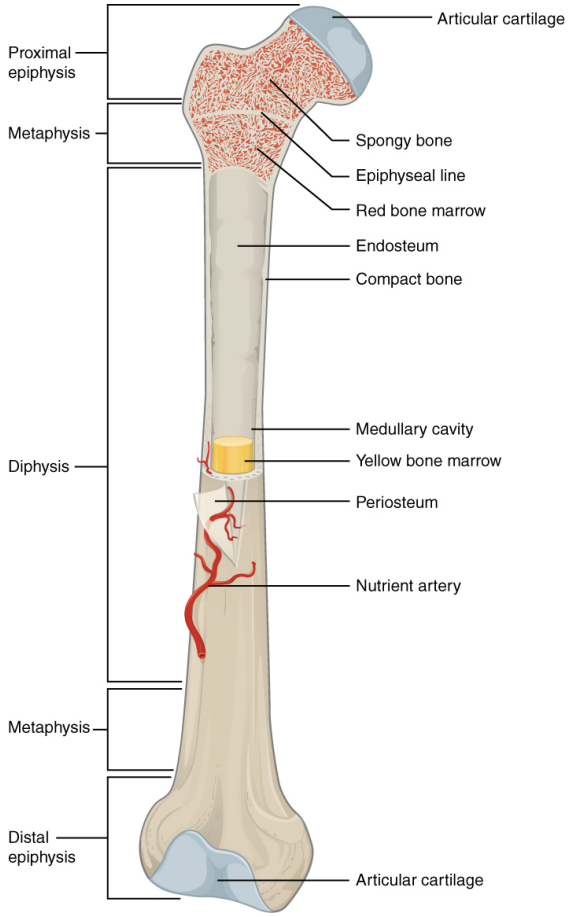


Figure 3.1: Example of human long Bone: tibia [35].

3.1.2 **Level 2 -** Cortical and TB

In the centimeter-micrometer range one can distinguish between two types of bone mass arrangements: cortical and TB. The cortical bone coats the entire bone providing shape, stability, and fracture resistance, while the TB is generally situated at the ends of long bones (some irregular bones make exceptions) as seen in Figure 3.2.

The bulky epiphyses that we introduced in Section 3.1.1 can manage stress efficiently through their larger surface area. The TB which lies inside the epiphyses improves this further through their special 'spongious' arrangement that facilitates stress dissipation. In this manner the joint loads are efficiently transfered to the midshaft of long bones and to the next joint while avoiding high stress concentrations (as shown in Figure 3.3).

The cortical bone and the TB are principally distinguished by their porosity, bone-volume-to-total-volume ratio (relative volume), and by the internal surface area (Figure 3.1).

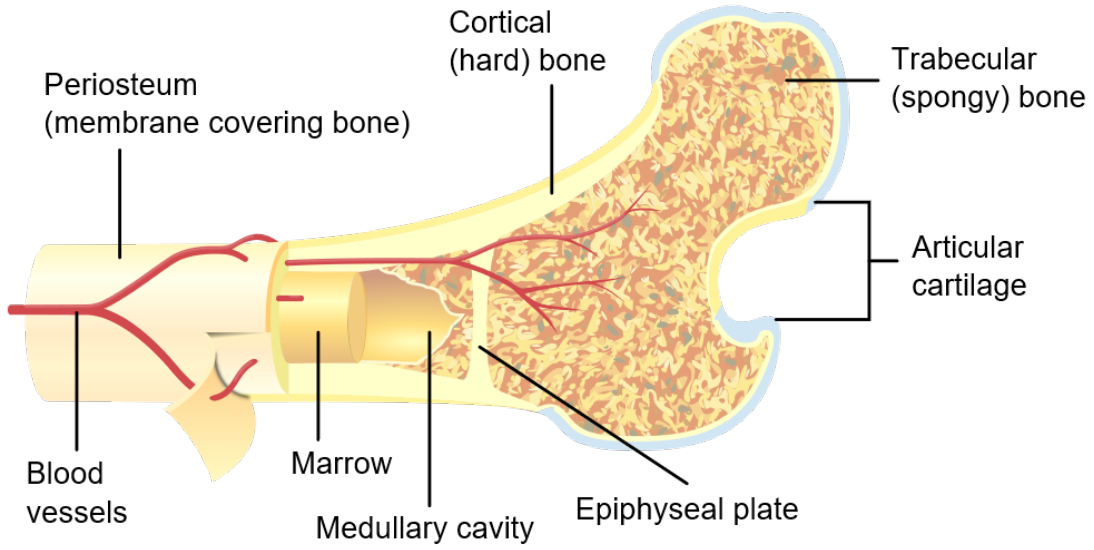


Figure 3.2: Cross-section of a long bone [36].

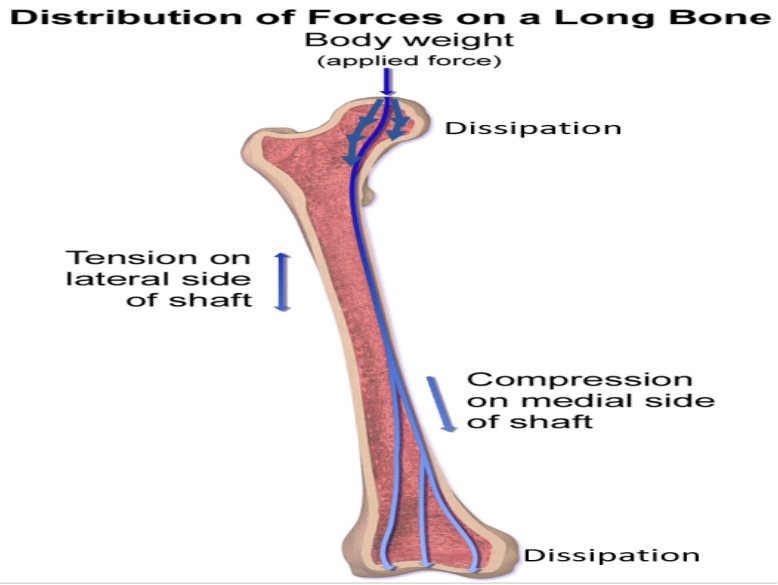


Figure 3.3: Stress dissipation in long bones. Reproduced with permission [37].

Table 3.1: Trabecular and cortical bone morphology [31].

Tissue	Compact	Trabecular
Relative Volume	80%	20%
Relative metabolic activity	50%	50%
Porosity	5 – 10%	55 – 95 %
Total volume	4.2 dm^3	1.05 dm^3
Internal surface	3.5 m^2	7.0 m^2
Density	2.0 g/cm^3	2.0 g/cm^3

3.1.3 **Level 3 -** Cortical and TB Microstructure

In the next level we can find the osteons, the trabecular packets, and the lamellae (as shown in Figure 3.4). The osteons are cylindrical structures composed of concentric lamellae. Inside the osteons, blood, lymphatic vessels, and nerve axons reside. Inbetween lamellae, lacunae filled with osteocytes can be found. Cement lines border osteons and contain minerals and non-collagenous proteins (NCPs), but less than the osteons [38].

The trabecular packets can be seen in Figure 3.4b. They are also composed of lamellae

bordered by cement lines. A mineralisation gradient can be observed in this case: the core of the trabecula is denser while its shell is thinner. The reason for this is unknown at the point of writing this work.

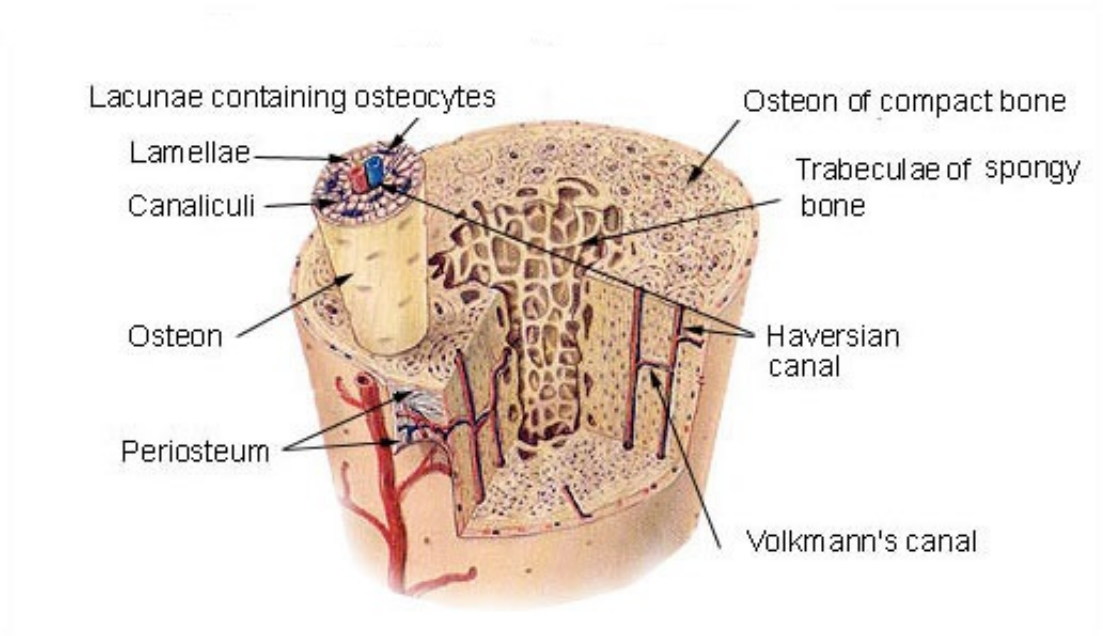
As mentioned before, the bone is a metabolically active tissue. The centers of metabolism in bones are the different bone cells (osteoblasts, osteocytes and osteoclasts) and are also found at this level in the hierarchy.

Osteoblasts are created after mesenchymal stem cells differentiate into them. The role of osteoblasts is to synthesize the organic components needed for the bone matrix. The so called 'bone lining cells' that cover the entire bone matrix are also osteoblasts, but in a dormant state, ready to generate new bone matrix if required.

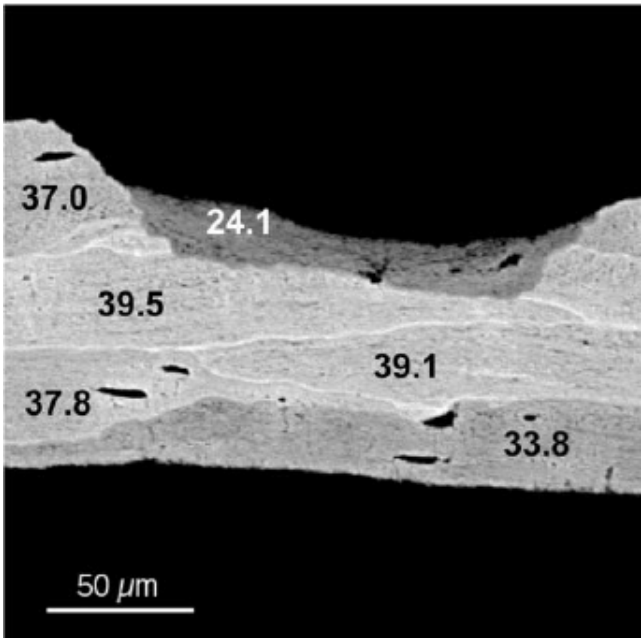
Once the osteoblasts become completely trapped by their products, they turn into osteocytes and serve as mechanosensors to detect damage and to initiate remodelling. It is unclear how these cells are capable to sense damage, but two hypotheses exist. The damage is thought to be sensed based on:

1. cilia found on the cell membrane, which increase fluid flow due to microcracks.
2. strain amplification on the surface of the cell due to cracks and other voids caused by them.

The osteocytes also detect the absence of cracks or loading and as a consequence will reduce the amount of bone in that specific region. The osteoclasts are derived from hematopoietic stem cells and are large and multinucleated. They secret lysosomal enzymes that resorb bone tissue. If the osteocytes detect damage in the bone, they initiate the so-called *'ARF Sequence'*: Activation, Resorption, and Formation. First, the osteocytes send a signal that activates the creation of osteoclasts. The osteoclasts resorb the affected tissue and consequently, the 'dormant' bone lining osteoblasts begin to reconstruct the missing bone matrix, cementing themselves in the process and becoming new osteocytes (as shown in Figure 3.5).



(a) Cortical bone osteons [39].



(b) Trabecular packet with lamellae and cement lines. The numbers represent the average mineral content. Lower values indicate newer formation [40].

Figure 3.4: Bone micro structure.

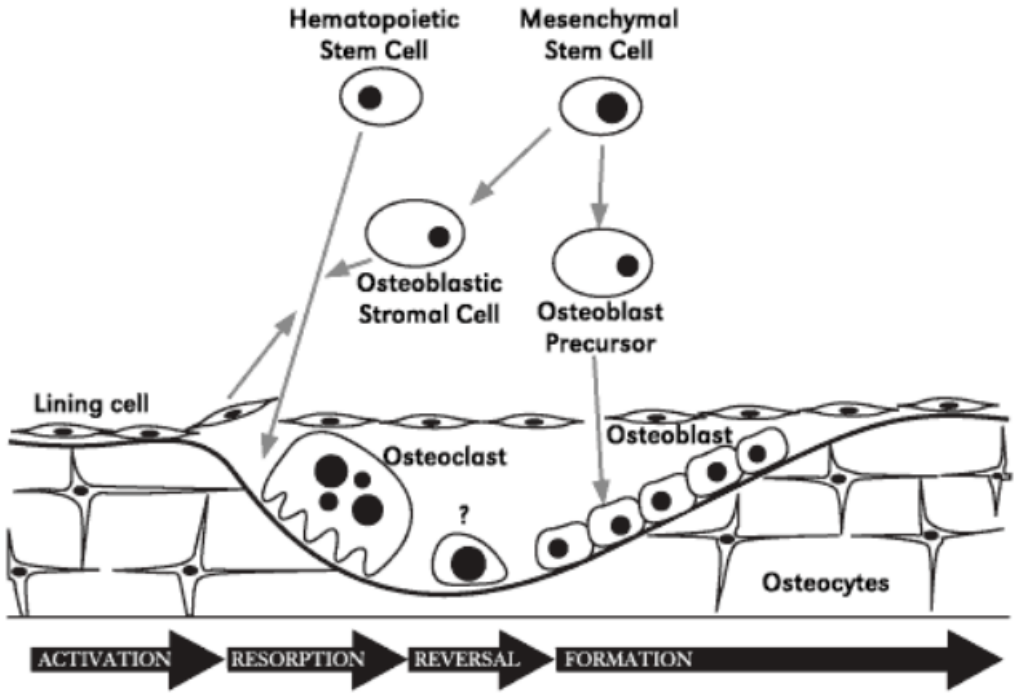


Figure 3.5: Bone repair ARF Sequence: Activation, Resorption and Formation [31].

3.1.4 **Level 4 -** Lamellar/Sublamellar Bone

The osteons are composed of bundles of collagen fibrils disposed differently (as seen in Figure 3.6). Depending on the loads on the bone, different spatial arrangements of collagen fibrils confer protection against strains. For example the twisted plywood osteon is efficient at resisting shear strains and not only transverse or longitudinal ones like the orthogonal osteons. The morphology of these osteons depend on the site and function of the specific bone they are part of.

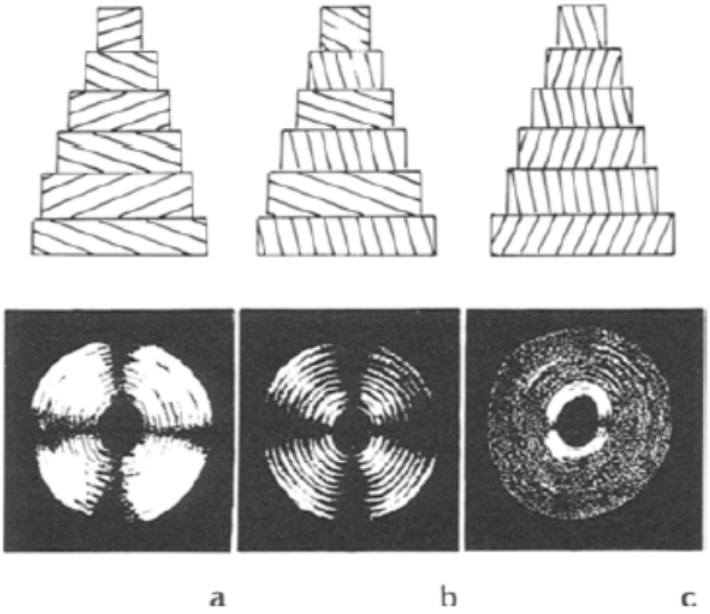


Figure 3.6: Osteon types. (a) Orthogonal. (b) Twisted plywood. (c) Plywood. [41].

3.1.5 **Level 5,6,7 -** Collagen Fibrils, Minerals, NCPs

Bone tissue is composed of organic elements, inorganic elements, and water. Among the organic materials one finds collagen type I, NCPs, proteoglycans and lipids, and other forms of collagen, whose presence facilitate the formation of functional collagen type I bonds. On the other hand, the inorganic materials (minerals) are mainly the hydroxiapatite crystals and magnesium, sodium, potassium and other.

Generally, the collagen takes the form of small, cylindrical fibers, called fibrils, due to self-assembly processes that involve ionic, Van-der-Waals, and hydrogen interactions. They can reach 500 nanometers in diameter and tens of micrometers in length. They form bundles and in order to be able to fill the whole space, they also assume different diameters.

The collagen fibrils are embedded in minerals, which are held in place by NCPs (see

Figure 3.7). NCPs are also found in the interosteonal or interlamellar cement lines. The minerals and NCPs are important micro-mechanical modulators that can deflect cracks in healthy bones (as seen in Figure 3.8).

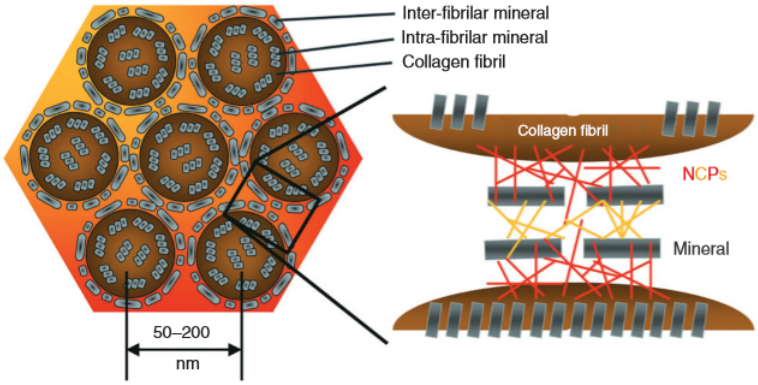


Figure 3.7: Fracture surface of trabecular bovine bone exhibiting collagen fibrils coated with minerals. Minerals are also found intra-fibrillar [42].

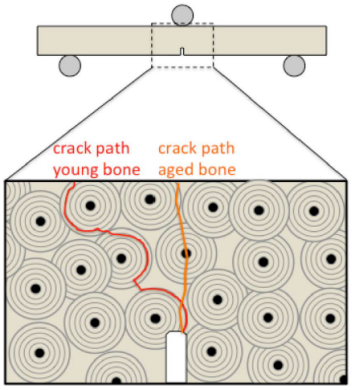


Figure 3.8: Crack propagation paths differ between young and elderly bones [43].

3.2 OA Pathophysiology

In Section 3.1 we presented the basics of human bone anatomy, which are needed to understand how OA develops and progresses. The pathophysiology of OA in general is discussed in this section.

The bone-cartilage-synovial fluid-cartilage-bone complex can be regarded as a continuum, whitout which movement would not be possible. Its structures however are organized differently depending on the body location and function. Since the focus of this work

is the knee OA, we will look closely at how OA develops in the knee joint. Six different basic structures can be observed within this joint [44] (as seen in Figure 3.9):

1. ligaments — passive elastic structures that resist tension
2. musculotendinous units — active elastic structures that act under tension
3. cartilage — passive elastic structure
4. subchondral bone — passive structure that together with the cartilage supports the compressive loads in the joint
5. medial and lateral menisci — passive fibrocartilaginous structures that resist tension and torsion
6. bursae — passive structures filled with synovial fluid that act as a buffer between tendons and bones or between muscles and bones to reduce friction and ensure free movement.

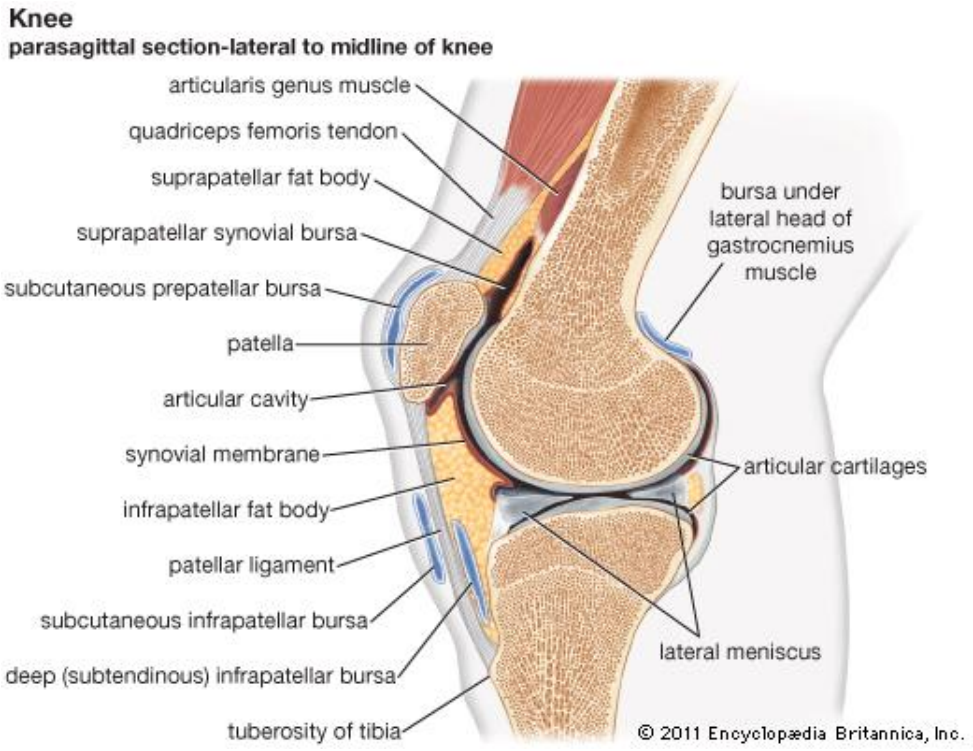


Figure 3.9: Knee parasagittal section [45].

3.2.1 Cartilage Degradation

Osteoarthritis is a disease that affects all joint structures to some extent, but mainly the cartilage [46]. The articular cartilage degenerates gradually, while subchondral bone sclerosis, osteophytes, and synovial inflammation will most certainly occur (as seen in Figure 3.10). A hypothetical model for initiation and perpetuation of OA can be seen in Figure 3.11. However, all models agree on the fact that there is a combination of risk factors and ageing that lead to the initiation of OA. While OA is not only a degenerative disease that occurs as a result of gradual wear and tear, one can differentiate two mechanisms that lead to abnormal remodelling of joint structures and finally to OA [44].

The first mechanism involves normal loads on abnormal cartilage. A cartilage can become 'abnormal' due to ageing or injuries. At the same time genetic factors can play an important role in disrupting chondrocyte (cartilage cell) differentiation and thus be responsible for abnormal mechanics [8]. This mechanism is usually the cause of OA in younger people due to repeated joint traumas.

The second mechanism involves abnormal loads on normal cartilage. This is the case of subjects with high BMI or different skeleton deformities (varus and valgus).

The two mechanisms can be differentiated however only in the early phases of OA. As OA progresses a combination of the two mechanisms appears that leads to a massive degeneration of the joint. For example the process can begin due to processes of the second mechanism, meaning that the subject has a high BMI, but healthy cartilage. Through the years, as the high loads affect the joint, the cartilage can become abnormal and the disease advances much faster from this point.

In healthy patients the homeostasis of the articular cartilage is controlled by chondroblasts and chondrocytes. The chondroblasts secrete the structural matrix that consists of collagen and proteoglycans. At some point they surround themselves completely with matrix and become trapped in lacunae. From this point they are called chondrocytes and can not migrate anymore. Since the cartilage is not supplied with blood, the nutrition is completely dependant on diffusion processes through the matrix. This is a slow process meaning that damaged cartilage has reduced healing capabilities as opposed to the bone tissue [30].

OA has been characterized in many studies by a disturbed/delayed repair process of damaged cartilage due to biochemical and biomechanical changes in the joint [47]. In patients with OA the chondrocytes can not synthesise as much matrix as is destroyed [48]. This eventually leads to a complete loss of articular cartilage and because it is aneural, the patient does not experience any symptoms related to its absence until the damage extends deeper into the subchondral bone.

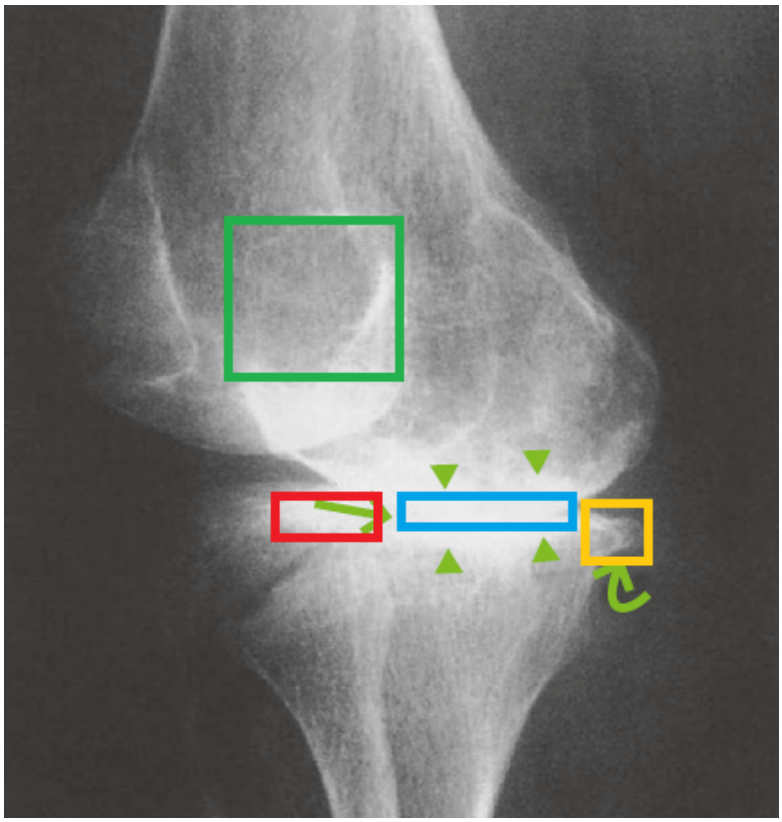


Figure 3.10: Radiographic manifestations of OA. Joint space narrowing (blue), osteophytes (yellow), bone cysts (green) and sclerosis (red) visible [44].

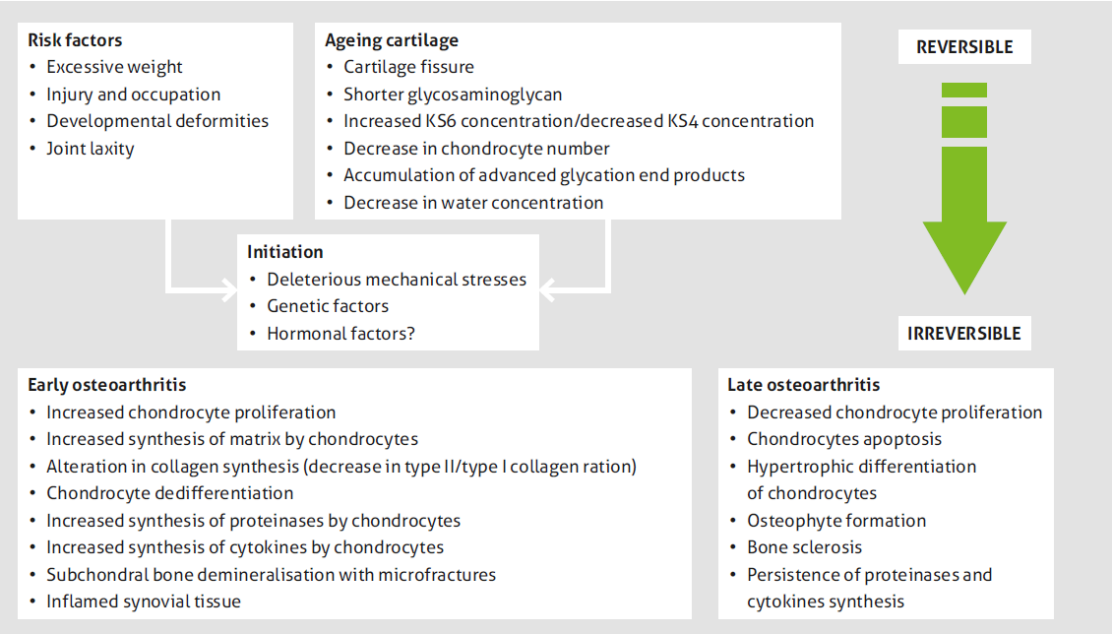


Figure 3.11: OA initiation and perpetuation hypothetical model. *KS* stands for keratan sulphate [44].

3.2.2 Subchondral Bone Changes

Distal to the articular cartilage lies the subchondral bone. This is composed of epiphyseal cortical bone and spongy bone, i.e. the TB. The subchondral bone provides support, absorbs shocks, and supplies the joint with nutrients. The surface of the subchondral cortical bone is less stiff than the diaphyseal cortical bone for more efficient nutrient transport out of the bone matrix [44]. It is not yet known with certainty if subchondral bone is altered before the most popular sign of OA, which is cartilage loss, but recent animal studies suggest that a certain degree of microstructural reformation is possible years before cartilage damage occurs [13]. The human bones were shown to indicate similar changes prior to the actual cartilage loss [11, 49].

Early OA sees an increase in subchondral bone remodelling rate and porosity for yet unknown reasons, leading to reduced thickness of the subchondral plate [13]. However some theories for this exist:

1. Interleukin 1 and 6 have been detected in abnormal amounts in deteriorating cartilage [13]. They are inflammation mediators and at the same time stimulants of bone remodeling.

2. It is known that early OA produces changes in the microarchitecture of the capillaries within the subchondral bone [50]. This vascular invasion can penetrate deep into the articular cartilage producing catabolic enzymes, which degenerate the cartilage. This produces a feedback loop with the subchondral bone, which needs to adapt in order to support the same loads without the same volume of cartilage present as before [13].

3. Burr et al. showed [13] by in vitro studies that there is a cross-talk between bone cells and cartilage cells due to micro-cracks in the subchondral plate. This false signaling may lead to increased bone resorption and consequently remodeling.

With the progression of the disease, the remodelling rate decreases, but overall there is an imbalance of resorption and bone formation leading to a net increase in bone volume and density [51]. This phenomenon is known as bone sclerosis and is detected as a condensation in radiographies (as seen in Figure 3.10) due to thicker bones. It is thought that the newly-formed bone is less mineralized, which leads to a reduced mechanical stiffness and consequently to the deterioration of the articular cartilage [13].

In the early stage of OA, osteophytes can also develop, which are outgrowths of osseus tissue covered in cartilage. They usually form at the docking places of tendons or ligaments (traction spurs) and their role is not fully understood. However there are studies that have found that limb osteophytes may be helpful in the stabilisation of the joint [52]. There exist also inflammation spurs, which usually occur between the vertebrae, but are also painful.

3.3 Radiographic Imaging

As stated in Chapter 1, OA is routinely assessed based on X-ray images. This technique 'remains the most accessible tool in the evaluation of the OA joint' [53, p. 1]. In radiography, electromagnetic radiation, also called Röntgen radiation, with wavelengths that range from 0.01 to 10 nanometers, are projected towards an object from an X-ray generator. The generator consists of a cathode, an anode, and a vacuum tube. The cathode directs a high-speed stream of electrons towards the anode, which is made out of tungsten for better heat dissipation. After the collision, 1% of the energy is released as X-rays and 99% as heat. These resulted rays are then projected through an object in order to visualize its internal structure. This is achieved due to the fact that the target object absorbs a certain amount of the X-rays depending on its composition. The energy that passes through is measured with a detector at the other end, which is a photographic film or a digital detector.

For the assessment of OA, the knee joint is generally imaged in an extended state, with joint-perpendicular X-rays and with the patient standing exerting full pressure on the joint (also called *weight-bearing* position). However, in order to improve intra-articular visualization, the knee joint can be flexed by various degrees. At the same time the X-rays' projection angle can be adapted. The grading of OA is done by evaluating osteophyte formation, JSN and other factors using predefined grading schemes such as KL [54] or OARSI [55]. An example of an X-ray image from a patient with osteophytes and reduced JSN is shown in Figure 3.12.

Figure 3.12: A) Antero-posterior weight-bearing radiographs of a patient with JSN and osteophyte formation consistent with bilateral medial osteoarthritis of the knee. B) A magnified view of the right knee joint. The arrow denotes medial JSN. Osteophyte formation can be seen on the femur and tibia [53].

CHAPTER 4

Data Sets and Hypotheses Definition

In this Chapter we introduce the real image data sets that we use to test our texture algorithms presented in Chapter 5. At this point we will also provide details on the planned tasks that should be solved by the AI models defined in Chapter 8. The tasks are formulated based on the structure of the data, which is specific to each data set. One of the data sets is a longitudinal study and therefore enables us to formulate research questions related to the development of the disease. Within each task we formulate appropriate study hypotheses, which we test for validity with statistical methods that are presented in Chapter 7.

In Chapter 3 we have shown that the subchondral regions of the tibia show structural changes during the OA development. Therefore, we assume that these regions may produce significant features that could be used for prediction and detection of OA. In a previous study this theory was confirmed and it was shown that the regions below the tibia plateau produced the most significant features for OA prediction [56]. The selection of ROIs from the X-Ray that will be used for feature extraction and analysis is done automatically using the IB Lab Analyzer Software developed at IB Lab (`https://imagebiopsylab.com`). The ROIs are positioned relative to the two landmarks *MED* (marking the medial tibial bone condyle) and *LAT* (marking the lateral tibial bone condyle, i.e. the side where the fibula is located) that are detected by deep learning means. An example of landmark placement is shown in Figure 4.1. This relative positioning of the ROIs can be viewed as a registration technique to assure that the ROIs are placed at roughly the same positions for the same patient across several visits. The software provides the possibility to automatically assess the OA grading based on the KL score by evaluating the joint space width, the presence of osteophytes and sclerosis. The marked ROIs can be used for in-place feature engineering using some built-in algorithms (currently BSV and BEV

are implemented) or can be extracted with a view of being assessed externally by other algorithms as well, which is the case of BCV and BVV.

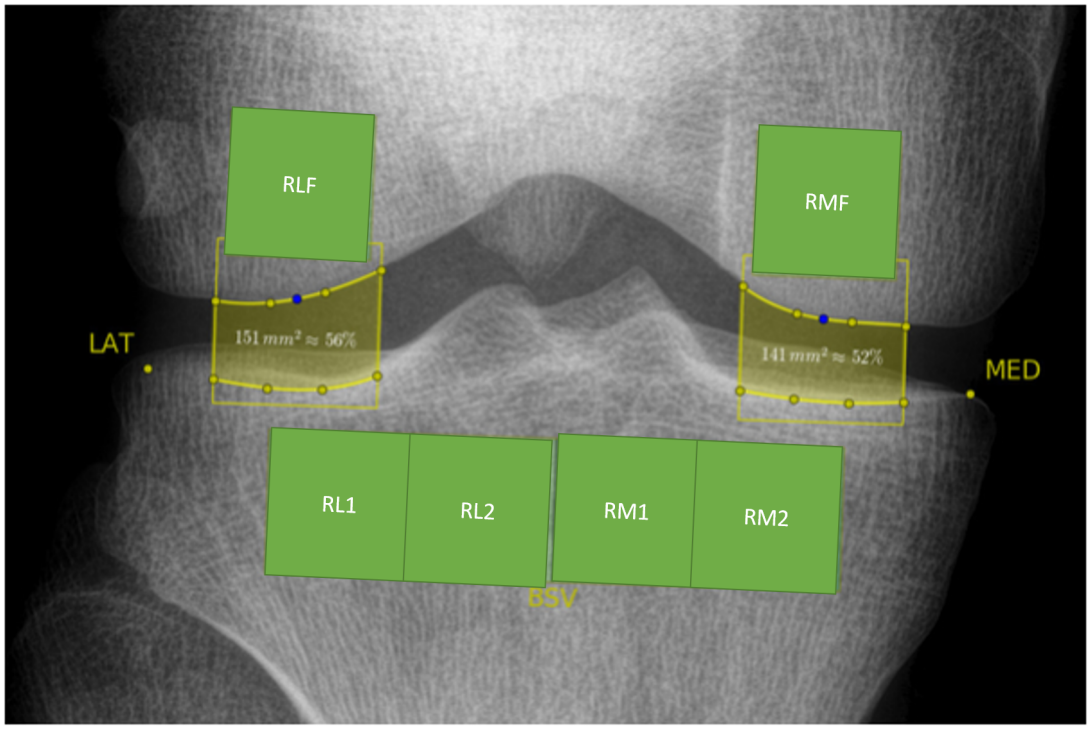


Figure 4.1: The different ROIs detected by the IB Lab Analyzer Software. The naming convention is as follows: *R* stands for *region*, *M* stands for *medial*, *L* stands for *lateral* and *F* stands for *femur*.

4.1 Portugal (EpiReumaPt)

The *EpiReumaPt* study started in September 2011 (and lasted until December 2013) in Portugal out of the lack of well-designed and consistent epidemiologic studies previously available in the country [57]. With this study, the Portuguese Society of Rheumatology (SPR) wanted to fulfill their mission of increasing knowledge and raising awareness about the Rheumatic and musculoskeletal diseases (RMDs) in Portugal.

Among the main aims of the EpiReumaPt are to estimate the prevalence of:

1. hand, knee and, hip OA,
2. low back pain (LBP),
3. RA,
4. fibromyalgia, and

5. osteoporosis (OP).

in the adult Portuguese population. The secondary aims were to determine the impacts of the diseases on the quality of life (such as general well-being, mental health, work status, etc.).

The study target population was composed of non-institutionalized adults and consisted of three phases. In the first phase, each participant was verbally interviewed at his or her house about any possible symptoms due to RMDs, using standardized questionnaires. In the second phase, the participants were observed in a clinic by three radiologists which set a diagnosis and also requested laboratorial tests and different types of image data from each participant, again, following standardized procedures defined previously by the study designers for each disease that was observed. In the third and last phase, the diagnosis set in the second phase were validated using the laboratory results and the images (dual-energy X-ray absorptiometry (DXA) and X-ray modalities). The knee OA was assessed using the ACR criteria [58].

Our goal concerning the Portugal data set was to test whether the algorithms introduced in Chapter 5 can engineer useful features that could aid in the control-case discrimination task regarding OA. For this, out of the 930 patient knee X-Rays that we had access to, we selected a representative sample of patients for which the certain diagnostic, the X-Ray tube voltage (measured in kV), the exposure (measured in mAs), the ethnicity, the age, and the BMI were available. Thus, we selected 171 white women, which are divided in two groups: 86 cases and 85 controls (as shown in Table 4.1). These persons had X-rays recorded at 65 kV. This selection was made to ensure uniform data for the upcoming analyses. The resolution of the Portugal images is 1944 x 3072 pixels, while the pixel spacing is 0.075 micrometer. The number of allocated bits for the displaying of intensities is 16, but only 14 are used.

Table 4.1: Number of controls and cases available in the Portugal data set for model building.

group	case	control
number	86	85

The BMI, voltage, exposure, and age were used to verify their effect on the texture features, as we do not want contributions from any confounding variables. The influences of the confounding variables and the process of removing them will be discussed in more detail in Chapter 9.

For the experiment using this image set all the ROIs (RM1, RM2, RL1, RL2, RMF, RLF) were used. This yielded a total of 126 features: 21 features per ROI. The means for obtaining these features are presented in detail in Chapter 5. We formulated the following hypothesis that we later attempt to invalidate using statistical tests:

There is no significant difference between the means of the OA and non-OA groups in terms of entropy, fractal dimension, and Haralick features.

4.2 MOST

As opposed to the EpiReumaPt study, which targeted a larger number of diseases (see Section 4.1), the MOST study focused solely on knee OA. The MOST is a longitudinal, prospective and observational study funded by the National Institute on Aging in the United States [59]. The main goal of the study was to determine possible new or modifiable risk factors (that could help prevent the development of the disease) for radiographic and symptomatic OA and to verify whether the risk factors of a newly-developed OA differ from the ones of an OA at a more advanced stage. In total, 3026 men and women with preexisting knee OA or at high risk of developing OA were recruited.

MOST completed a baseline (where risk factor screening was carried out) and five follow-up contacts at 15, 30, 60, 72, and 84 months. At each time point, clinical assessments were conducted and radiological data (DXA, X-ray, and MRI of both feet) were collected following standardized procedures. Also other measures like for example the KL) were recorded. The 72-month visit did not generate interesting data for our purposes, as it was a telephone interview only. Our goal with the MOST data set was to test whether the features that our algorithms engineer are suitable for:

1. discriminating OA from non-OA patients,
2. predicting knee OA and
3. tracking the progression of knee OA

based on only 2D X-Rays of knees. To achieve this, for each task we created different study groups that we selected from the large pool of participants based on different criteria that would fit the goal of each task in turn. For our experiment we focused on the baseline (BL), 30 months (30m) and 84 months (84m) visits only. We carried out the same experiments for men and females separately, i.e. in a gender-stratified fashion. We limit our study to the three mentioned time points in order to assure the largest time intervals possible between the visits. This leads to more detectable differences among the time points for the OA cases. Our study design thus led to the formulation of eight general hypotheses that we hope to invalidate through statistical tests, based on the three tasks enumerated above:

1. **OA discrimination task**:
   a) *The cases' trabecular structure does not differ significantly from the controls' in terms of entropy, fractal dimension (FD), and Haralick features for the male group.*

b) *The cases' trabecular structure does not differ significantly from the controls' in terms of entropy, FD, and Haralick features for the female group.*

2. **OA early prediction task**:

   a) *There is no significant difference between the measurements recorded for healthy patients at BL and measurements recorded for the same patients at 30m or 84m if they developed OA, in terms of entropy, fractal dimension, and Haralick features for the male group.*

   b) *There is no significant difference between the measurements recorded for healthy patients at BL and measurements recorded for the same patients at 30m or 84m if they developed OA, in terms of entropy, fractal dimension, and Haralick features for the female group.*

3. **OA progression task**:

   a) *There is a significant difference between patients that indicated constant KL scores across visits in terms of entropy, fractal dimension, and Haralick features both for:*

      i. *males*

      ii. *females*

   b) *There is no significant difference between visits of patients whose KL grade worsened in terms of entropy, fractal dimension, and Haralick features both for:*

      i. *males*

      ii. *females*

For the experiment using the MOST image set, all ROIs (RM1, RM2, RL1, RL2, RMF, RLF) were used. We were interested in discovering which ROI shows the largest contribution to the detection of the disease in the texture. This yielded a total of 126 features (21 texture features per ROI). The methods for obtaining these features are presented in detail in Chapter 5. Only images recorded with constant voltage, exposure, known BMI, and age were used for all the tasks. We used Computed Radiography (CR) images only, all taken with the same machine (AGFA Adc Solo with a pixel spacing of 0.1699), obtained from the center at Birmingham, AL in order to avoid eventual digitization, and image-stitching/manipulation artifacts.

The resolution of the MOST images is 2530 x 2048 pixels, while the pixel spacing is 0.17 micrometer. The number of allocated bits for the displaying of intensities is 16, but only 12 are used. The X-rays were recorded for each patient at each visit for both knees. The KL labeling of the data is available for both knees. In our experiment, we always extracted the features from the left knees for the patients that were controls and we extracted the features from the knee labeled as ill for the patients that were classified as cases.

4.2.1 Diagnosis

For diagnostic purposes we treated each visit separately to investigate where the discrimination is better. The general selection criteria for the control groups were $KL_t == 0$ and $KL_{t>BL} == 0$ (i.e. the patients were healthy throughout the study) and for the case groups were $KL_t > 1$, where $t \in \{BL, 30m, 84m\}$. Thus, for BL we had 189 controls and 138 cases in the female group and 265 controls and 217 cases in the male group. At 30m, 59 controls and 95 cases from the female group and 127 controls and 150 cases from the male group were available. At visit 84m, 257 controls and 245 cases from the female group and 294 controls and 386 cases from the male group were available. The data presented in this paragraph is available in Table 4.2.

Table 4.2: Number of controls and cases available at each visit for the male and female groups separately for the diagnostic task. The number of controls varies across studies due to interrupted visits of the persons involved in the study and the selection criteria applied on the images (in terms of exposure, voltage, etc.).

	BL				**30m**			
gender	male		female		male		female	
group	case	control	case	control	case	control	case	control
number	217	265	138	189	150	127	95	59

Table 4.2: (continued)

	84m			
gender	male		female	
group	case	control	case	control
number	386	294	245	257

Early Prediction

To inspect the early prediction capabilities of our algorithms, we selected as controls the knees that remained healthy at all visits($KL_{BL} = KL_{30m} = KL_{84m} = 0$). The cases group was composed of patients that at BL were healthy ($KL_{BL} = 0$), but at some later point in time developed OA ($KL_{t>BL} > 1$). These criteria yielded 189 and 265 controls and 226 and 343 cases (female and male). The data presented in this paragraph is available in Table 4.3.

Table 4.3: Number of controls and cases available at each visit for the male and female groups separately for the early prediction task.

gender	male		female	
group	case	control	case	control
number	343	265	226	189

Progression

To test if our algorithms can track the disease progression, we constructed four groups of participants:

- patients that will become ill ($KL_{BL} = 0$ and ($KL_{30m} > 0$ or $KL_{84m} > 0$)). This selection yielded: 273, 180 and 345 females and 457, 291, 528 males at BL, 30m and 84m. We called this the ***OA-incidence group***.
- patients that stay healthy throughout the study ($KL_{BL} = 0$ and $KL_{30m} = 0$ and $KL_{84m} = 0$). This selection yielded: 189, 124 and 257 females and 265, 127, 294 males at BL, 30m and 84m. We called this the ***stay-healthy group***.
- patients whose KL score worsened throughout the study ($KL_{30m} > KL_{BL}$ or $KL_{84m} > KL_{BL}$ or $KL_{84m} > KL_{30m}$). This selection yielded: 48, 33 and 170 females and 90, 60, 267 males at BL, 30m and 84m. We called this the ***KL-worsening group***.
- patients whose KL score remained constant throughout the study ($KL_{BL} = KL_{30m} = KL_{84m} = 0$). This selection yielded: 384, 241 and 410 females and 570, 290, 503 males at BL, 30m and 84m. We called this the ***KL-constant group***. The number of participants selected varies across the study due to inconsistent visits and the selection criteria of the images (no stripes as in Figure 4.2, constant exposure, constant pixel spacing, constant voltage, etc.).

4.2.2 Group Balancing

Due to the fact that the data sets that we have constructed were unbalanced, we used the *SMOTE* technique to enrich the smaller groups. For example, for the diagnosis task at 30m we had 95 cases, but only 59 controls, so we equalized the numbers [60]. This technique assures that the newly-generated samples for the smaller group do not leave the boundaries of the so called *elliptic envelopes* of the original, base groups. The main reason for the reduced number of data are our very specific selection criteria. The secondary reason is that after a thorough visual inspection of the images, we found a lot of images with stripe artifacts that would have disturbed the correct application of the algorithms (as shown in Figure 4.2). The power spectra observed are obtained after summing up the image rows into a single 1D signal. The artifacts generally appear on such a power spectrum as high frequency peaks (as shown in Figure 4.2c). We have detected these peaks by convolution of the signal with wavelets of different widths. The peaks that appear on most scales were the peaks corresponding to the stripes. All the images that indicated such artifacts were removed from further analysis, as a consequence.

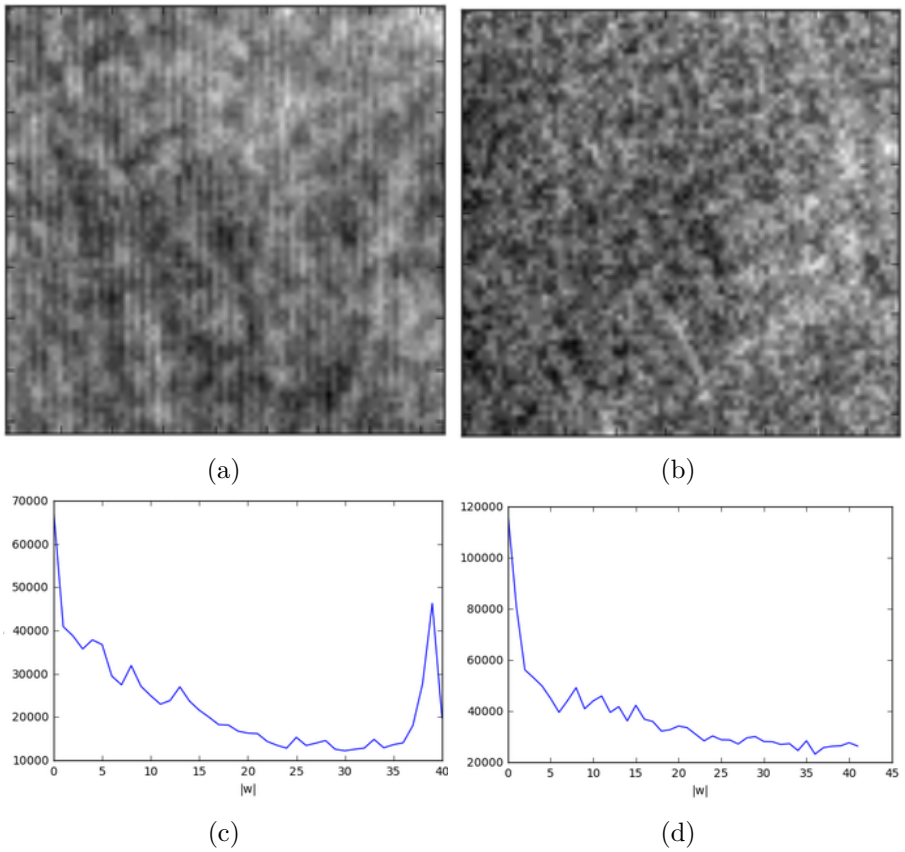


Figure 4.2: Example of ROIs from X-ray images with (a) and without (b) stripe artifacts. The power spectra of the ROI with artifacts (c) and of the ROI without artifacts (d) are also depicted.

CHAPTER 5

Methods

We first introduce each algorithm that we will use to analyze the presented data sets in more detail. We attempt to trace back the origins of each algorithm in order to better understand why they are indeed suitable for characterizing knee TB from radiographs.

First, we look at two fractal-based algorithms (Bone Score Value (BSV) [18] and Bone Variance Value (BVV) [61]) given the fractal properties of the TB [16]. Second, an algorithm based on Shannon Entropy, which stems from information theory, is presented: the Bone Entropy Value (BEV) [62]. Last but not least, an algorithm that computes different image descriptors not directly from the image histogram as the BEV algorithm, but from the so called coocurrence matrices of the image is described (Bone Coocurrence Value (BCV) [19]).

5.1 Fractals

Pentland raised the problem of computational representation of complex natural shapes, such as 'a crumpled newspaper', 'a clump of leaves' or 'a jagged mountain'. He stated that Plato's notion of ideal forms (e.g., spheres, cylinders, cubes) are too 'primitive' to achieve this [63]. Moreover, using only the ideal forms it would be impossible to extract 3-D information from the image of a 'rough' or 'crumpled' surface if all the available models assume smooth surfaces. As a consequence, fractal functions have been proposed as better generators of naturally-looking surfaces. This is due to the fact that basic physical processes produce fractal surfaces: formation of clouds, leaves growth (see Figure 5.1), tree growth, TB (see Figure 5.2) etc. Thus, fractals are common in nature [64, 65], with Mandelbrot showing that fractal surfaces are indeed produced by basic physical processes that range from the aggregation of galaxies to the curdling of cheese.

In one of our previous works we have provided an extensive definition of fractals [23]. In this work we only reiterate the basics to refresh the memory of the reader. A structure,

a surface, or a shape is thus considered to be a *fractal* if it possesses a defining set of features, called *fractal properties*. These properties are put together by Falconer [66]. He states that if F is a fractal, the following properties will apply in most of the cases:

P1. F has a fine structure, i.e., detail on arbitrarily small scales.

P2. F is too irregular to be described in traditional geometrical language, both locally and globally.

P3. Often F has some form of self-similarity, perhaps approximate or statistical.

P4. Usually the *FD* of F (defined in some way) is greater than its *topological dimension – D*.

P5. In most cases of interest, F is defined in a very simple way, perhaps recursively.

One can thus notice that a fractal is too complex to be described by simple, traditional Euclidean geometry. For this matter, Mandelbrot introduced the field of *fractal geometry* [65], which provided not only some new insight into the intricate properties of fractals, but also tools for analyzing and characterizing their structure. In traditional geometry one must only apply the following generalized formula to measure some metric property M (such as length, area, volume):

$$M = nr^D \tag{5.1}$$

where r is the 'measuring stick' (size of the measuring unit), D is the topological dimension of the measuring instrument (e.g., 1 – a line, 2 – an area, 3 – a volume) and n is the number of such units needed to 'cover' M completely [63, p. 663].

The importance of fractal geometry and especially of the *FD* becomes clear once we present the most popular example that illustrates the need of such a dimension: measuring the length of the coastline of an island (see Figure 5.3). The smaller we choose to be the size of our measuring tool (can also be seen as magnification), the more of the coastline is covered (see Figure 5.4), but there still remain regions of the coastline where it can not fit and thus those features are missed. This means that at this point the measurment does not depend only on the subject but also on the measuring tool. Mandelbrot stated that in order to compensate for the length lost due to the smaller details than the measuring unit, a fractional power (dimension) must be introduced [67]:

$$N \propto r^{-FD} \tag{5.2}$$

where N is the number of non-overlapping segments of size r in which the previous segment (measuring stick) is divided in a process of recursively 'decreasing' the size of the measuring stick. Rearranging Equation 5.2, which has the form of a general scaling law, yields:

$$\log N \propto -FD \log r. \tag{5.3}$$

And:

$$FD \propto \frac{\log N}{\log 1/r}. \tag{5.4}$$

Finally:

$$-FD \log r \propto \log N. \tag{5.5}$$

At this point it is clear that *FD* is proportional to the slope of the log-log plot of r against N (see Equation 5.5). Indeed, if we calculate this rate of change, we obtain approximately 1.21 as the *FD* of the Great Britain coast line. Thus, the *FD* acts as an 'adjustment' factor for the details smaller than the chosen measuring unit that were lost; 'it may also be viewed as a measurement of the shape's roughness' [63, p. 663]. As a consequence, for natural shapes, any description that does not contain the *FD* as correction factor will not be correct at more than a single scale.

Since the *FD* of a surface can not be calculated directly just by applying Equation 5.4, other approaches of approximating this value must be employed. Additionaly, given the fact that the TB is known to feature fractal properties [16], in the next two sections, two algorithms that calculate the *FD* of a surface slightly differently are presented. These algorithms also do not calculate *FD* directly, but calculate a so called Hurst exponent [68, 69], which relates to the *FD* as follows [70]:

$$FD = D + 1 - H, \tag{5.6}$$

where D is the topological dimension of the surface and H is the Hurst coefficient (H). In case of 2D surfaces (radiographs of TB), Equation 5.6 becomes:

$$FD = 3 - H. \tag{5.7}$$

This means that a fractal whose *FD* lies between 2 and 3 is too complex to be described by only two topological dimensions (e.g., length and width), but is not complex enough that another 'depth' dimension should be added. Its complexity lies somewhere in between and the degree of complexity is fully described by H. Given the fact that H lies between 0 and 1, a value smaller than 0.5 indicates a 'wild' randomness. For example, in Equation 5.7 if $H < 0.5$ the *FD* comes closer to 3, which means that the complexity of the signal almost reaches three dimensions. However, value bigger than 0.5 indicates a 'mild' randomness of a signal. For example, in Equation 5.7 if $H > 0.5$ the *FD* stays closer to 2, meaning that the complexity of the signal is not much higher than only two dimensions; there is not much more information needed than two dimensions to describe the underlying surface. The two cases can also be interpreted as 'negative' and 'positive' autocorrelation of a signal. In case of H taking the value of exactly 0.5, there is no correlation between a signal's past and current state, but complete randomness is observed [71].

With this in mind, we will first present two algorithms based on slightly different approaches that attempt to estimate the same parameter, namely the Hurst exponent.

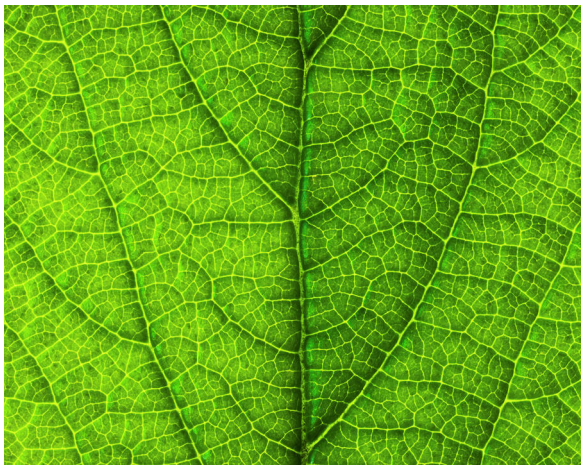


Figure 5.1: Fractals in nature: leaf [72]

.

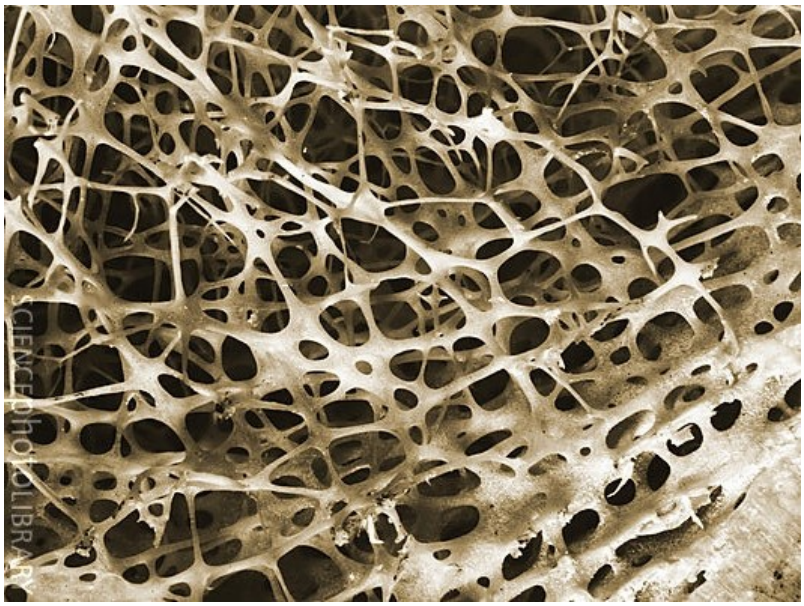


Figure 5.2: Fractals in nature: human TB [73].

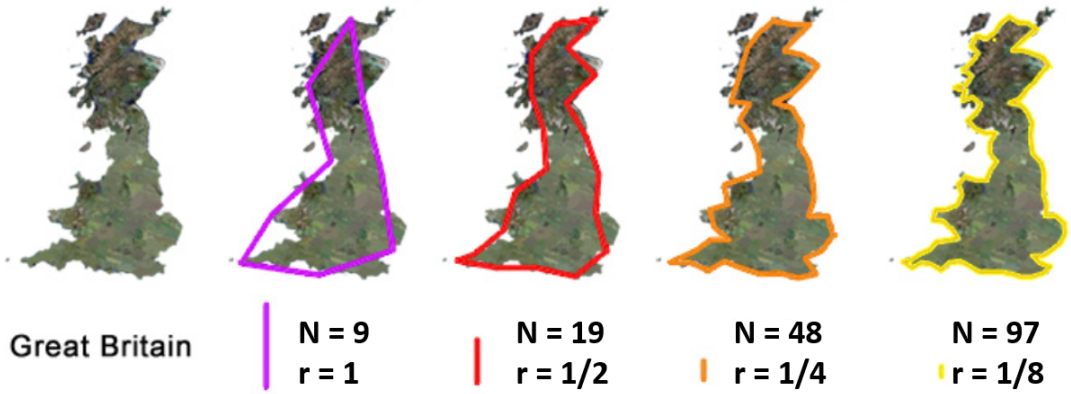


Figure 5.3: Great Britain coast line measured with different measuring units [74].

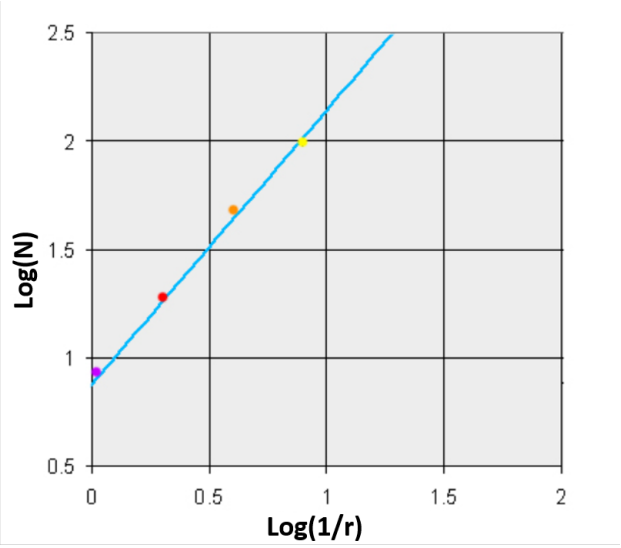


Figure 5.4: Log-log plot of different magnification levels r against perimeter N [74].

5.1.1 Bone Variance Value (BVV)

The BVV is the first fractal-based algorithm that we introduce and describe. In this case, is is assumed that the intensity surface of an image is generated by a stochastic process, more specific by a fractal Brownian function, which is the mathematical generalization of the Brownian motion [61]. In 1977, Mandelbrot stated that a random function $I(x)$ is a fractal Brownian function if for all x and Δx the probability Pr is given as[64]:

$$Pr\left(\frac{I(x+\Delta x)-I(x)}{||\Delta x||^H} < y\right) = F(y), \tag{5.8}$$

where $F(y)$ is a cumulative distribution function and H is the Hurst exponent. Like for most of the parameters of any natural phenomenon or shape, it is assumed that the intensities of the image representing TB are drawn from a normal distribution, which means that differences of intensities are also drawn from a normal distribution. Due to the fact that the fractal Brownian motion is the only self-similar Gaussian process [75], we can rearrange Equation 5.8 in the following manner [63, p. 665]:

$$VAR\left[\frac{I(x+\Delta x)-I(x)}{||\Delta x||^H}\right] = VAR[I(x+1)-I(x)], \tag{5.9}$$

which shows how the second-order statistics of the image change with scale. By applying the basic property of the variance:

$$VAR[aX+b] = a^2 VAR[X], \tag{5.10}$$

where X is a random variable and a a linear scaling factor and b a constant, the scaling constant $\frac{1}{||\Delta x^H||}$ can be extracted and Equation 5.9 becomes:

$$\frac{1}{||\Delta x^{2H}||} VAR[I(x+\Delta x)-I(x)] = VAR[I(x+1)-I(x)], \tag{5.11}$$

which finally leads to the statement

$$VAR[I(x+\Delta x)-I(x)] \propto ||\Delta x^{2H}|| \tag{5.12}$$

In other words, the variance of the distribution of differences is proportional (up to a constant, which are the differences if $\Delta x = 1$ pixel) to the chosen spatial distance or scale Δx between the intensities, 'corrected' by the power $2H$. By applying *log* to Equation 5.12 we obtain

$$\log\left(VAR[\Delta I_{\Delta x}]\right) \propto 2H \log\left(\Delta x\right) \tag{5.13}$$

which means that one can estimate the Hurst exponent H by fitting a line to the log-log plot of variances versus differences and calculating 1/2 of the slope of it.

Equation 5.13 shows how powerful this approach of estimating H is. By varying Δx values, one can employ different 'measuring sticks' and by assigning a direction to these units, a fractal can be described in a complex manner. This is useful due to the fact that TB is an anisotropic structure and its properties do not only vary with direction, but also with scale [76]. As a consequence, Wolski et al. employed a version of this variance method where they calculate H values for 24 directions (i.e., every 7.5°) and for six scales along each direction (as shown in Figure5.5). This is called the Variance Orientation Transform (VOT). However, we have decided to restrict the number of directions so that the algorithm is applicable also for X-rays with lower resolutions. As a consequence we limited the algorithm to calculate H values for only eight directions (i.e., every 45°). In reality, these directions reduce to four due to symmetry. At the same time, as opposed to how we presented the algorithm in its first iteration [23], we discard the different scales. In other words, we compute a single H along every mentioned direction.

We cut each beam along each direction to be five pixels long. Wolski et al. did not consider the first four pixels due to a possible 'digitization error' [17, p. 213]. Also, Wolski et al. cut off their beams at 16 pixels, making the beams 12-pixels long. By defining beams with lengths of five pixels, we disregard the low frequencies, that correspond to larger structures, from the calculations. We have shown that these regions have a negative impact on the fractality of the image (as shown in Figure 5.6). In other words, the intensities in the X-rays that we analyze appear to follow the power rule of fractals (see Equation 5.12) only at small distances, but not over longer ranges.

This algorithm computes an approximation of H at $0°, 45°, 90°$ and 135° to account for the fractality of the TB. A mean H is also computed. This yields a total of five BVV features per image.

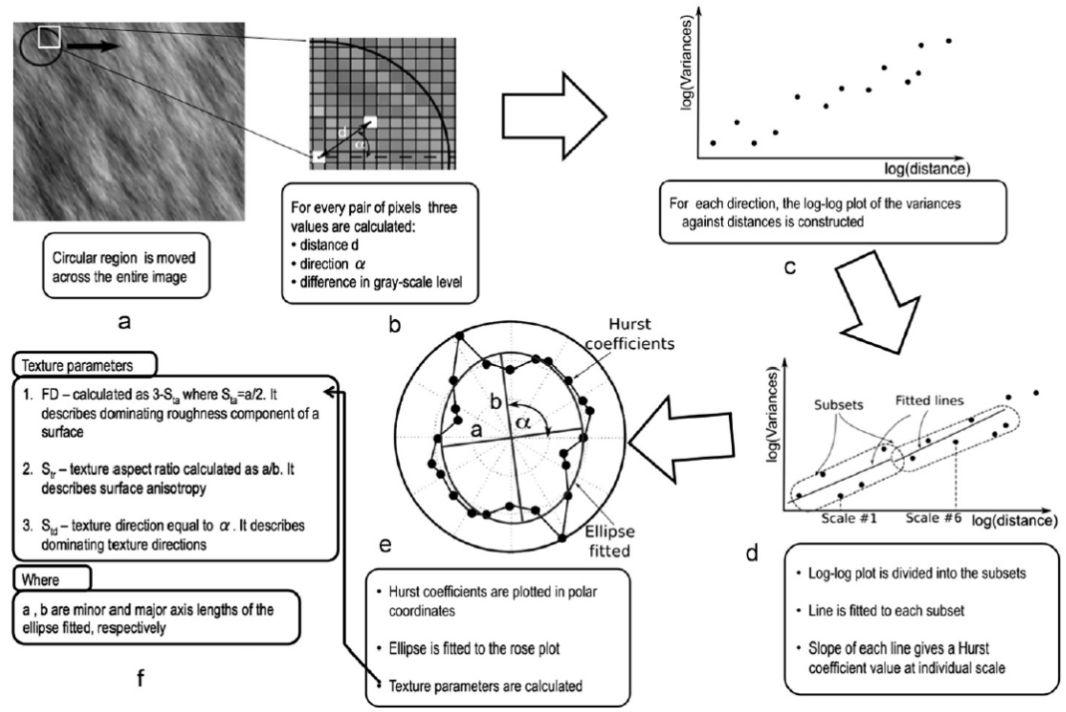


Figure 5.5: A schematic illustration of the VOT method: (a) a search region that moves across the image, (b) values calculated for a pair of pixels within the region, (c) a log-log plot, (d) lines fitted to the plot, (e) a rose plot of Hurst coefficients and (f) texture parameters calculated from the ellipse fitted [77].

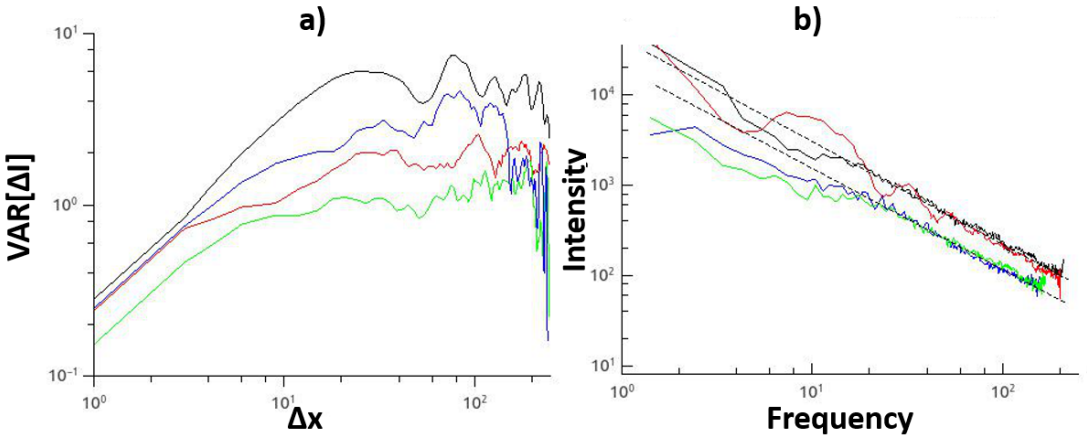


Figure 5.6: Different, randomly selected ROIs from our data sets (color-coded) to show where the 'decoherence' of variances of differences and scale begins. a) log-log plot of $VAR[\Delta I_{\Delta x}]$ against scales Δx. b) Power spectra of the said ROIs.

5.1.2 Bone Score Value (BSV)

The second fractal-based algorithm that we introduce is the BSV. This algorithm is based on the work of Lundahl [18]. In this case, the assumption that the function describing the intensity surface of the image is a stochastic process generated by a fractal Brownian function (fBf) must also be made. A fractal Brownian function (fBf) is governed by a single parameter, the Hurst exponent [18, p. 152]. At the same time, the Hurst exponent was shown to be directly related to the *FD* of a function, describing its 'intuitive roughness' (cf. Equation 5.7) [18, p. 152]. In other words, a realization of a fractal Brownian motion (fBm) describes a fractal set.

In this section we first take a closer look at the background of a fBm in order to understand the approach by which the BSV algorithm is approximating the Hurst exponent.

Brownian Motion

A chaotic, random movement of particles within a medium is called a Brownian motion. This is a physical phenomenon, which was noticed by Robert Brown in 1827 while observing the behaviour of pollen grains in a watery milieu. The mathematical formalization of this process is called the *Wiener process* and is described and analyzed in much detail in the work of Moerters and Peres [78]. The extensive information presented there is out of the scope of this present work. However, Moerters and Peres proved a very simple definition of such a Wiener process in their book that is clearly formulated:

Definition 5.1. *A real-valued stochastic process $B_t : t \in T; t \geq 0$ is called a (linear) Brownian motion with start in $x \in \mathbb{R}$ if the following hold:*

- *$B(0) = x$; if $x = 0$ then the process is also called a 'standard Brownian motion',*
- *the process has independent increments, i.e., for all time points $0 \leq t_1 \leq t_2 \leq ... \leq t_n$ the increments $B(t_n) - B(t_{n-1}), B(t_{n-1}) - B(t_{n-2}), ..., B(t_2) - B(t_1)$ are independent random variables,*
- *for all $t \geq 0$ and $h > 0$, the increments $B(t + h) - B(t)$ are normally distributed with expectation 0 and variance h,*
- *the function $B(t)$ is continous.*

Fractal Brownian motion (fBm)

A fBm is a generalization of the Brownian motion introduced above for enabling applicability in other fields, such as probability theory. In comparison to the general Brownian motion, the fBm is defined as follows:

Definition 5.2. *A real-valued stochastic process $B_H(t)$ with t on $[0, T]$ is a fBm if the following hold:*

- *$B_H(t)$ is a continuous-time Gaussian process.*
- *its increments need not be independent.*
- *has expectation 0 for all $t \in T$.*
- *is a first-order stationary process, meaning that the process constructed from the first-order increments $X(t) = B_H(t+1) - B_H(t)$ are stationary. This process is called fractal Gaussian noise (fGn).*
- *is a second-order non-stationary process, as shown by the variance law: $Var(B_H[i]) = \sigma^2 i^{2H}$, where i is a discrete time index [79, p. 327]*
- *is statistically self-similar (cf. Equation 5.1) due to the fact that the covariance function (also autocorrelation function in this case) is a homogeneous function of order $2H$. The fBm is the only self-similar Gaussian process.*
- *is defined by the covariance function $E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H})$, where t and s are two different instances in time and H is the Hurst exponent.*

The Algorithm

At this point, after introducing the basic theory regarding the fBm, we can introduce the BSV algorithm. As we have seen in the previous sections, the fBm is a Gaussian process, whose probability density function (PDF) can be expressed as:

$$Pr(X) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)}, \tag{5.14}$$

where d is the number of dimensions of the multivariate process, X is the $1\ x\ d$ data vector (in our case, a row of intensities from the TB ROI) and μ is a $1\ x\ d$ vector containing all the means for each random variable in X and Σ is the $d\ x\ d$ symmetric positive definite covariance matrix. In the same manner, the increments of the fBm, the fGn, also build a Gaussian process. In other words, the fGn is multivariately normally distributed as well, whose PDF can be expressed simplified as:

$$P(G|H) = \frac{1}{\sqrt{(2\pi)^d |R|}} e^{-\frac{1}{2} G^T R^{-1} G}, \tag{5.15}$$

where H is the Hurst exponent, G is the fGn computed from the fBm and R is the Σ from Equation 5.14. It was renamed to point to the more special covariance (autocorrelation) function of a fGn, which must be derived from the covariance function of a fBm in the

following manner, by making use of the variance law introduced in Definition 5.2:

$$\begin{aligned}
&E[(B_H(t+1) - B_H(t))(B_H(t+k+1) - B_H(t+k))] = \\
&\quad = E[B_H(t+1)B_H(t+k+1)] + E[B_H(t)B_H(t+k)] \\
&\quad\quad - E[B_H(t+1)B_H(t+k)] - E[B_H(t)B_H(t+k+1)] \\
&\quad = \frac{\sigma^2}{2}[(|t+1|^{2H} + |t+k+1|^{2H} - |k|^{2H}) + (|t|^{2H} + |t+k|^{2H} - |k|^{2H}) \\
&\quad\quad - (|t+1|^{2H} + |t+k|^{2H} - |k-1|^{2H}) - (|t|^{2H} + |t+k+1|^{2H} - |k+1|^{2H})] \\
&\quad = \boldsymbol{\frac{\sigma^2}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H})}
\end{aligned} \tag{5.16}$$

Note that the mean μ is missing from Equation 5.15 due to Definition 5.2, which states that a fBm has expectation 0 at all time points. This obviously applies to the increments, the fGn, as well.

As we can see at this point, the covariance function of the fGn (which basically generates the entries of R) depends on H (see Equation 5.16). In other words the PDF of the fGn depends on H as well. We can now estimate the H by maximizing the likelihood $\mathcal{L}(H|G) = P(G|H)$. In [18] it was shown that $P(G|H)$ is unimodal, meaning that it has a unique maximum. We do not change this maximum by applying the logarithm to the PDF, since the log is a monotonic function. In other words we can maximize $log(P(G|H))$ instead, which leads to a more simplified form:

$$log(P(G|H)) = -\frac{d}{2}log(2\pi) - \frac{1}{2}log|R| - \frac{1}{2}G^T R^{-1} G, \tag{5.17}$$

where the elements of R are given by the autocorrelation function (cf. Equation 5.16). At this point we notice that the variance σ^2 is also unknown and must be estimated. We can, however, get rid of it by decomposing $R = \sigma^2 R^{prime}$ with the likelihood function becoming:

$$log(P(G|H)) = -\frac{d}{2}log(2\pi) - \frac{1}{2}log|R^{prime}| - \frac{1}{2}log(\sigma^2) - \frac{1}{2\sigma^2}G^T(R^{prime})^{-1}G \tag{5.18}$$

By computing the derivative of the likelihood with respect to σ^2 and letting it go to zero we obtain an approximation for σ^2, which maximizes the likelihood function:

$$\hat{\sigma}^2 = \frac{G'(R^{prime})^{-1}G}{d} \tag{5.19}$$

Inserting this approximation into Equation 5.18 yields the final function to be maximized with respect to H:

$$log(P(G|H)) = -\frac{d}{2}log(2\pi) - \frac{1}{2}log|R^{prime}| - \frac{1}{2}log(\frac{G'(R^{prime})^{-1}G}{N}) - \frac{d}{2}, \tag{5.20}$$

where the constants can further be removed for an even simpler form.

We observe that Equation 5.20 is not an explicit form for an estimate $\hat{H}$ as a function of the data G. In this case, numerical methods must be employed to find the maximum with respect to H. In this case, we used the Golden Section Search to narrow down the possible value of H.

For each input image, the algorithm computes a vertical, a horizontal and a mean approximation of H, yielding three features.

5.2 Information Theory and other Image Properties

As opposed to the fractal approaches introduced and presented in Section 5.1, in this section we introduce two approaches that deal with the raw, unprocessed pixel intensities to produce characteristic features. These features are directly related to the integrity, homogeneity, and correlation of the structure represented by the respective intensities.

5.2.1 Bone Entropy Value (BEV)

The first algorithm that we introduce in this section stems from the field of *Information Theory* and is called *Shannon's Entropy*. It bears the name of Claude Shannon, who first introduced the idea of *information entropy* in 1948 [62].

Information Theory

Before we can explain how the BEV algorithm works based on information entropy, we first need to give a short and intuitive introduction to the field of *Information Theory* as summarized by Carter [80]. Suppose there is an event, which occurs with probability p and carries a certain information content or information amount. We can then intuitively develop some rules that must apply to this information as a function I:

- Information is non-negative: $I(p) \geq 0$.
- If the probability p of an event is 1, than it means that this event carries no "surprise" or no additional information, therefore $I(1) = 0$. The reverse also holds for the case if $p = 0$.
- If two independent events occur, then the information content of both is the sum of the individual information contents: $I(p_1 * p_2) = I(p_1) + I(p_2)$.
- I must also be a continuous function of the probability. "Slight changes in probability should result in slight changes in *information*" [80, p.16].

At this moment we observe that the only function that fulfills all of the rules specified above is the *log* function! This means that:

$$I(p) = -log_b(p), \tag{5.21}$$

where b is a placeholder for any possible base. Due to the fact that the p lies between 0 and 1 we needed to negate the *log* so the information content remains non-negative.

Information Entropy

With a clear definition of the information content of an event (message, byte stream, etc.), we can now define the entropy held by that particular event.

Definition 5.3. *Given a probability distribution Pr, the entropy of the distribution P is defined by:*

$$E(Pr) = -\int Pr(x)log(Pr(x))\,\mathrm{d}x. \tag{5.22}$$

in case of a continuous probability distribution, or

$$E(Pr) = -\sum_{i=1}^{n} p_i log(p_i). \tag{5.23}$$

in case of a discrete distribution (such as the histogram of an image for example).

In other words, we can say that the entropy of a probability distribution of an event is the expected value of the information content of that distribution:

$$E(P) = \mathbb{E}(I(P)) \tag{5.24}$$

Based on Definition 5.3, we can now introduce the definition for *Shannon's Entropy*, which is:

$$S(P) = -\sum_{i=1}^{n} p_i log_2(p_i), \tag{5.25}$$

i. e., the base two logarithm was used for Equation 5.23. Due to this, the units of this entropy are usually referred to as *bits*.

The Algorithm

Building on previous thoughts and definitions we can now construct an algorithm that can describe an image in terms of its information complexity. Multiple circular masks with given, preset radius are placed on each input image. The length of the radii depend on image resolution and observed structure sizes, i.e., TB dimensions lie in the range 1μm - 100μ (see Chapter 3). Histograms are computed for each region and using the information provided by the histogram, local entropies are computed. On the obtained distribution of entropies we can then compute different statistics (see Figure 5.7). The chosen metric for the purpose of this work was the mean of entropies. This yields a

'global' entropy measure descriptive of the whole image. We consider that the mean is a good descriptor of the entropy distribution because we work with real data. As a consequence, we assume a single peak in the distribution of the entropies, which when tested proves to be true. However, if there are more peaks, the mean would still be a good descriptor for the average information content in the image. The higher the value, the more 'chaotic' the underlying signal is and vice-versa.

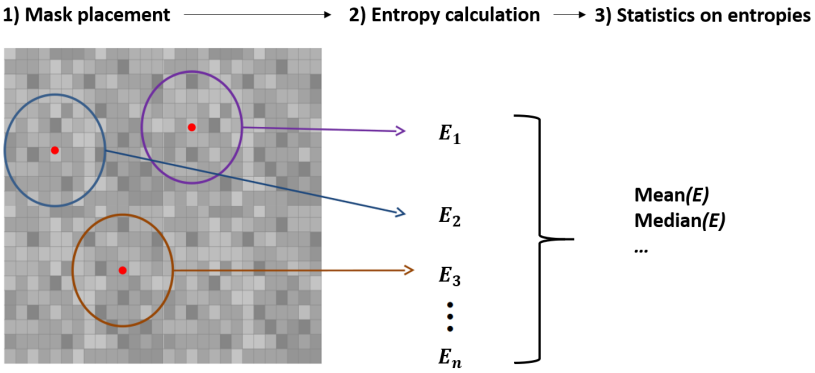


Figure 5.7: Pipeline showing how Shannon's Entropy can be used to characterize an image.

5.2.2 Bone Coocurrence Value (BCV)

The second algorithm that we describe in this s ection stems from the original work of Haralick ([19]). The author based his findings on the premise that "texture and tone bear an inextricable relationship to one another. Tone and texture are always present in an image, although one property can dominate the other at times" [19, p. 611]. This algorithm uses the so-called *Co-occurrence Matrix (CM)* to characterize a given surface. As a consequence, we must first introduce how the CM is computed.

Co-Occurence Matrix

The CM is a matrix computed over the intensity distribution of an image. It is used to describe the spatial relationship between pairs of pixels at a fixed given offset, in any dimension. Due to the dimensionality and color map of radiographs (i.e. 2D and gray-scale, in this work we will only consider a particular case of CM, namely the Gray-Level Co-occurrence Matrix (GLCM). A general formula for computing a set of GLCM for an image is as follows:

$$C_{\Delta x,\Delta y}(i,j) = \sum_{x=1}^{n}\sum_{y=1}^{m} \begin{cases} 1, & \text{if } I(x,y) = i \text{ and } I(x+\Delta x, y+\Delta y) = j \\ 0, & \text{otherwise} \end{cases} \tag{5.26}$$

where $I(x, y)$ is the intensity value at location (x, y) in an mxn image, $(\Delta x, \Delta y)$ is a known offset defined in x and y directions, i.e., horizontally, vertically, and diagonally (see Figure 5.8) and i and j are intensity values. In other words, a CM holds at position (i, j) the number of pairs of *reference* (at location (x, y) in the image) and *neighboring* (at location $(x + \Delta x, y + \Delta y)$) pixels that have intensities i and j respectively, at a given offset $(\Delta x, \Delta y)$. At this point, we notice that for every combination of any possible offset (that lies between the boundaries of an image) and of the four directions shown in Figure 5.8 a CM can be computed. Moreover, "higher order" matrices can be built if more than a single neighboring value is considered. For example, a third-order CM would take into account a single reference location as before, but two neighboring values for comparison and so on.

A CM is usually also expressed as a probability. To achieve this, each entry is divided by the sum of the entries. This way, each entry of the normalized CM holds the probability of occurrence of a specific pairing of pixels.

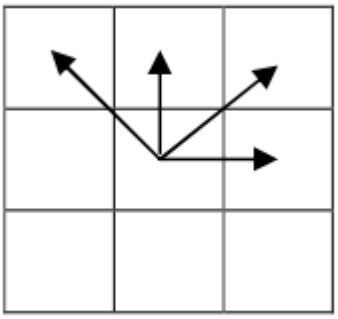


Figure 5.8: The four adjacency directions. Illustrated is the case only for a fixed offset of on [81]

.

The Algorithm

A CM has the same number of rows and columns as the quantization level of the image. In other words, if an image is 16-bit encoded, the resulting matrix will be $(2^{16} - 1) = 65535$ x $65535 = (2^{16} - 1)$. This alone poses storage and performance problems, let alone computing more than one matrix. As a consequence, generally the algorithms that make use of the CM for texture analysis first perform a down-scaling of an the images to 4- or 5-bit depths. By definition, the CM is a symmetrical matrix. Another trick for speeding up the computation of the matrix is to first generate the entries of the half above the main diagonal and then add this resulting matrix to its transpose. This is a significant improvement over the classical double-counting method that becomes noticeable in case of large and/or many matrices.

Features

As stated above, for each combination of offsets and angles a CM is computed. Each of these matrices is then used to compute different features describing the texture of an image. These features can be classified in three groups, according to the effects of the used weights in the formulas ([82]). In the following listings, $p(i,j))$ represents the normalized version of the (i,j) entry in the CM and Q is the quantization level of an image.

1. **Contrast group:**

   a) **Contrast** (Sum of Squares: Variance): $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} p(i,j)(i-j)^2$

   When the considered pixels share the same intensity (i.e. $i=j$), the weight attributed in this case is 0. If the differences in intensities are bigger than 1, the weights increase polynomially by the power of 2. In other words, larger differences are weighted more.

   b) **Dissimilarity**: $\sum_{i=0}^{Q-1}\sum_{j-0}^{Q-1} p(i,j)|i-j|$

   As compared to the polynomially increasing weights in the case of contrast and dissimilarity, here the weights only increase linearly.

   c) **Homogeneity** (Inverse Difference Moment): $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} \frac{p(i,j)}{1+(i-j)}$

   As compared to the polynomially increasing weights of power 2 in the case of contrast and dissimilarity, homogeneity weights by the inverse of intensity differences.

2. **Orderliness group:**

   a) **Angular Second Moment**: $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} p(i,j)^2$

   b) **Entropy**: $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} p(i,j)(-ln(p(i,j))))$

   This measure is not related with our BEV. The BEV is derived from the histogram of the raw image data, while this measure is derived from the CM.

3. **Descriptive statistics group:**

   a) **Mean (reference: σ_i)**: $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} i\ p(i,j)$ The CM mean is a different measure from the image mean. While the usual mean is weighted by the frequency of appearance of a quantity by itself, in this case the weighting represents the frequency of occurrence of a quantity strictly in combination with another one. This mean can be defined both with respect to reference pixels and with respect to neighboring pixels (see below).

b) **Mean (neighbor: σ_j)**: $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} jp(i,j)$

c) **Variance (reference: σ_i^2)**: $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1}(i-\sigma_i)p(i,j)$

d) **Variance (neighbor: σ_j^2)**: $\sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1}(j-\sigma_j)p(i,j)$

e) **Correlation**: $\sum_{i=0}^{Q-1}\sum_{j-0}^{Q-1} p(i,j)\frac{(i-\sigma_i)(j-\sigma_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}$

From the features enumerated above we have selected dissimilarity, homogeneity, and correlation as representative features for the BCV algorithm that we will test on the available data sets. In total, 12 features, i.e., in four directions for each group, are generated with the BCV method.

5.3 Feature Summary

In this section we provide an overview of all the features that the algorithms presented above generate. In Table 5.1 the feature codes and the number of features per algorithm are recorded.

Table 5.1: Summary of features produced by the methods used. With trailing *H* we mark features measured in the horizontal direction, with *V* in the vertical direction, with *D1* in the direction of the first diagonal (i.e., 45°) and with *D2* the features measured in the direction of the second diagonal (i.e., 135°). *M* stands for the mean value and *DISS*, *HOM* and *CORR* for dissimilarity, homogeneity and correlation in the case of the BCV algorithm.

	Feature Code	**Total Number**
BEV	BEV	1
BSV	BSV:M, BSV:H, BSV:V	3
BVV	BVV:M, BVV:H, BVV:V, BVV:D1, BVV:D2	5
BCV	BCV:DISS:H, BCV:DISS:V, BCV:DISS:D1, BCV:DISS:D2, BCV:HOM:H, BCV:HOM:V, BCV:HOM:D1, BCV:HOM:D2, BCV:CORR:H, BCV:CORR:V, BCV:CORR:D1, BCV:CORR:D2	12

CHAPTER 6

Method Validation

In the previous Chapter we have introduced in detail the algorithms that we are employing in the hope of producing useful features from simple 2D knee radiographs that could characterize the degenerated TB due to the presence of (early) OA. However we have not yet got an idea whether the algorithms are working as they are supposed to. Thus, in this Chapter we describe our approach to validate the said algorithms using artificially generated images with known theoretical values. For example, in case of BSV and BVV there are methods to build artificial fractals with known FD, which are then compared with the FD produced by the algorithms through the applied heuristics. Ideally, the computed FD should match the theoretical FD.

6.1 BSV and BVV Validation

We will first attempt at validating our two fractal algorithms. We are treating them together due to their similar nature and goal and the fact that the same image data, namely artificial fractals, can be used to validate both. To generate artificial fractals (isotropic and anisotropic) we use the *power spectrum method* as suggested by Russ [83]. According to Russ, the general requirement of a surface to be of fractal nature is that

$$I \propto |w|^{-(H+1)/2}, \tag{6.1}$$

where I is the intensity surface (of the power spectrum), H the Hurst coefficient (H) and w the frequency. Generally

$$|w|^2 \propto w_x^2 + w_y^2 \tag{6.2}$$

where w_x and w_y are the signal frequencies in the x and y direction respectively. If we also want to control for the direction of the fractal, the procedure is slightly more complex

due to the fact that another parameter is required, namely the angle a with respect to the horizontal direction, which specifies the dominant orientation of the surface:

$$|w|^2 \propto \frac{w_x}{w_y}((w_x cos(a) + w_y sin(a))^2 + (w_x sin(a) - w_y cos(a))^2) \tag{6.3}$$

The intensities calculated applying the equations introduced above belong to the power spectrum of the fractal. If the inverse Fourier transform is applied on this spectrum, the desired fractal is obtained. If more than one such spectra are superimposed, a surface with different FDs in different directions can be obtained. To illustrate the results of this procedure, in Figure 6.1 we can observe an isotropic fractal ($w_x = w_y$) with a theoretical H of 0.2. In Figure 6.2 an isotropic fractal with a theoretical H of 0.7, and in Figure 6.3 an anisotropic fractal with a theoretical H of 0.3 in the 15° direction and its associated power spectrum are shown.

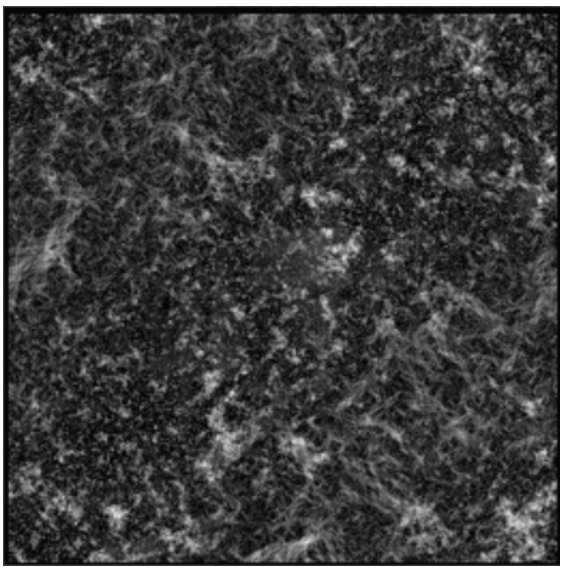


Figure 6.1: Isotropic fractal example with a theoretical Hurst exponent of 0.2 generated with the power spectrum method.

To test the mathematical correctness of the BSV and BVV algorithms we have generated a set of 1600 isotropic fractals of sizes 32x32, 64x64, 128x128, 256x256 with Hurst coefficients between 0.1 and 0.8 in steps of 0.1 (i.e. 50 images for each size and Hurst coefficient combination). We then calculated the H using the algorithms and compared them in H-H plots with the theoretical H values set at generation (as shown in Figure 6.4). The diagonally-measured (45°) Hs of the BVV algorithm are covered in Figure 6.5 separately since BSV has no capability at the moment to determine the Hs at intermediary angles. We observe that in general, the algorithms are more stable with images of larger sizes (128x128 and 256x256) in the case of isotropic fractals.

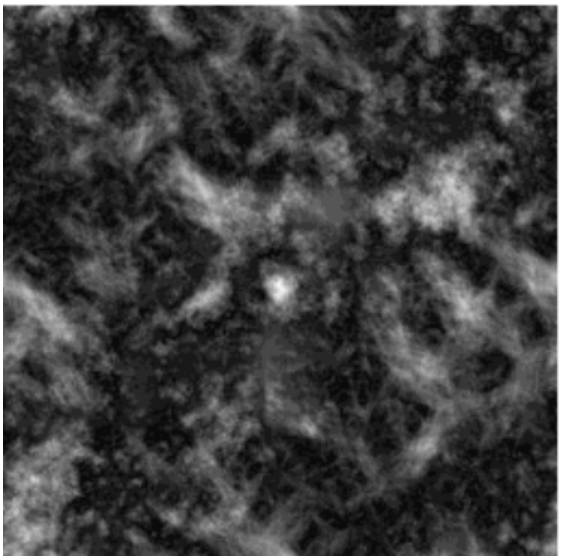
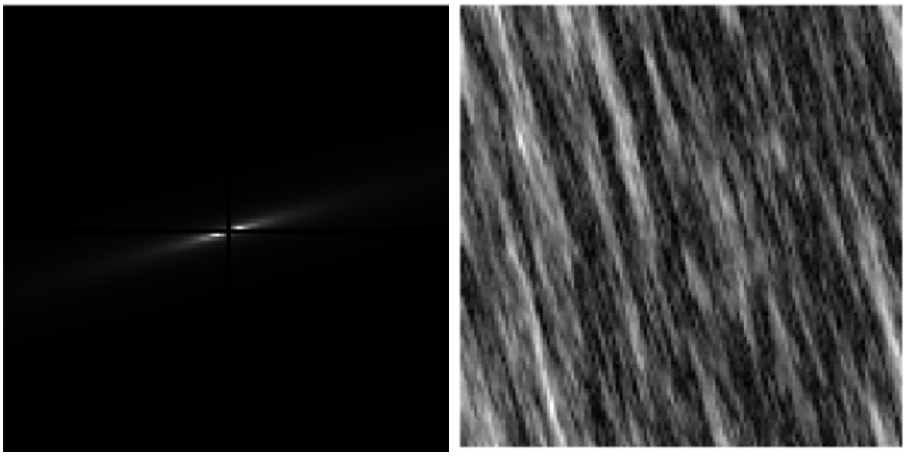


Figure 6.2: Isotropic fractal example with a theoretical H of 0.7 generated with the power spectrum method.

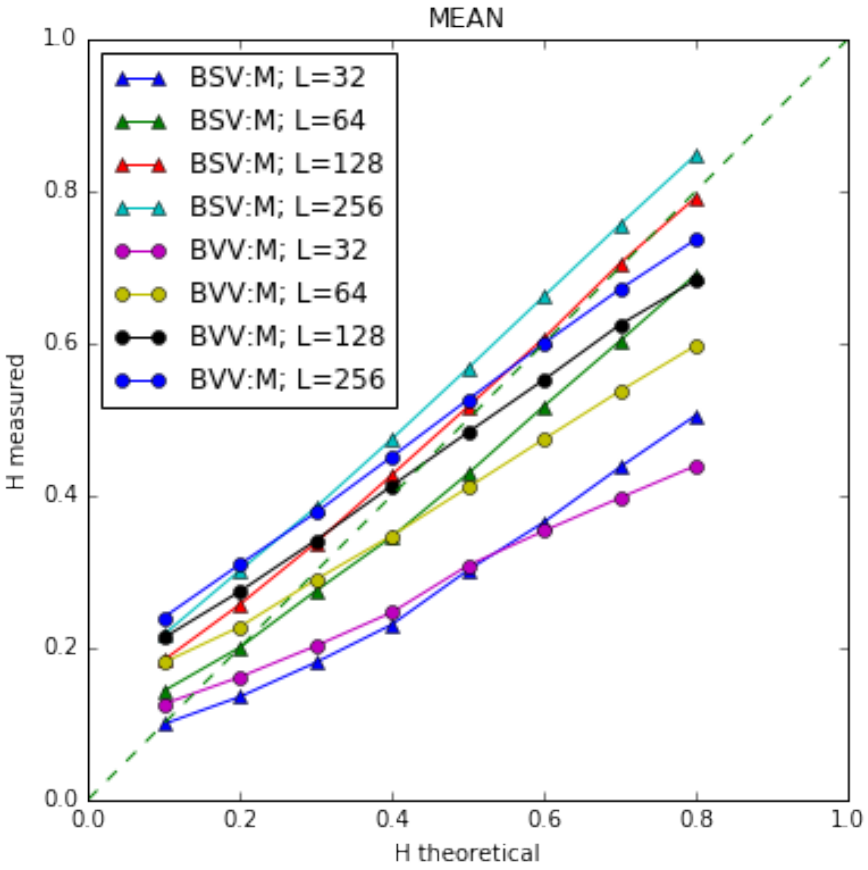


(a) (b)

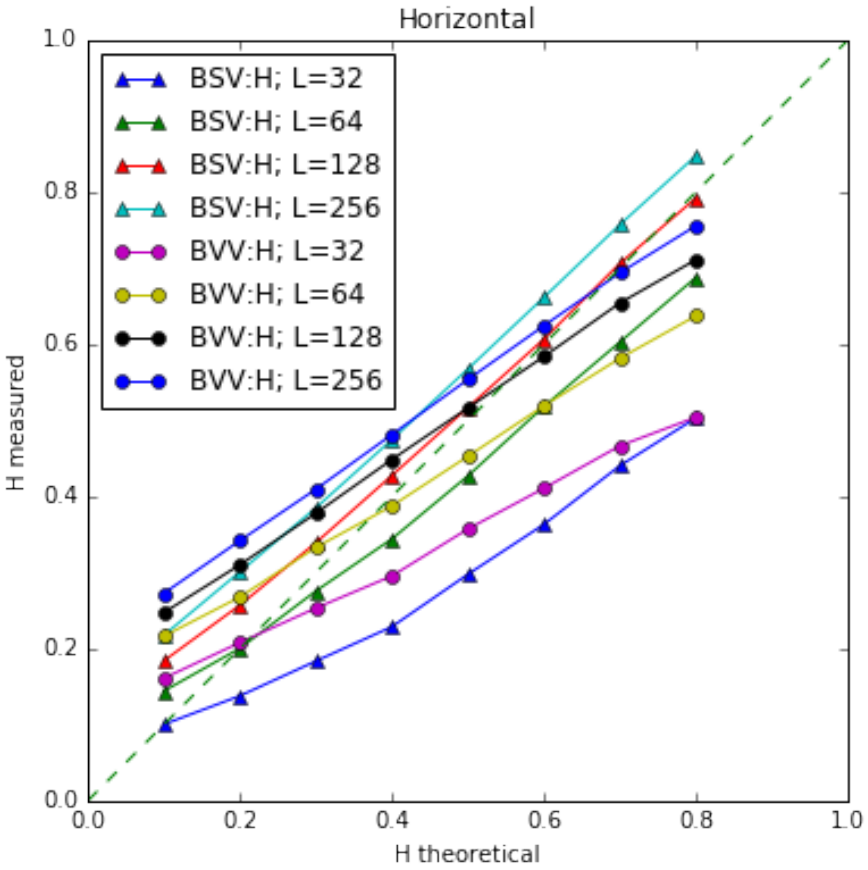
Figure 6.3: Anisotropic fractal example with a theoretical Hurst exponent of 0.3 in the direction 15°. (a) power spectrum of the fractal. (b) the resulted fractal.

In the case of anisotropic fractals we have generated a set of 800 images: 400 with an H of 0.3 in the 0° direction and an H of 0.7 in the 90° direction for sizes 32x32, 64x64, 128x128 and 256x256 (50 per size) and 400 with H of 0.3 in the 15° direction and an H of 0.7 in the 165° direction for the same sizes. The images were analyzed with the BSV and BVV algorithms and the results were reported as plots showing the image size against the computed H (as shown in Figure 6.6). We observe in Figure 6.6a that in case the

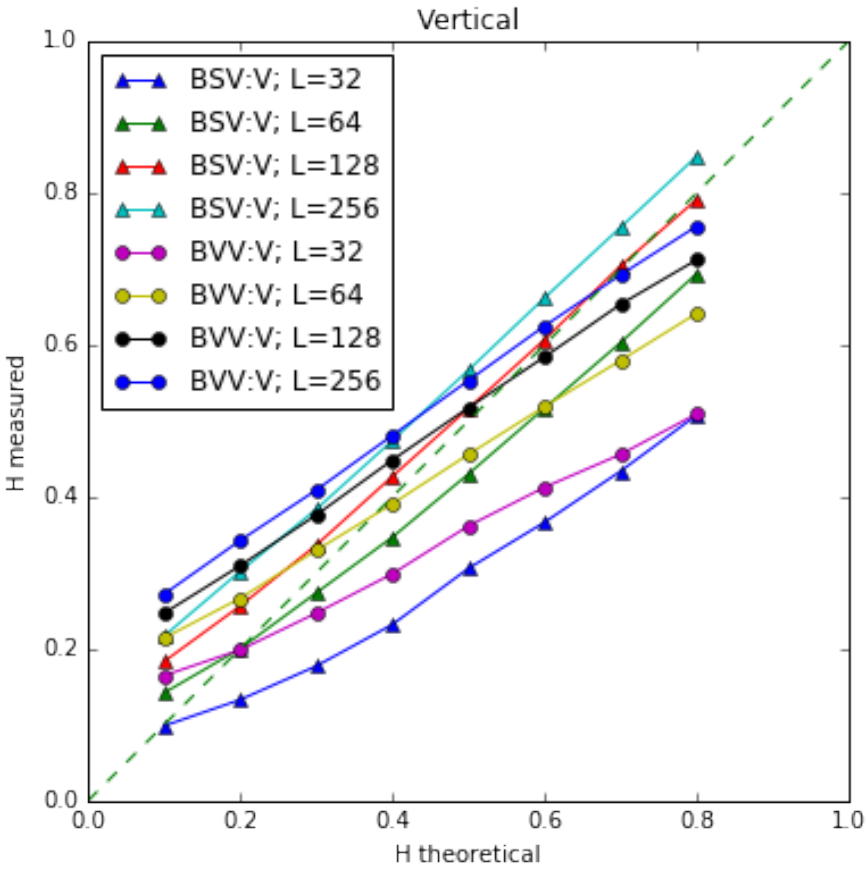
dominant Hurst coefficients are along the main axes, they are largely underestimated with small images (32x32, 64x64), but the algorithms perform better with larger images (128x128 and 256x256) as it was also the case with isotropic fractals. In the case the dominant Hurst coefficients are set along intermediary directions, the mean H values are more stable (see6.6b), since the dominant Hurst coefficient directions are not completely aligned with the main axes' directions, but rather lie in between (i.e., we do not have a feature that measures the H exactly in the 15° and 165° direction as compared to the directions 0° and 90°). However, the algorithms tend to overestimate the real H values with large image sizes. We remind ourselves that the fractal texture algorithms provide only an approximation of the true FD.


MEAN
BSV:M; L=32
BSV:M; L=64
BSV:M; L=128
BSV:M; L=256
BVV:M; L=32
BVV:M; L=64
BVV:M; L=128
BVV:M; L=256
H measured
H theoretical
0.0
0.2
0.4
0.6
0.8
1.0

(a)


Horizontal
BSV:H; L=32
BSV:H; L=64
BSV:H; L=128
BSV:H; L=256
BVV:H; L=32
BVV:H; L=64
BVV:H; L=128
BVV:H; L=256
H measured
H theoretical
0.0
0.2
0.4
0.6
0.8
1.0

(b)



(c)

Figure 6.4: *H*-*H* plots (theoretical vs. computed) of isotropic fractals for different image sizes as measured by the BVV and BSV algorithms horizontally and vertically. Each point on the lines represents a mean of all the Hurst coefficients calculated for the images created with the corresponding parameters. (a) the mean H as computed *H* on the y-axis. (b) the horizontally-computed (0°) *H* on the y-axis. (c) the vertically-measured (90°) *H* on the y-axis.

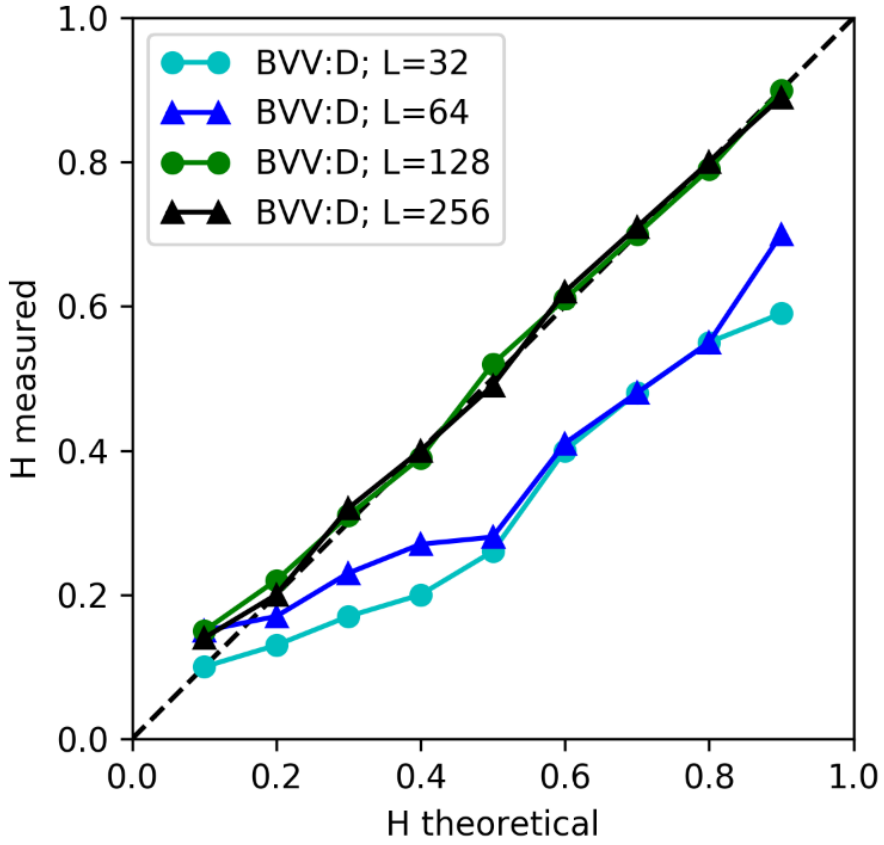
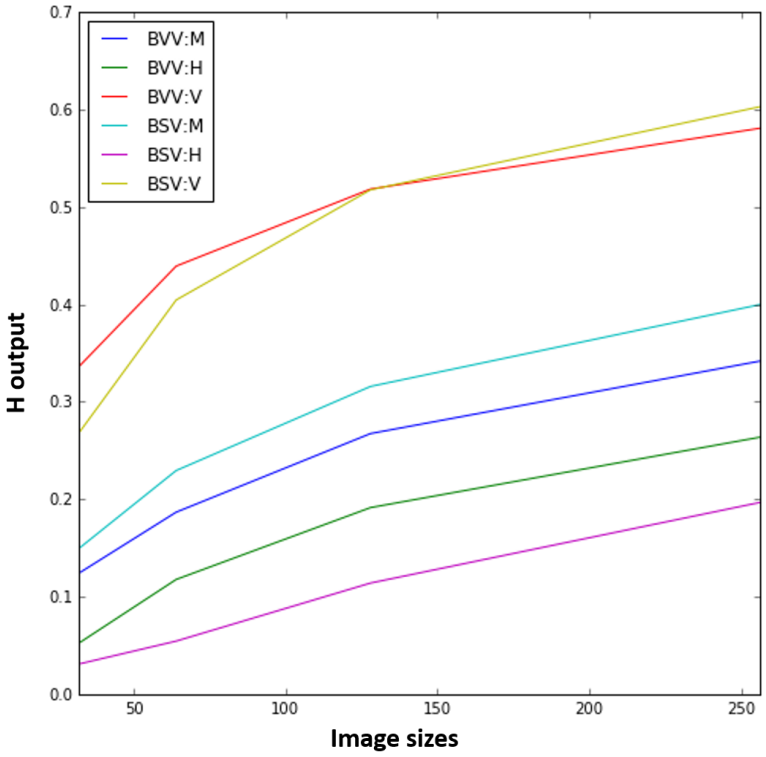
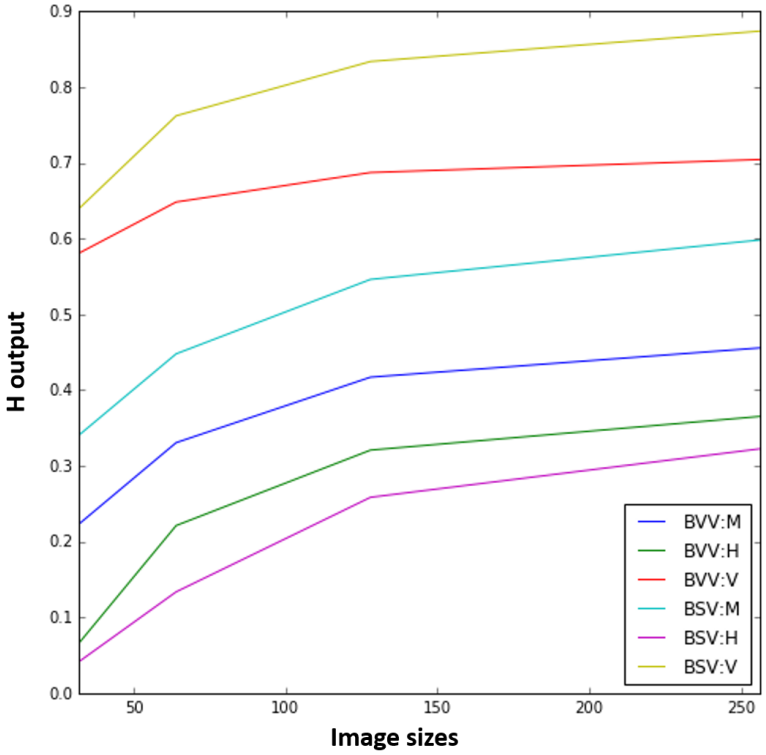


Figure 6.5: *H*-*H* plot of anisotropic fractals for different image sizes as measured by the BVV algorithm diagonally.



(a)



(b)

Figure 6.6: H against image size plot of anisotropic fractals for different image sizes (L) as measured by the BVV and BSV algorithms. Each point on the lines represents a mean of all the Hurst coefficients calculated for the images created with the corresponding parameters. a the H measured for 50 images per size of anisotropic fractals with H of 0.3 in 0° direction and H of 0.7 in 90° direction. b the H measured for 50 images per size of anisotropic fractals with H of 0.3 in 15° direction and H of 0.7 in 165° direction.

6.2 BCV Validation

The BCV algorithm can unfortunately not be validated using the generated images from the previous sections since it is not of a fractal nature. To illustrate the correct detection of the chosen BCV feature, we are using the example code provided with the scikit-image Python library documentation [84]. The execution of the code yields the images in Figure 6.7. Eight patches were selected, four in the sky region (blue rectangles) of the original image and four on the ground (green rectangles). The correlation and dissimilarity of these patches were calculated in turn and the results were drawn in a scatter plot. Each extracted individual patch is also illustrated separately. The patches that are smoother (i.e., sky) produce a lower dissimilarity that tends to zero with increasing smoothness and a higher correlation of the structure that tends to 1 (full correlation) with increasing smoothness. On the other hand, the patches that are rougher produce lower correlation (rougher structure), but higher dissimilarity. Our version of the algorithm computes a homogeneity feature as well, but it is not represented in the Figure, since it is only an inverse of the dissimilarity measure.

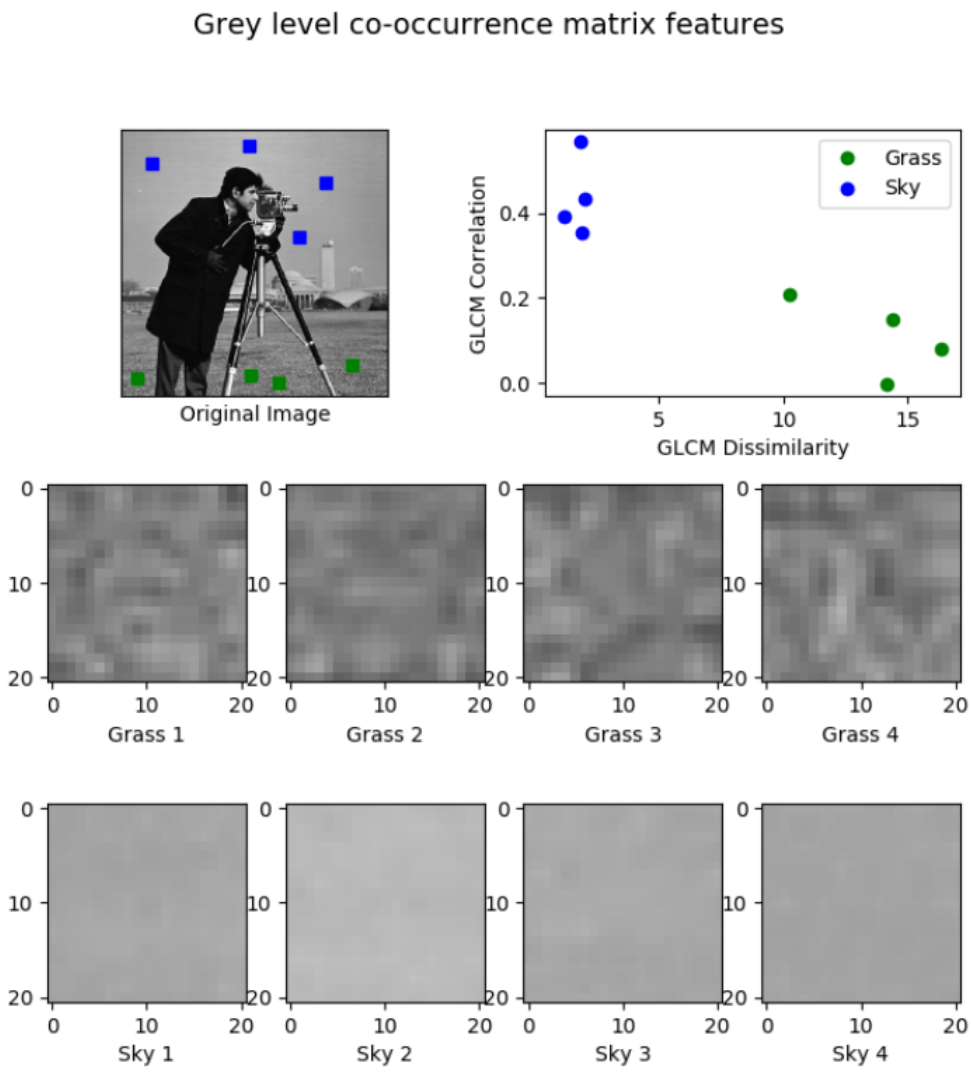


Figure 6.7: Illustration of BCV (GLCM) features measured on a sample image.

6.3 BEV Validation

Similar to the BCV Validation presented in Section 6.2, the validation of the BEV algorithm can not be performed on the artificially generated fractals introduced in Section 6.1 since BEV does not approximate the FD of an image. The BEV is more of a statistical measure directly related to the distribution of intensities in an image. Thus, we approach the validation in a different manner. We know that the BEV is a measure that describes the information complexity in terms of bits of information in a signal. As such, we can generate simple images with known and fixed intensity ranges for which the

BEV value is straightforward. In Figure 6.8 we can see examples of such test images. For example, in Figure 6.8c we observe a 128x128 image that was generated to contain only intensities in the range 0-63 (i.e., 64 values). Therefore, the information content in terms of Shannon's Entropy must equate to 6 bits ($2^6 = 64$). Indeed, when passing this image to the BEV algorithm it measures an information content of exactly 4 bits. The argumentation for the other images covering the other cases in Figure 6.8 is similar.

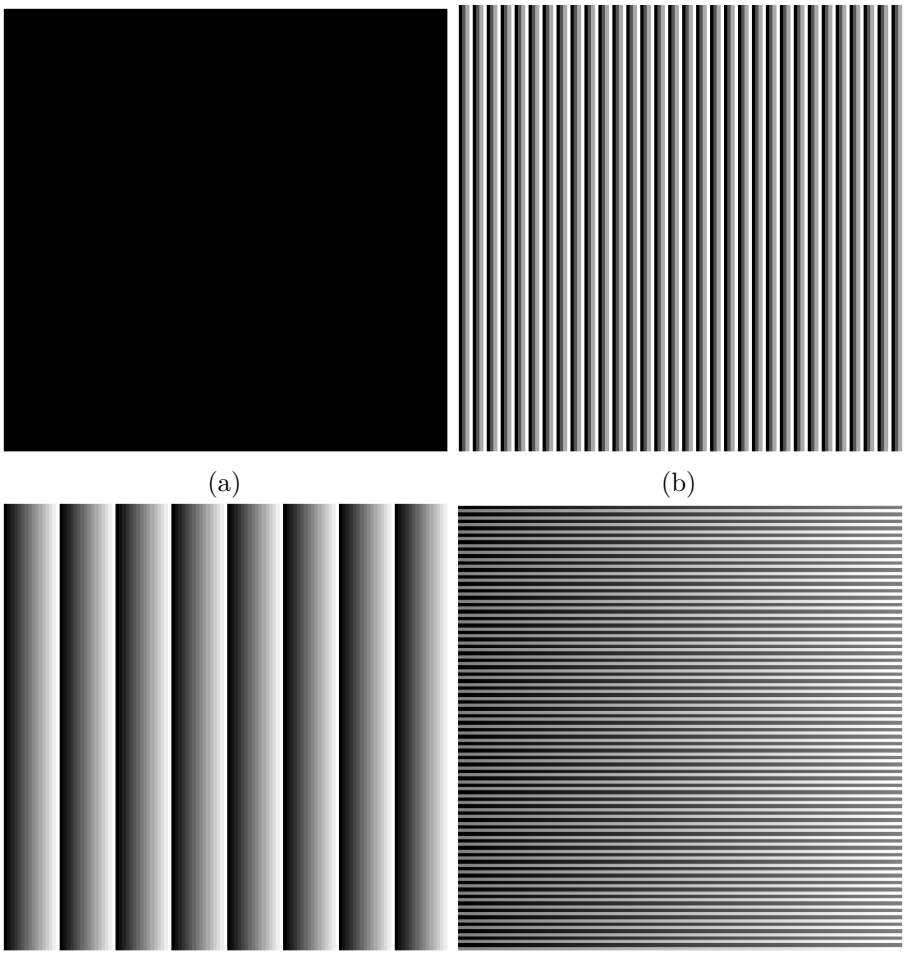


Figure 6.8: Sample images generated with known intensity ranges for the validation of the BEV algorithm. (a) shows a sample image that consists only of zero-intensities with a BEV measure of 0. (b) shows a sample image that consists only of intensities between 0-3 with a BEV measure of 2. (c) shows a sample image that consists only of intensities between 0-63 with a BEV measure of 6. (d) shows a sample image that consists only of intensities between 0-255 with a BEV measure of 8.

CHAPTER 7

Statistical Methods for Hypothesis Testing

In the previous Chapters 5 and 6 we have introduced and validated the algorithms that we employ to characterize the texture of knee TB extracted from the data sets presented in Chapter 4. In this Chapter we present the statistical methods that are used to explore the features provided by the four algorithms. A statistical analysis is needed to determine possible correlations between features, significant differences and other patterns in the data. We will mainly apply statistical tests to find whether our algorithms actually measure features that are different in persons affected by knee OA and in persons without knee OA, or between patients that were healthy at the beginning of a study, but developed OA at a later point. All statistical tests follow a specific pipeline and we will stick to this five-step process when presenting the tests that we have used. The meaning of the following terminology and the procedures involved in each step will become clear when we will describe actual statistical tests that serve certain purposes:

1. **Definition of the significance level of interest.**
   The *significance level* (or α level) is a measure between 0 and 1 that represents the maximally allowed probability of making a mistake when picking a decision in a statistical test.

2. **Definition of the null hypothesis and alternative hypothesis.**
   Every statistical hypothesis test defines a null hypothesis (denoted as $\mathcal{H}_0$, i.e., the hypothesis to be nullified/rejected) and an alternative hypothesis (denoted as $\mathcal{H}_1$ or $\mathcal{H}_a$) that assumes what we want to demonstrate. Following a strategy that stems from the well-known *reductio ad absurdum* principle, the goal is to provide enough evidence to reject the null hypothesis (unwanted effect), so that the alternative hypothesis is assumed and thus the desired outcome is supported.

3. **Calculation of a test statistic.**
   A statistical test generally reduces the whole data set to a single characteristic measure. This measure is called a *test statistic* and is used to support the decision-making process. Based on this measure, the decision whether to reject the $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ or not is later made.

4. **Comparing of the test statistic to known critical values and calculating the p-value.**
   Generally, there is a so-called *probability table* available in literature for each type of test, that holds information about the critical test statistics at different levels of significance and for different sample sizes. The previously calculated test statistic is compared to these values and the critical range is found. Based on this range, the so-called *p-value* is also calculated.

   The p-value shows the probability that the effect observed with the given samples occurred 'by chance' in case the null hypothesis was actually true and we rejected it. A p-value of 0.05 (5%) and below is usually accepted to mean the data is valid. In other words, in that case there would be a probability of less than 5% that we would make a so-called *Type I error* when we reject the null hypothesis; it is very unlikely that we rejected it and it is true in reality. This can be expressed in terms of *confidence levels* as well: *we are 95% confident, that we are not doing a mistake if rejecting the null hypothesis.* The desired confidence limit (also called the *significance level*) α is set at the beginning of a test such that $\alpha + confidence\ level = 100$.

5. **Reporting the results and deciding whether there is enough evidence to support one hypothesis or another.**
   Generally, provided that the calculated p-value is less than the fixed significance level, the $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. However, if the p-value is greater than the set significance level, the $\mathcal{H}_0$ is not accepted, but the result in this case is that there is not enough evidence to reject it, so it is assumed to be true. A statistical test does never accept a hypothesis. This is the reason why the test hypotheses are defined in a way such that the desired outcome lies with $\mathcal{H}_1$. If there is enough evidence to reject $\mathcal{H}_0$, then $\mathcal{H}_1$ is assumed to be true. Generally, so-called effect sizes are also reported along with the p-values. The p-values are often not enough to describe how strong the effects observed are. There are a variety of effect size calculation methods available. Among them, the *Cohen's d* is widely known and employed [85]. However, in this work we aim at showing that there is indeed an effect in the data, but we do not quantify this effect in detail.

All of the methods that we present in this chapter are employed by using available Python 3.6 libraries. Self-written scripts were built by using the modules numpy 1.12.0, scipy 0.18.1, pandas 0.19.2 and scikit-learn 0.18.1. Regression analysis was partly performed in R. The algorithms that we presented in Chapter 5 were implemented in C++, but for our software and for the purposes of the experiments presented in this work, C-object Python wrappers are employed to import and execute the .dlls of the procedures.

7.1 Shapiro-Wilk Test

The first statistical test that we introduce and describe is the Shapiro-Wilk test, which is one of the many normality statistical tests available in the literature [86]. The Shapiro-Wilk normality test was deemed the most powerful (even though it is among the oldest) and stable normality tests in a recent study [87]. Generally such tests are used to check the fulfillment of certain requirements that are imposed by more stringent and complex tests, or by some models as we will see in the following sections.

After setting the desired significance level, the test hypotheses are defined and have the following general form, applicable to all normality tests:

$\mathcal{H}_0$: *the sample in question is drawn from a normally distributed population.*
$\mathcal{H}_1$: *the sample is drawn from a population that is not normally distributed.*

Next, this test calculates its specific statistic, also called the *W-statistic*, that supports the decision making process of whether the underlying data set is drawn from a normally distributed population or not, by applying the following simple formula:

$$W = \frac{(\Sigma_{i=1}^{n} a_i x_{(i)})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}, \text{ with } a_i = \frac{m_i^T V^{-1}}{(m_i^T V^{-1} V^{-1} m_i)^{1/2}} \text{ , and } m = (m_1, ..., m_n)^T \tag{7.1}$$

where $\bar{x}$ is the sample mean, $x_{(i)}$ is the i^{th} order statistic of the sample (i.e., the i^{th} smallest number in the sample), m is a vector containing the expected values of all the order statistics and V is the co-variance matrix of the same order statistics.

In the next and last step, the measured test statistic is matched in the so-called *Shapiro-Wilk table* that shows different critical W-values for different levels of significance and for different sample sizes. A p-value is also approximated in this step from the same table data. If this calculated p-value is smaller than the set α level, then $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$.

7.2 F-test

This statistical test was introduced in the 1920s by Sir Ronald A. Fisher as the *variance ratio* and later received the name of *F-test* in George W. Snedecor's book in honor of the original author [88]. An F-test is any statistical test where the computed test statistic follows a F-distribution. These types of tests are generally used to test two independent samples for equal variances if other statistical tests impose this kind of requirement on the given samples. An F-test can be safely performed if and only if, the samples in question were previously shown to be derived from normally distributed populations. This can be achieved by employing a normality test, such as the one introduced in Section 7.1.

After setting a desired significance level, the test hypotheses are usually defined as:

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2$$
$$\mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2 \text{ (for a two-tailed test)}$$
$$\mathcal{H}_1 : \sigma_1^2 < \sigma_2^2 \text{ (for a lower-tailed test)}$$
$$\mathcal{H}_1 : \sigma_1^2 > \sigma_2^2 \text{ (for an upper-tailed test)}$$

where σ_1^2 and σ_2^2 are the variances of the first and second population. The test statistic is then defined as:

$$F = \frac{s_1^2}{s_2^2} \tag{7.2}$$

where s_1^2 and s_2^2 are the variances of the first and second samples. The larger the deviation from 1, the stronger is the evidence that the samples are drawn from populations of unequal variances. In the next step, the F-statistic is matched in an F-table that holds the critical values for different α levels and samples sizes and the possible ranges of the p-value are found. If the p-value is less than the set significance level, the $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$ and $\mathcal{H}_1$ is assumed as true.

7.3 Levene's Test

The Levene's test is used to test the equality of variances (also called homoscedasticity) among two or more groups [89]. We employ this type of test in our work prior to the application of the K-means clustering algorithm to fulfill the homoscedasticity assumption (see Chapter 8).

The first step to employ this test is to define the level of significance. This is usually 0.05 or 0.01.

In the second step, the test hypotheses are defined as:

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$
$$\mathcal{H}_1 : \sigma_i^2 \neq \sigma_j^2 \text{ for at least one pair } (i, j)$$

In the next step the W-statistic is computed as follows:

$$W = \frac{(N-k)}{(k-1)} \frac{\Sigma_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\Sigma_{i=1}^{k} \Sigma_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2} \tag{7.3}$$

where:

- k is the number of groups,
- N_i is the number of subjects in the i^{th} group,

- N is the total number of subjects,
- $Z_{ij} = |Y_{ij} - \bar{Y}_i|$ with $\bar{Y}_i$ being the mean of the i^{th} group and Y_{ij} being the measured variable of the j^{th} case in the i^{th} group,
- $Z_{i.} = \frac{1}{N_i}\Sigma_{j=1}^{N_i} Z_{ij}$ is the mean of the Z_{ij} for the i_{th} group and
- $Z_{..} = \frac{1}{N_i}\Sigma_{i=1}^{k}\Sigma_{j=1}^{N_i} Z_{ij}$ is the mean of all Z_{ij}.

The computed W-statistic follows an F-distribution with $k-1$ and $N-k$ degrees of freedom. As a consequence, this W-statistic is compared with the corresponding entry in the F-table and if the W-statistic is larger than the critical F-value, the $\mathcal{H}_0$ that the group variances are all equal is rejected in favor of $\mathcal{H}_1$ and $\mathcal{H}_1$ is assumed as true.

7.4 Student's T-test

Introduced in 1908 by William Sealy Gosset under the pseudonym *Student*, the *Student's t-test* is a statistical population hypothesis test. It is used to test two averages to determine whether there are any significant differences between them. Generally, the Student's test is employed in case information about the original population, like standard deviation for example, are not known [90]. We employ t-tests in our work to be able to make statements about how different the distributions of features coming from two different groups of patients are. These groups can be: OA vs. non-OA, BL vs. 30m etc.

Each t-test calculates a test statistic (called *t-value* in this case). The t-value (or t-score) represents the ratio between the difference of the group means and the differences within the groups. The larger the t-score, the clearer the difference between the two respective distributions.

There are three types of t-tests commonly used and each type assumes different null and alternative hypotheses and expects different requirements about the data in question before it can be applied:

1. **One-sample t-test**
   This kind of t-test is used to compare a mean of a sample against a given, known population mean. For example, one could compare the emissions of a car across a period of time against a known maximally allowed limit. In this case the hypotheses are defined:
   $$\mathcal{H}_0 : \mu = \mu_0$$
   $$\mathcal{H}_1 : \mu \neq \mu_0 \text{ (for a two-tailed test)}$$
   $$\mathcal{H}_1 : \mu < \mu_0 \text{ (for a lower-tailed test)}$$
   $$\mathcal{H}_1 : \mu > \mu_0 \text{ (for an upper-tailed test)}$$

where μ is the sample mean and μ_0 is the known limit that we compare against. The next step is to find the t-value which is calculated as:

$$t = \frac{\mu - \mu_0}{s/\sqrt{n}} \tag{7.4}$$

where μ is the mean of the available sample, s is its standard deviation and n the sample size. This t-value is then compared with known critical t-values from a so-called *t-table*. For this, the *degrees of freedom* of the measurement must also be calculated. In the case of the one-sample t-test this is as simple as $n - 1$. The degrees of freedom express how many variables are there available that can vary and thus change the state of the system, or in other words, the outcome. Now, the last step is to calculate the p-value. This is done as following:

$$\begin{aligned} p &= 2 \cdot Pr(T > |t|) && \text{(for two-tailed tests)} \\ p &= Pr(T < t) && \text{(for lower-tailed tests)} \\ p &= Pr(T > t) && \text{(for upper-tailed tests)} \end{aligned}$$

where T is a random variable that comes from a *t-distribution*. If the calculated p-value is smaller than the significance level, then we can reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$. In other words, the data provides enough evidence to safely reject the null hypothesis. However, this type of test assumes that the provided sample is normally distributed. To test for normality, the Shapiro-Wilk test that was presented in Section 7.1 can be used.

2. **Two-sample t-test (also called the *Independent Sample T-Test*)**
   As opposed to the one-sample t-test, the two-sample t-test compares two means of two independent samples. For example, these tests can be used to compare the mean production of two different (so independent) gardens of apple trees. In this case, the test hypotheses are defined as follows:

$$\begin{aligned} &\mathcal{H}_0 : \mu_1 = \mu_2 \\ &\mathcal{H}_1 : \mu_1 \neq \mu_2 \text{ (for a two-tailed test)} \\ &\mathcal{H}_1 : \mu_1 < \mu_2 \text{ (for a lower-tailed test)} \\ &\mathcal{H}_1 : \mu_1 > \mu_2 \text{ (for an upper-tailed test)} \end{aligned}$$

   where μ_1 is the mean of the first sample and μ_2 is the mean of the second sample. Similarly to the pipeline presented for the one-sample t-test, the t-value calculation follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{ with } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{7.5}$$

   where n_1 and n_2 are the sizes of the first and second sample, $\bar{x}_1$ and $\bar{x}_2$ are the means of the first and second sample and s_1^2 and s_2^2 are the variances of the first

and second samples. Again, the obtained t-value is compared with the t-table using the known degrees of freedom $(n_1 + n_2 - 2)$ and a p-value is approximated or calculated in the same way as was the case for one sample t-tests. If the p-value is smaller than the originally-set significance level, then $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. However, generally a two-sample t-test also assumes a set of facts about the underlying data before it can be applied:

a) Both samples must be derived from normal distributions. This is checked using the Shapiro-Wilk test from Section 7.1.

b) Both samples should have equal variances. This is checked using an F-Test as described in Section 7.2.

3. **T-tests for paired samples (also called the *Dependent Sample T-Test*)** This version of a t-test is a special case of the two-sample t-test in that the compared samples are now dependent. For example, such a test could be applied if the production of an apple tree garden is measured, then the trees are treated with some kind of chemical boosting solution and the production is recorded again. The paired-samples t-test is capable of detecting the difference between dependent (same trees before and after the treatment) samples. In the case of the dependent samples t-test, the statistical hypotheses are defined as follows:

$$\begin{aligned}
&\mathcal{H}_0 : \mu_d = D_0 \\
&\mathcal{H}_1 : \mu_d \neq D_0 \text{ (for a two-tailed test)} \\
&\mathcal{H}_1 : \mu_d < D_0 \text{ (for a lower-tailed test)} \\
&\mathcal{H}_1 : \mu_d > D_0 \text{ (for an upper-tailed test)}
\end{aligned}$$

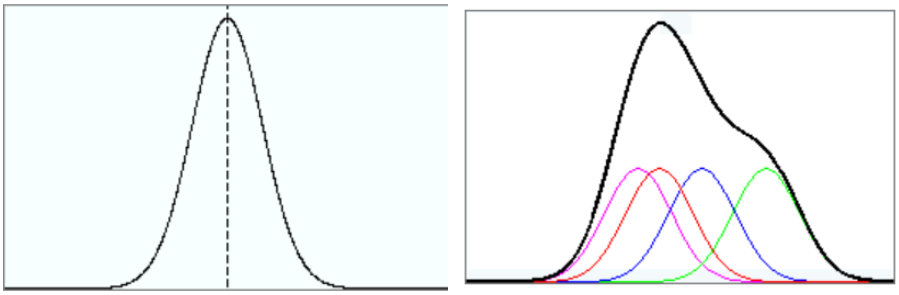
where μ_d is the *mean difference*, or the mean of the distribution of sample differences, between the paired samples. In other words, $\mathcal{H}_0$ assumes that there is no difference between the measured dependent samples. The calculation of the t-value is then done in the following manner, similarly to the previous cases:

$$t = \frac{\bar{d} - D_0}{s_D/\sqrt{n_D}} \tag{7.6}$$

where $\bar{d}$ is the mean of the difference sample, the D_0 is the set difference that we are comparing against (usually 0), s_D is the standard deviation of the difference sample and n_D is the size of the difference sample. This t statistic is then compared to the values in a t-table and with its help the range or the exact value of the p-value may be calculated as previously done. If the p-value is below under the specified level of significance, $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. As opposed to the first two types of t-tests, the dependent samples t-test assumes that the samples' difference is normally distributed. This assumption is easily checked with the help of the Shapiro-Wilk test presented in Section 7.1.

7.5 Analysis of variance (ANOVA)

The *analysis of variance (ANOVA)* was introduced by R. Fisher as a generalization of a t-test to more than two groups. ANOVA is used to compare three or more means of variable distributions for statistical significant differences. We use this type of test in our work to be able to make statements about the differences of the distributions' averages of the same patients across the duration of the MOST study. The idea behind it is that the population means are reflected in the variances of the samples. If the samples have the same mean and variance, then the joint sample distribution will be just a superposition of identically placed distributions (as shown in Figure 7.1a). However, if the means of the samples are different, the joint distribution becomes broader (as shown in Figure 7.1b) and this can be detected by the ANOVA hypothesis test.



(a) (b)

Figure 7.1: (a) the joint sample distribution of smaller distributions with the same mean and variance. (b) the joint sample distribution of smaller distributions with equal variances but different means; the shape of the joint distribution in this case is impacted by the variance of the samples [91].

There exist different versions of ANOVA based on the number of independent variables taken into account. One-way (or one-facor) ANOVA is computed only with respect to a single independent variable. For example, three groups of patients are treated with a different approach each and the differences of the mean responses are then compared using ANOVA. In this example, the used medicine per se is the independent variable and it has three levels: first medicine that went to the first group, second medicine for the second group, and the last medicine to the third group. A two-way ANOVA would, in comparison, take into consideration more than one independent variable. In the example above, the second independent variable could be the life style of the treated patients with three levels: inactive, mildly active, very active. However, to apply these versions of ANOVA, the samples drawn must be completely independent from each other. The ANOVA that are applied using this logic, are called *between-subjects ANOVA* or *between-samples ANOVA*.

As with the other presented statistical tests, in case of the ANOVA a null and an alternative hypothesis must also be defined, after setting a desired significance level, in

the following way:

$$\begin{aligned}\mathcal{H}_0 &: \textit{ the means of the groups are all the same.}\\ \mathcal{H}_1 &: \textit{ the means of the groups are different.}\end{aligned}$$

The next step is to compute the test statistic of the ANOVA test, which follows an F-distribution, just as with the F-test presented in Section 7.2. This test statistic is also called *the variance ratio (V.R.)* and is computed as following:

$$VR = \frac{Among - groups\ MS_b}{Within - groups\ MS_w} \tag{7.7}$$

where the MS are the mean-squared distances computed from the data. The two MS stem from the total variability of measurements that is partitioned into among-groups variability and within-groups variability (as shown in Figure 7.2). The among-groups $M.S._b$ is built by computing the deviations of all group means to the total groups' 'grand' mean (SS_b) and dividing it by the degrees of freedom of the independent variable (DF_b):

$$\begin{aligned}MS_b &= \frac{SS_b}{DF_b}\ ,\ with\\ SS_b &= \Sigma_{i=1}^{k} n_i(\bar{x_i} - \bar{x})^2\ ,\ and\\ DF_b &= k - 1\end{aligned} \tag{7.8}$$

where k is the number of levels of the independent variables (in Equation 7.8, k would be equal to 3 and thus the degrees of freedom would be $3 - 1 = 2$), n_i is the number of measurements in the i^{th} group, $\bar{x_i}$ is the mean of the i^{th} group and $\bar{x}$ is the mean of all groups together. The within-group MS_w is obtained by first calculating the within-group deviation SS_w and by dividing it by the degrees of freedom, which is total number of measurements *minus* the number of groups (see Equation 7.9).

$$\begin{aligned}MS_w &= \frac{SS_w}{DF_w}\ ,\ with\\ SS_w &= \Sigma_{i=1}^{k}\Sigma_{j=1}^{n_i}(x_{ji} - \bar{x_i})^2\ ,\ and\\ DF_w &= \Sigma_{i=1}^{k}(n_i) - k\end{aligned} \tag{7.9}$$

The resulting F-statistic (VR) is then compared with an F-table and the p-value range is approximated. If the p-value is below the significance level that we selected initially, $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. To compute this test safely, the group samples must come from a normally distributed population, a fact that can be tested prior to the employment of the actual ANOVA with the Shapiro-Wilk test that we have already discussed in Section 7.1. Also, the variances of all the groups must be equal. In theory, we could test for this using the F-test presented in Section 7.2, however the need of repetition of the F-test will inflate the real significance level of the entire test as a whole. For this reason, other,

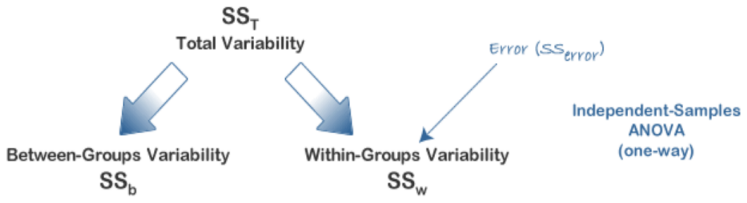


Figure 7.2: Total variability of measurements partitioning for between-subjects ANOVA [92].

related tests must be used. The O'Brien variance homogeneity test is a more complex extension of the F-test, but we will not go into details about this specific test. Another example is the Levene's test that was introduced in Section 7.3.

As mentioned above, independent-samples ANOVA works only if the analyzed groups are independent from one another. However, in our work we will deal with measurements recorded for the same patients across several visits (MOST study, as shown in Chapter 4). In other words, the samples will be dependent of each other. For this reason, the ANOVA approach that was described up to this point is not suited for the significance test. The *repeated-measures analysis of variance (RMANOVA)* deals with dependent data. The F-statistic is computed in a similar manner in this case:

$$F = \frac{MS_{conditions}}{MS_{error}}, \tag{7.10}$$

where $MS_{conditions}$ is the same as MS_b from independent-measures ANOVA and MS_{error} is the equivalent of the independent-sample ANOVA within-subject source of error (MS_w) but smaller (as shown in Figure 7.3). The smaller error is due to the fact that the measurements are produced from the exact same sources at each of the independent variable's levels. $MS_{conditions}$ is calculated as the previous MS_b and MS_{error} can be calculated directly (complicated formula), but it is usually faster to extract it from already-existing knowledge of the other sums. For example (as shown in Figure 7.3):

$$\begin{aligned} MS_{error} &= \frac{SS_{error}}{DF_{error}} \ , \ with \\ SS_{error} &= SS_w - SS_{subjects}, \ and \\ DF_{error} &= (n-1)(k-1) \end{aligned} \tag{7.11}$$

where n is the number of subjects (participants) and k is the number of levels of the independent variable (usually time). As mentioned, SS_w is obtained as with the independent-measures ANOVA and to calculate the $SS_{subjects}$, the subjects of the test are treated as levels of another independent variable which is *subject*. From this follows:

$$SS_{subjects} = k \cdot \Sigma_{i=1}^{n}(\bar{x}_i - \bar{x})^2 \tag{7.12}$$

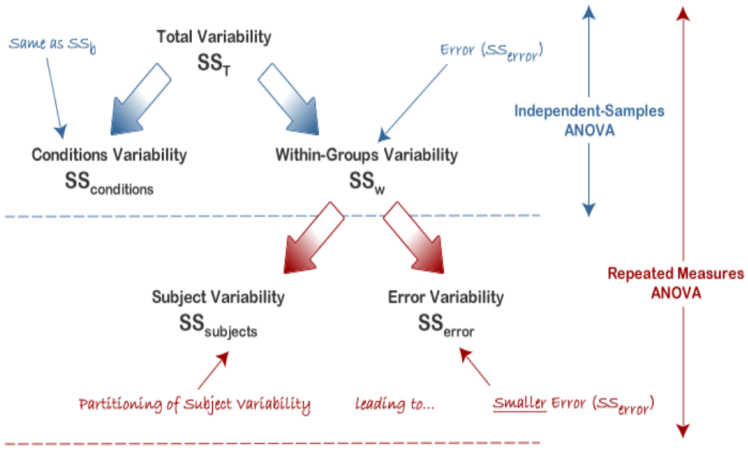


Figure 7.3: Total variability of measurements partitioning for dependent-subjects ANOVA (repeated-measures ANOVA) [92].

where k is the number of levels of the usual independent variable *time*, $\bar{x_i}$ is the mean measurement across visits of the i^{th} subject and $\bar{x}$ is the total mean of the visit means (the grand mean). After SS_w and $SS_{subjects}$ are computed, the SS_{error} is obtained as shown in Equation 7.11. The F-statistic can then be calculated as shown in Equation 7.10. The F-statistic is then compared against an F-table as usual and the test procedure from this point on is identical to any other statistical test that we presented. The requirements of the ANOVA apply for RMANOVA as well.

The tests used for checking homogeneity (also called *sphericity* or *homoscedasticity*) of the data, which is a crucial requirement for ANOVA, will not be described here in detail due to their more complex nature. For every statistical test that we employ in this work we make use of pre-programmed Python libraries that offer the required functionality.

The problem with ANOVA tests is that they are capable of recognizing significant statistical differences among groups, but are not capable of indicating where exactly, between which groups, do these differences occur. For this matter, generally so-called *post-hoc* tests are employed after receiving confirmation of a significant difference from ANOVA. An example of this would be the *Tukey's Honest Significant Test (HSD)* [93], which is a generalization of a t-test for multiple comparisons. If we were to apply more pairwise t-tests, the real significance level of the end result would be much higher than the initially specified value (5%). In other words the risk of a Type I error increases if increasing the number of comparisons. This phenomenon is known as *Family-Wise Error Rate* or *Alpha Inflation* and depending on the number of repeated comparisons, the new, adapted significance level is computed as following:

$$\bar{\alpha} = 1 - (1 - \alpha_0)^m \tag{7.13}$$

where α_0 is the initially-specified significance level (usually 5%) and m is the number

of total comparisons or the number of total hypothesis pairs tested. Tukey's HSD is one of the approaches that corrects these inflations. Another example would be the Holm-Bonferroni method [94].

CHAPTER 8

Model Building for Classification and Early Prediction of OA

In this chapter we present our approaches at building and training artificial neural networks that are able to learn from the features provided by our algorithms from labeled image data and make a prediction on new, unknown data. We will thus describe the process of significant feature selection and we will give details on the chosen network and model types.

8.1 K-means Clustering

The K-means clustering algorithm is one of the simplest and earliest form of clustering. It was first proposed in 1957 at Bell Labs, but publicly presented only in 1982 [95]. It is based on the widely-used approach of vector quantization. The goal of the procedure is to divide given data into k partitions (also called *clusters*) in an unsupervised manner (i.e., the algorithm is able to separate and classify unlabeled data). To achieve this, two important steps are repeated in a loop, after k initial random points have been chosen from the data as initial *centroids*, until intra-cluster variance is minimized and inter-cluster variance maximized:

1. **Assignment step:** Each data point in the set is assigned to the nearest centroid. This is done by minimizing the Euclidean distances between the point and the centroid. In other words a partitioning of the observation is done based on the Voronoi diagram spanned by the centroids.

2. **Update step** After each assignment step, new means are calculated for each cluster as the new centroids.

The algorithm stops when no assignments change from one loop to another. In general it is very hard to choose the right number of clusters if no external constraints are available. Usually other algorithms (such as DB-scan for instance) are employed in those cases that first estimate the number of clusters. However, in our case, we have the previous knowledge from our data sets that there must be only two groups (clusters) of patients: with and without OA. At the same time we assume that our data is normally distributed, since it is real data. This fact is confirmed by Shapiro-Wilk tests. The other two assumptions of the K-means algorithm is that the feature variances are equal and that the number of the subjects in all groups are approximately the same. We test the former condition with the Levene's test as presented in Section 7.3 and the latter is taken care of by the SMOTE technique as shown in Chapter 4 .

In this work, we use K-means clustering as a first step in our experiment. This step indicates whether the separation in our feature space is extremely good or poor. If there are clearly separable clusters in the hyperspace, the K-means clustering will produce a high classification score. If the boundary between the two groups, i.e., OA and non-OA, is not linear and does not have a large margin (i.e., distance to the closest points in each group from the separation boundary), the K-means classifier will show a poor performance.

8.2 Support Vector Machine (SVM)

As opposed to the K-means, the SVM is a supervised learning technique that is trained to classify data based on a labeled training set. The method searches the best separation boundary between the two groups of data (as shown in Figure 9.5). The separating hyperplane can be defined by:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0, \tag{8.1}$$

where $\boldsymbol{w}$ is the coefficient/weight vector (which is perpendicular to the hyperplane) and $\boldsymbol{x}$ is the vector of observations with each entry x_i being multidimensional. Ideally, the hyperplane will be at maximum distances from the closest points (referred to as the *Support Vectors*) of the two classes at the same time in case of linearly separable data. This distance is called the *margin* of the system and is equal to $\frac{1}{||\boldsymbol{w}||}$, a fact derived from the equations of the planes passing through the support vectors by computing a perpendicular distance from a point on one of them to a point on another:

$$\begin{aligned} P_{-1} &: \boldsymbol{w}\boldsymbol{x} + b = -1 \\ P_{+1} &: \boldsymbol{w}\boldsymbol{x} + b = 1 \end{aligned} \tag{8.2}$$

In other words, to maximize the margin, we must minimize the $||\boldsymbol{w}||$ so that every observation is correctly classified, which is the same as minimizing $\frac{1}{2}||\boldsymbol{w}||^2$ so that:

$$y_i(x_i \cdot \boldsymbol{w} + b) - 1 \geq 0, \forall i, \tag{8.3}$$

where y_i is the known label (+1 or -1) of the data vector x_i. This formulation facilitates the later optimization by means of Quadratic Programming (QP) in the coming steps.

The maximization is done by introducing the so-called *Lagrangian multipliers* $\alpha \geq 0$:

$$L_P = \frac{1}{2}||\boldsymbol{w}||^2 - \boldsymbol{\alpha}[y_i(x_i \cdot \boldsymbol{w} + b) - 1], \tag{8.4}$$

which after short manipulations leads to:

$$L_P = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{N} \alpha_i y_i (x_i \cdot \boldsymbol{w} + b) + \sum_{i=1}^{N} \alpha_i. \tag{8.5}$$

where N is the number of observations. In the next step, Equation 8.5 must be minimized with respect to $\boldsymbol{w}$ and b, but maximized with respect to $\boldsymbol{\alpha}$ (with $\alpha_i \geq 0 \ \forall_i$). Note that $\alpha_i = 0$ only in the case of the support vectors. Differentiating with respect to $\boldsymbol{w}$ and b and replacing in Equation 8.5 finally leads to:

$$L_D = \sum_{i=1}^{N} -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i H_{ij} \alpha_j, \tag{8.6}$$

with $H_{ij} = y_i y_j x_i x_j$. L_D is only dependent on $\boldsymbol{\alpha}$ and must be maximized so that $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$. This can be solved by means of a quadratic optimization and $\boldsymbol{\alpha}$ is returned. Earlier, when setting the derivative of L_P with respect to $\boldsymbol{w}$ to 0, an expression for it is found:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i x_i. \tag{8.7}$$

By replacing the determined α_i in this equation, we find $\boldsymbol{w}$. To find b, one must have a support vector verify Equation 8.3 and the calculation of the constant b becomes straightforward. At this point, the variables $\boldsymbol{w}$ and b are known and with them the optimally separating hyperplane is also found. A limitation of the SVM is that in its original form it can not make any smart selection of the features it uses for finding the optimal hyperplane. However, there is an approach called *SelectFromModel*, which treats the learned weights $\boldsymbol{w}$ as feature importances and based on that it iteratively selects an increasing number of best features, which then are used for model building. The features are added to the feature pool one at a time until all are consumed and for each subset, an accuracy is computed. In this way, the best combination of features (in terms of their importance) can be found.

The SVM generally assumes independent and identically distributed (IID) data. However it has been recently shown that the SVM is also stable under dependent data [96]. We assure that our data is identically distributed by normalizing the data before the training of the model. Subsequently we train a linear SVM that attempts to learn the differences between two independent groups of patients in the discrimination tasks with the Portugal and MOST data sets: OA vs. non-OA. We train a similar classifier for the early prediction task, but with dependent data, since we compare OA vs. non-OA features coming from the same patients that were healthy at BL, but became ill over the duration of the study. The outcomes are presented in Chapter 9.

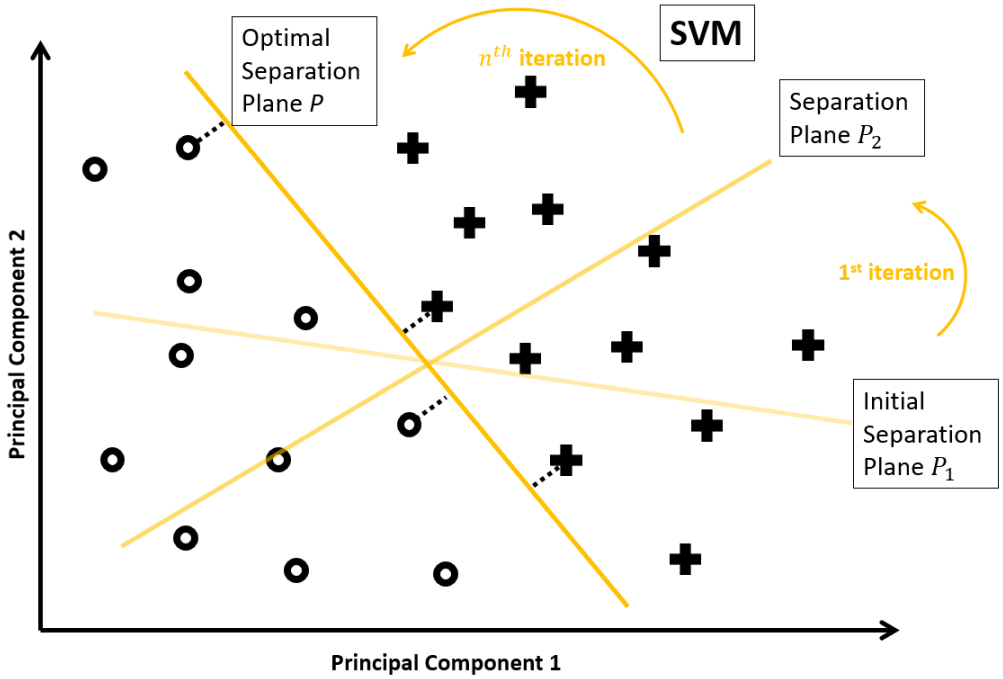


Figure 8.1: SVM procedure illustration. Planes P_1 and P_2 separate the two classes, but not in an optimal manner, i.e., the squared distances to the plane are not maximized. The optimal separation is done with plane P.

8.3 Random Forests

The Random Forests are a statistical learning method based on ensembling and are generally used for classification or regression [97]. It is based on a bagging approach since at training, many trees are 'grown' based on a random subset of the initial observations. The first step in a classification task is to select a random subset of observations with known labels. This is the preparation step for learning. In the next step, each feature is verified for its separation power based on a simple threshold, against the ground truth. In other words, the feature that separated the subset best (with the least false positives and negatives), is considered the decision feature at the current splitting node. In the third and the next steps, the obtained subsets are split again based on other features that perform the best until a stopping criterion is met. All the steps are repeated until the desired number of trees are trained. One of the stopping criterion that are used, is called the *Gini Impurity* which aggregates the misclassification rates of the nodes of the trees. The goal is to minimize this impurity, but once there is no change in the impurity after splitting, the stopping criterion was reached. The magnitude of the update in Gini impurity that takes place at each node is also a direct measure of feature importance. Another stopping criteria would be a minimum number of observation that a subset should contain after splitting. This is usually given either relative to the size of the entire

data set or relative to the size of the selected subset.

Once the forest is trained, new inputs can easily be classified by voting. All the trained trees will return an output whether the new observation belongs to class one or class two. Not all the trees are built the same and thus they will not return the same result. As a consequence, generally a majority voting is applied to determine the new observation's class. The single real drawback of the Random Forests is the model size that can get extremely large extremely fast. A prediction in that case becomes slow.

In this work we employ Random Forests with a view of eliminating insignificant or correlated features from the feature space. This procedure is illustrated in Chapter 9.

8.4 Principal Component Analysis (PCA)

The earliest form of Principal Component Analysis (PCA) was introduced in 1901 by Pearson [98], but was later reintroduced under different names with slight improvements or additions, such as eigenvalue decomposition, singular value decomposition, etc. The PCA is a statistical procedure that 'rotates' the system of coordinates of possibly correlated variables so that the new system is more 'efficient'. After the so called *orthogonal transformation*, the new system will represent the observations as linearly uncorrelated. These new, transformed variables are called *principal components*.

The system rotation takes place in such a manner that the projection on the axis of the first principal component accounts for the largest variability in the data. The rotation vector (also called *the loadings*) for the first principal component is computed as follows:

$$\boldsymbol{w}_{(1)} = argmax\Big\{\frac{\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}\Big\} \tag{8.8}$$

where $\boldsymbol{X}$ is the mean-centered data matrix containing all the observations on the rows and the features along columns. The quantity in the curly brackets is also referred to as the *Rayleigh quotient*. With known $\boldsymbol{w}_{(1)}$, the transformed coordinates of an observation can be given as:

$$t_{1(i)} = \boldsymbol{x}_{(i)} \cdot \boldsymbol{w}_{(1)} \tag{8.9}$$

which are called the *scores*. The back-transformation is achieved by multiplying the score t again with the weight vector $\boldsymbol{w}$. The next components also account for the 'next largest' variability in the data that is not already covered by the first component. To determine these, the previously computed components must be subtracted from the original data matrix, which returns a new matrix. Using this matrix, the Equation 8.8 is applied again to find the next rotation weights corresponding to the component and finally computing the corresponding scores and so on.

The PCA procedure is sensitive to the scaling of the original observations. Due to this, a standardization of the data is a compulsory preparation step prior to the computation of the transformation.

In this work we use PCA for visualization purposes. The dimensionality reduction allows a projection of the feature space into 2D. We use this procedure to visualize the results of the K-means clustering.

CHAPTER 9

Results

In this chapter we present the results that we obtained on both the data sets that we presented in Chapter 4. Based on the results of statistical tests we investigate whether we can reject or accept the hypotheses that we have previously defined given the available data samples. We also discuss which feature (set) is better for discriminating, predicting, and tracking the progression of OA based on the accuracy measures of the models that we have built.

9.1 Portugal Data

Before we applied any sort of statistical test, we first investigated the texture features for influence from other confounding variables, such as *age, BMI, voltage (kV), gender,* or *exposure (mAs)*. After a simple visual inspection of the plots shown in Figure 9.1a we notice slight correlations between, for example, BSV:H and BMI for all ROIs. The hypothesis is confirmed if computing pairwise Spearman correlation coefficients and testing them for being significantly different from 0 using a t-test. Most of the features showed significant correlations with age or BMI, which together with gender are known risk factors for OA incidence [99, 100]. As a consequence, we decided to adjust all our features, as we are only interested in the texture factor contained in each feature, uninfluenced by any other clinical covariates or machine parameters. To compute the correct adjustment parameters we employed a multi-linear regression (MLR). We assumed the following general model:

$$feat = BMI + age + gender + kV + mAs + feat_0 \tag{9.1}$$

which leads to:

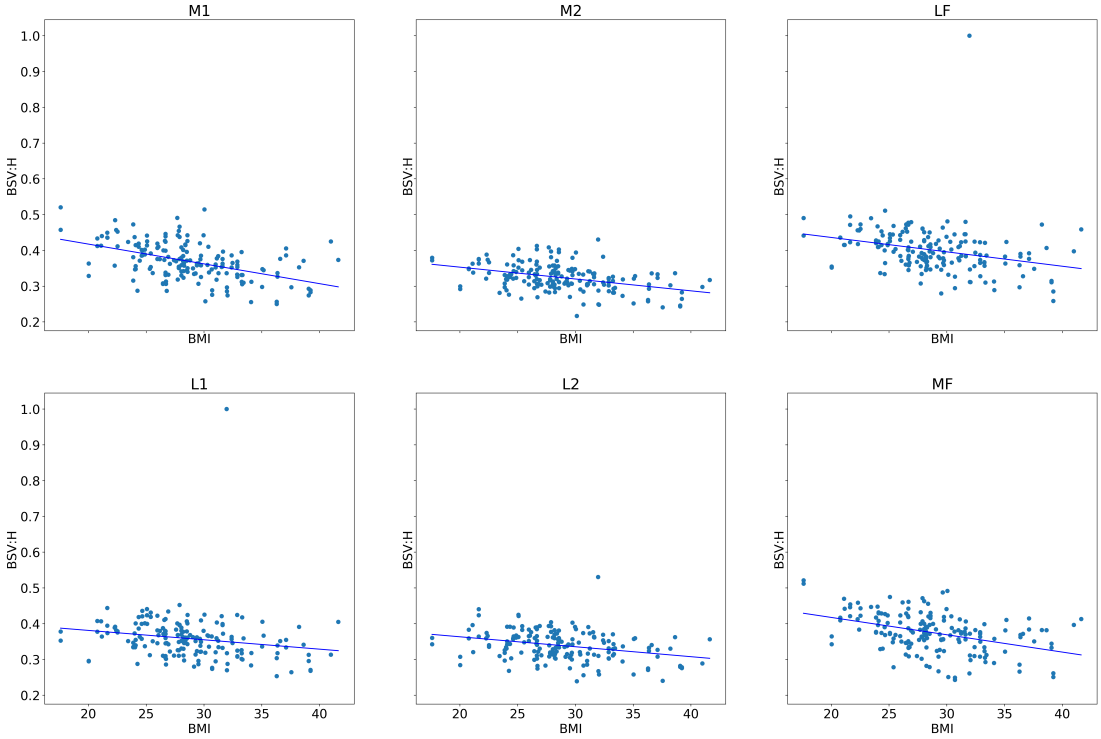
$$feat_0 = feat - BMI - age - gender - kV - mAs \tag{9.2}$$

where $feat_0$ is the uninfluenced BSV, BVV, BEV or BCV feature. In other words, using MLR we find the contributions of BMI, age, kV and mAs (gender is already taken care of since in the experiment based on the Portugal data we only selected females) and we subtract these contributions from the originally calculated features. The result can be observed in Figure 9.1b based on the example of BMI correlation. After adjustment, the computed Spearman correlation coefficients do not differ significantly from 0 anymore. Through this adjustment we want to ensure that the effects that we find based on the computed features come solely from the bone texture and are not influenced by anything else.

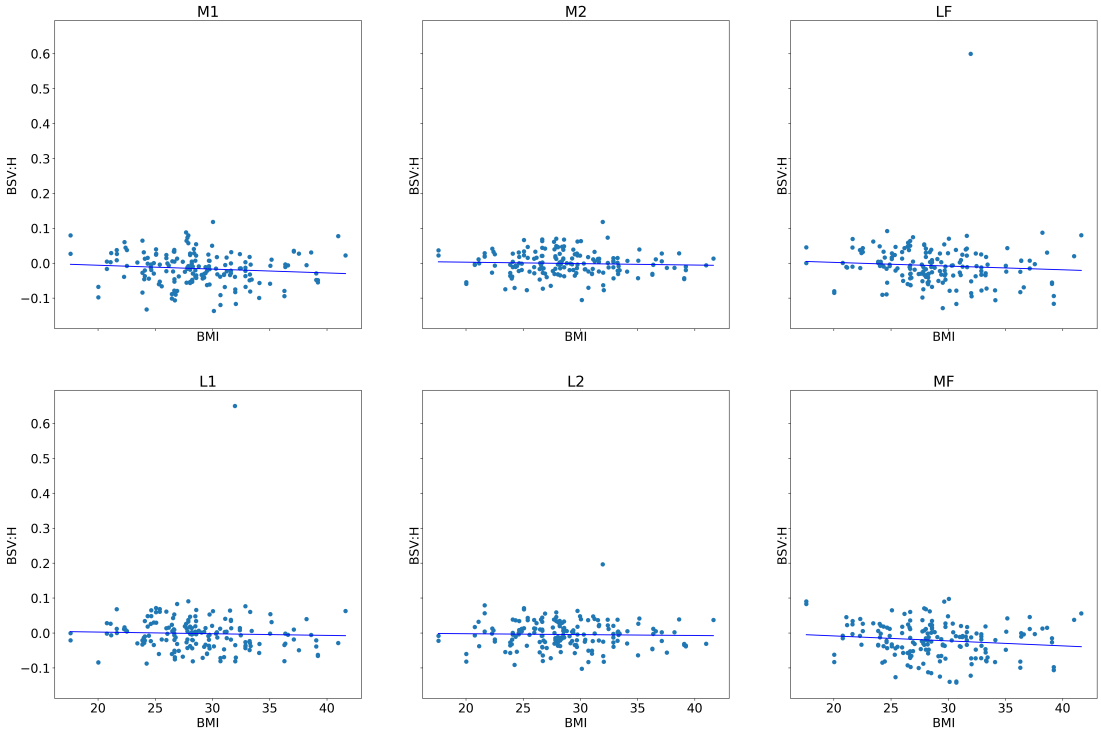
After feature adjustment for confounding influences, we first employed an unsupervised learner to investigate its performance without any training. We employed a K-means classifier, as the features were found to be homoscedastic with a Levene's test (p-value $<$ 0.05). The results after using the K-means classifier can be seen in Figure 9.2. The best classification accuracy of 64% is obtained if using all 126 features in combination. To improve this result we decided to employ a type of feature selection based on Random Forests. In Figure 9.3 we observe the top ten features sorted by ascending importance according to the random forest model. Note that in the case of the BEV only six features were given in total. We observe that the most important feature in all cases is the one measured at the MF ROI, which is the femural medial compartment. This is somewhat expected since the medial compartments theoretically bear the highest loads . Considering only these top-ten scoring features for the K-means algorithm calculation, the results improved slightly (as shown in Figure 9.4). The classification accuracy if using all selected 36 features (top-ten from each of BSV, BVV, BCV and six from BEV) has reached 71%, while the classification scores per feature group did not improve significantly.

Considering supervised learning with a linear SVM and using all 126 features, the classification scores improve significantly for all feature groups over the scores obtained with unsupervised learning (as shown in Figure 9.5). The best separation of cases from controls is achieved by taking all the features together, which produces a classification score of 84% in terms of ROC-AUC. A model configuration can also be found along the ROC-curve that reaches the highest sensitivity, specificity, precision, and accuracy of 73%, 85%, 83%, and 79% for this patient separation task (see the green point in Figure 9.5). Employing the same linear SVM model trained the top performing features according to the Random Forests approach does not improve the classifications significantly over the models without a prior feature selection step (as shown in Figure 9.6). In general the classification score and the other best-configuration metrics even decrease. This could be due to losing information while reducing the feature space. The SVM models were validated by a 100-fold cross-validation procedure, each fold randomly splitting the data into a training and a test set with the test set being 15% the size of the whole data set.

In another approach we have tested each feature individually whether its measurements are significantly different between the OA and the non-OA groups. For this we have used two-sample two-sided t-tests with a significance level of 0.05. Prior to the application of the tests, we must consider the requirements for a correct application. The normality of



(a)



(b)

Figure 9.1: Correlation of BSV:H with the BMI before (a) adjustment and after (b) adjustment.

the samples was tested using Shapiro-Wilk tests and the equality of variances was only tested for the feature samples which according to the Shapiro-Wilk test results ($p<0.05$) are drawn from normal distributions. The feature samples that showed equal variance between the OA and non-OA groups ($p<0.05$ according to Levene's tests) were selected and fed into t-tests. The feature samples that generated p-values smaller than 0.05 at the end of the t-test were selected and finally a SVM model was built using only these features:

- BCV:CORR:H_M1
- BCV:CORR:D1_M1
- BCV:CORR:V_M1
- BCV:CORR:D2_M1
- BCV:CORR:H_LF
- BCV:CORR:D1_LF
- BCV:CORR:V_LF
- BCV:CORR:D2_LF
- BCV:HOM:H_MF
- BCV:HOM:D1_MF
- BCV:HOM:V_MF
- BCV:HOM:D2_MF
- BCV:CORR:H_MF
- BCV:CORR:D1_MF
- BCV:CORR:V_MF
- BCV:CORR:D2_MF

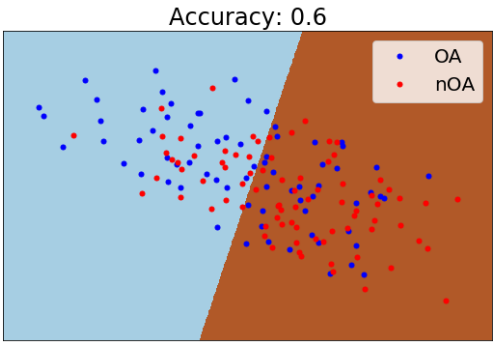
The classification score of this model can be seen in Figure 9.7. The classification score did not improve over other models, while the other best-model metrics decreased significantly as compared to the best configuration found in the original feature space with the entire feature pool. However, we observe that the features that were significantly different between the patient groups are all BCV features and 75% of them are measured at the medial compartments. Also, approximately 75% of the features listed are measured on the femur and removing these features, significantly impacts the classification scores. This was tested by building two separate classifiers that were once trained with and once

without femur features. The mean classification scores were then compared by t-tests. As a conclusion, we can only partly reject the hypothesis that we have defined in Chapter 4, namely that there is no significant difference in the texture values measured by the fractal, entropy, and Haralick features between the OA and non-OA groups on the Portugal data set. This hypothesis can be rejected only in part in case of some of the BCV features (listed above). In the case of the other 110 features, there is not enough evidence given in the data set to reject the formulated hypotheses at a 5% level of significance.

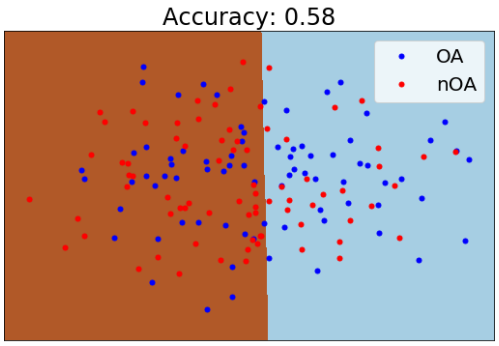
In Table 9.1, the results of the best-performing model are listed for a better overview. In conclusion of this section we note that the best configuration of a classifier is achieved when considering the entire feature pool for classification. Considering each feature group alone, the BSV-based models achieve the highest scores on the Portugal data set.

Table 9.1: OA detection maximum scores in terms of ROC-AUC obtained by all four algorithms if selecting features with *SelectFromModel*. The combined ('all') scores are ROC-AUC scores as well, but computed on models trained on the complete pool of features. The sensitivity, specificity, precision, and accuracy of the best classifiers (as seen in Figure 9.5) are also given. Portugal data set.

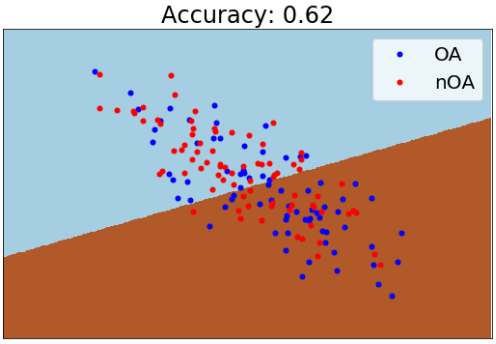
feature group	BSV	BVV	BEV	BCV	all
ROC-AUC	0.79	0.78	0.76	0.77	0.84
sensitivity	0.73	0.65	0.63	0.63	0.73
specificity	0.77	0.83	0.81	0.83	0.85
precision	0.76	0.80	0.76	0.78	0.83
accuracy	0.75	0.74	0.72	0.73	0.79



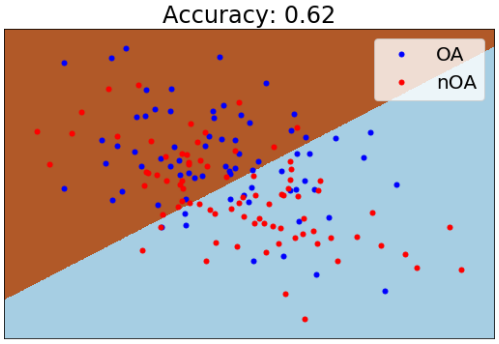
(a) BSV



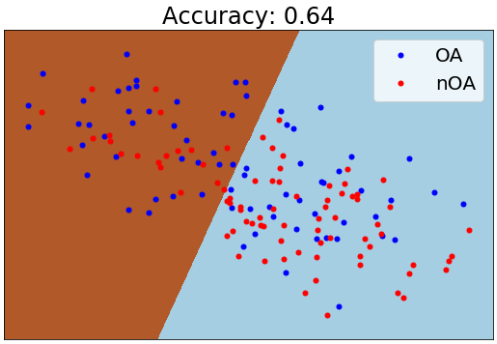
(b) BVV



(c) BEV

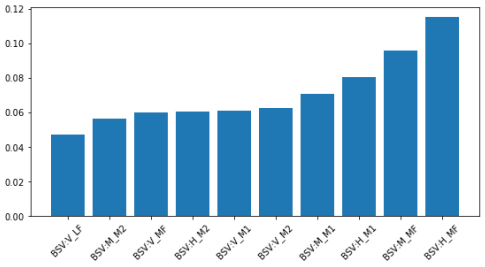


(d) BCV

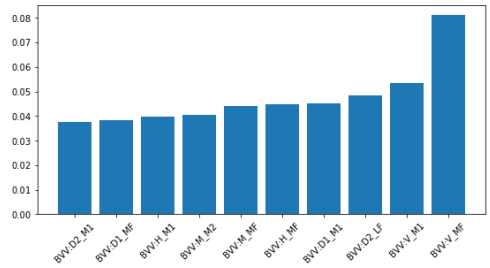


(e) all features

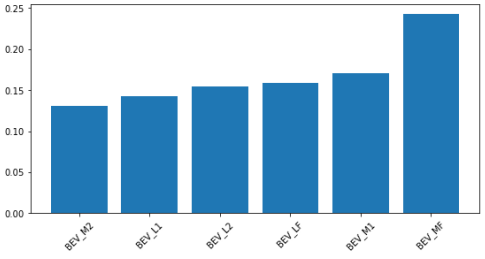
Figure 9.2: K-means classification result. 2D projection of the 126-dimensional hyperspace ((a)-(d)) for each feature group alone and (e) for all features combined using PCA. Portugal data set.



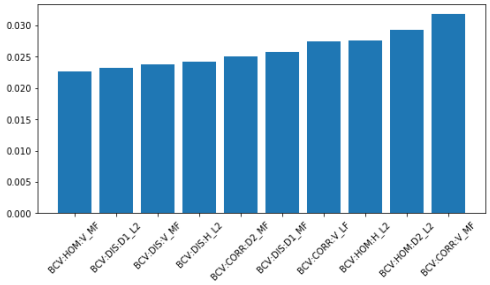
(a) BSV



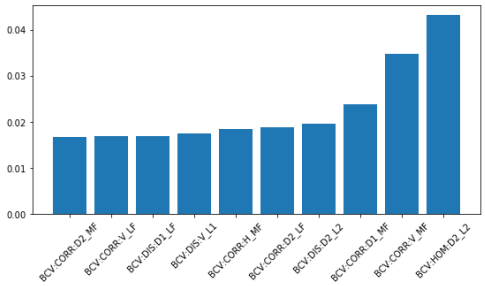
(b) BVV



(c) BEV

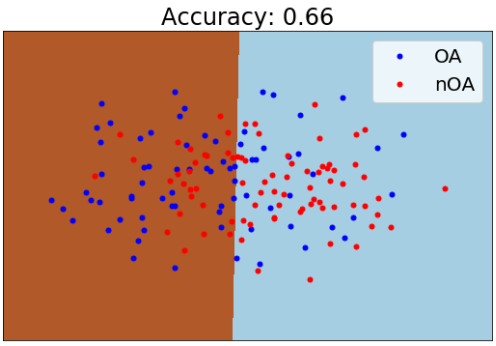


(d) BCV

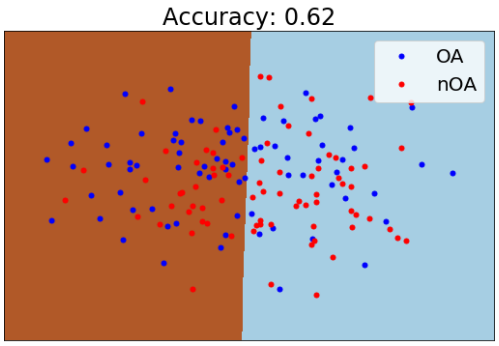


(e) all features

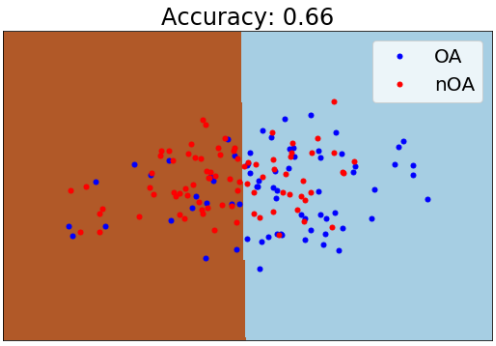
Figure 9.3: Top ten features sorted by ascending importance ((a)-(d)) for each feature group alone and (e) for all the features combined as reported by the Random Forest model. Portugal data set.



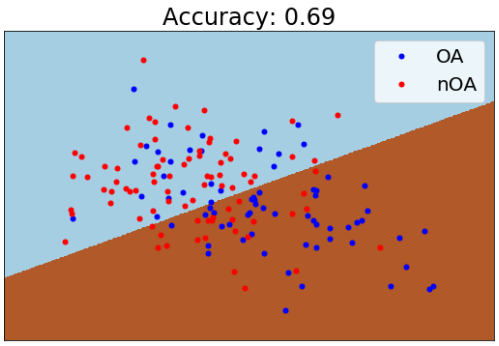
(a) BSV features



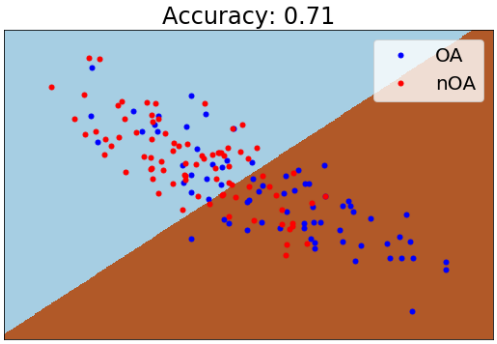
(b) BVV features



(c) BEV features

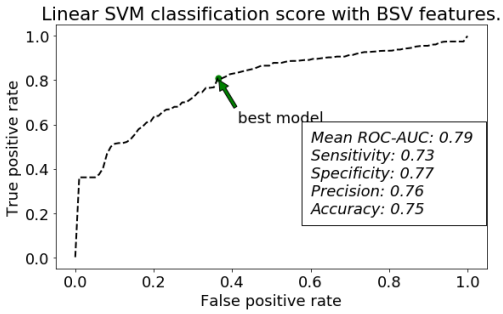


(d) BCV features

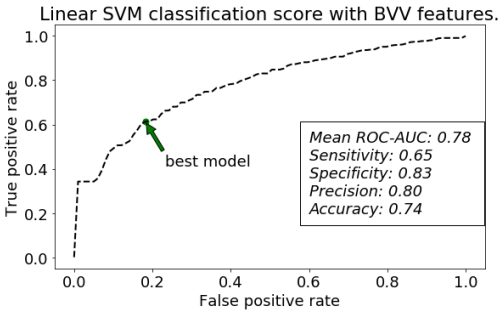


(e) all features

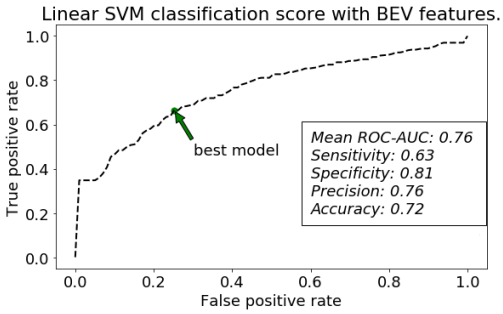
Figure 9.4: K-means classification result on the Portugal data. 2D projection of the 36-dimensional hyperspace, i.e., top-scoring features from each group only.(a)-(d) for each feature group alone and (e) for all the features combined. Portugal data set.



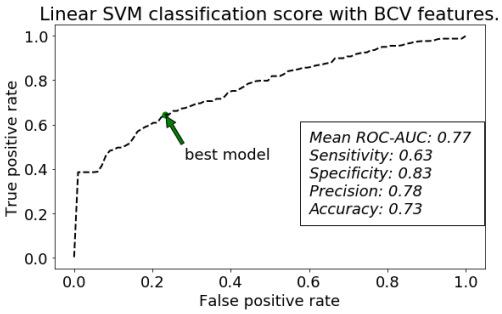
(a) BSV features



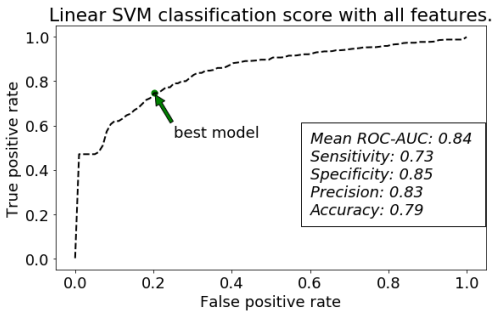
(b) BVV features



(c) BEV features

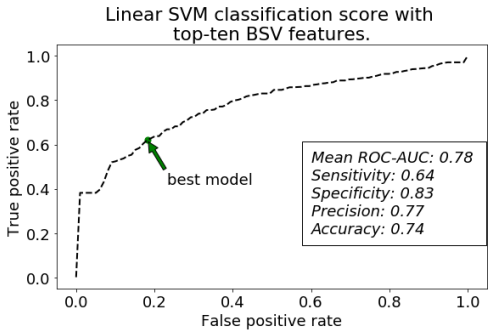


(d) BCV features

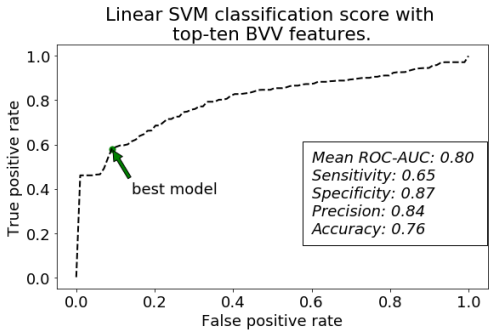


(e) all features

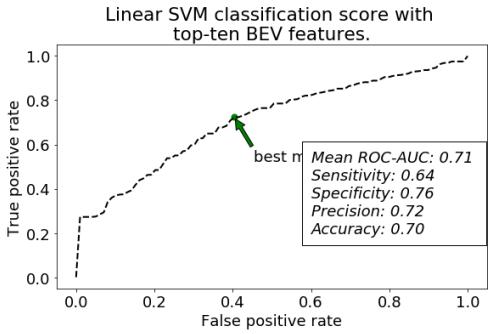
Figure 9.5: Support vector machine classification scores (a)-(d) for each feature group alone and (e) for all the features combined in terms of ROC-AUC. The green arrow points to the best model configuration that achieves the highest model metrics. These model metrics are also displayed. Portugal data set.



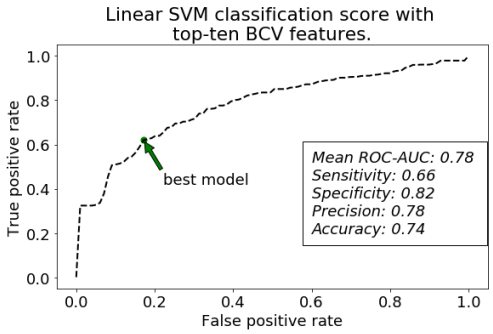
(a) BSV features



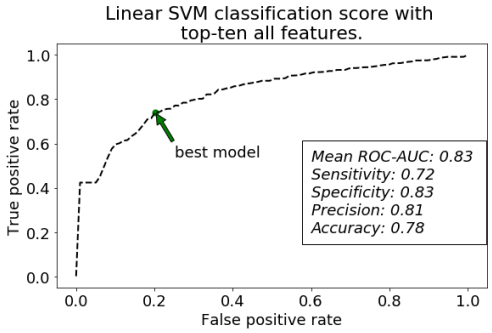
(b) BVV features



(c) BEV features



(d) BCV features



(e) all features

Figure 9.6: Linear SVM classification scores (a)-(d) for each feature group alone and (e) for all the features combined in terms of ROC-AUC. The feature pool was reduced by SelectFromModel. The green arrow points to the best model configuration that achieves the highest model metrics. These model metrics are also displayed. Portugal data set.

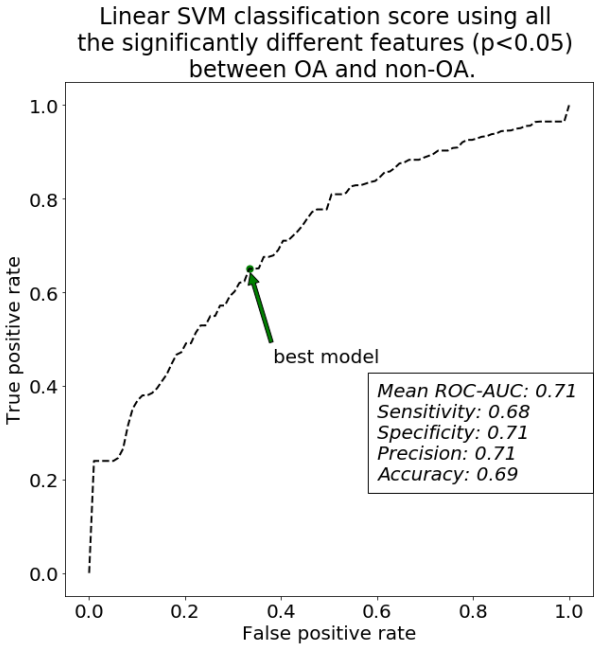


Figure 9.7: Linear SVM classification scores in terms of ROC-AUC if using the statistically different features between OA and non-OA for training. The green arrow points to the best model configuration that achieves the highest model metrics. These model metrics are also displayed. Portugal data set.

9.2 MOST Data

The features obtained from the MOST images are first adjusted with respect to the confounding variables in the same manner the Portugal data features were adjusted in Section 9.1. However, when investigating the relationships of, for example, BSV to BMI for all three visits at the same time, one can observe a strange displacement of the distribution of the 84m measures (as shown in Figure 9.8). The reason for this, we assume, is an update to the digitization software on the MOST X-Ray machines somewhere between visit 30m and 84m. To control for this behavior, we calibrate/adapt the features measured at 84m to the ones measured at BL prior to the adjustment step. The transformation that we use is based on the straightforward *z-transform*:

$$feat_{new\_84m} = \frac{(feat_{84m} - mean_{84m})}{std_{84m}} std_{BL} + mean_{BL} \tag{9.3}$$

The way the results of the classification models are reported, depend on the approached task. In the following subsections we show our results obtained on the MOST image data set for the diagnosis, early prediction, and progression tasks separately.

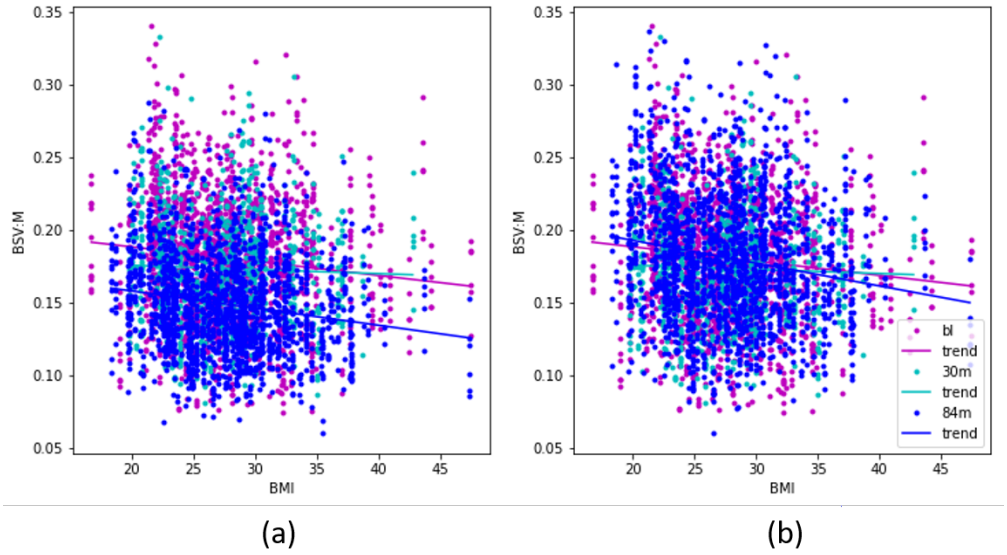


Figure 9.8: BSV vs. BMI non-calibrated (a) and calibrated (b).

9.2.1 Diagnosis

Employing K-means clustering as unsupervised learning, the classifications scores do not differ significantly from random (55% in terms of ROC-AUC at max) for BL, 30m and 84m, males and females, with or without feature selection through random forests. All the features fed to the K-means classifier were found to be homoscedastic by employing a Levene's test (p-value < 0.05).

In the case of supervised learning with a linear SVM and 100-fold cross-validation, without any feature selection we obtain at BL a maximum classification score of 82% and 80% in terms of ROC-AUC for the males and females using all features combined. At 30m we obtain for the males a maximum classification score of 75% and for the females 85% using all the features combined. At 84m the maximum scores are 82% and 75% for males and females using all features combined. Combining the linear SVM with a preceding variable selection by means of Random Forests does not improve the scores. We observe, that when taking each algorithm independently, the best classification metrics are generally obtained by the BCV features both for males and females.

Another approach that we made was to select the features according to the coefficients learned by the SVM model using the SelectFromModel class. These coefficients were treated as feature importances and the combination of features that achieves the largest scores is selected. With this model, at BL, the males obtain generally better scores, while the maxima for both males and females is obtained using the BCV values: 78% and 75% respectively. However, if we combine all the features and then select the best, we obtain a classification score of 82% and 80% respectively. At 30m, the best scores

are also obtained by the BCV features: 80% and 77% ROC-AUC for males and females respectively. A combination of features yields, however, 85% and 75% for males and females. At 84m the BCV features are once again the best performers, obtaining scores of 78% and 74% for males and females. Combining BSV, BEV, BVV and BCV features yields scores of 82% and 75% for males and females respectively.

The best scores that we have obtained can be seen in Table 9.2 together with the model metrics of the best configuration found from the ROC. When applying t-tests we find that there is no significant difference between the scores obtained for men and women at BL, 30m, or 84m.

Last but not least, we applied t-tests in the same fashion as with the Portugal data set to find the features that indicate significantly different measures between OA and non-OA. After normality and equality of variance tests, the t-tests returned the significantly different features. At BL, 60% of the returned features for the males are measured at the femur ROIs. For the females, 83% of the most different features are BCV features. At 30m, 76% and 50% classification scores in terms of ROC-AUC are obtained for males and females with the features selected being mixed. Building a SVM model using only these features, a score of 75% and 67% in terms of ROC-AUC is obtained with 100-fold cross-validation for males and females. At 84m, 75% and 68% classification scores were obtained for males and females. 76% of the features that were significantly different for the males were BCV features with mixed positioning, while for females the features were mixed (both in terms of positioning and feature groups).

Table 9.2: OA diagnosis maximum scores obtained by all four algorithms for males and females separately by selecting features with *SelectFromModel*. (a) shows baseline model metrics, (b) shows metrics obtained from the second visit, and (c) contains metrics obtained from the third visit. The combined scores are ROC-AUC scores as well, but computed on models trained on the complete pool of features. MOST data set.

	BL									
gender	male					female				
feature group	BSV	BVV	BEV	BCV	all	BSV	BVV	BEV	BCV	all
ROC-AUC	0.61	0.77	0.53	0.78	0.82	0.64	0.67	0.57	0.75	0.80
sensitivity	0.57	0.71	0.39	0.71	0.72	0.64	0.67	0.62	0.72	0.73
specificity	0.67	0.72	0.78	0.78	0.82	0.63	0.65	0.57	0.71	0.74
precision	0.63	0.71	0.65	0.77	0.79	0.64	0.65	0.59	0.70	0.74
accuracy	0.62	0.69	0.58	0.75	0.77	0.65	0.66	0.59	0.70	0.73

(a) BL

	30m									
gender	male					female				
feature group	BSV	BVV	BEV	BCV	all	BSV	BVV	BEV	BCV	all
ROC-AUC	0.63	0.80	0.63	0.80	0.85	0.69	0.76	0.63	0.77	0.75
sensitivity	0.55	0.71	0.64	0.62	0.75	0.68	0.66	0.55	0.58	0.60
specificity	0.72	0.80	0.63	0.84	0.84	0.69	0.80	0.54	0.82	0.83
precision	0.67	0.79	0.63	0.80	0.82	0.68	0.77	0.52	0.76	0.79
accuracy	0.63	0.75	0.63	0.73	0.80	0.68	0.73	0.53	0.70	0.71

(b) 30m

	84m									
gender	male					female				
feature group	BSV	BVV	BEV	BCV	all	BSV	BVV	BEV	BCV	all
ROC-AUC	0.67	0.76	0.61	0.78	0.82	0.63	0.71	0.62	0.74	0.75
sensitivity	0.60	0.64	0.44	0.70	0.70	0.68	0.64	0.55	0.64	0.70
specificity	0.68	0.80	0.78	0.77	0.81	0.58	0.70	0.69	0.73	0.72
precision	0.65	0.77	0.66	0.74	0.78	0.62	0.68	0.64	0.70	0.72
accuracy	0.64	0.73	0.61	0.73	0.76	0.63	0.67	0.62	0.69	0.71

(c) 84m

9.2.2 Early Prediction

Trying to predict the OA from its earliest stages using only texture features, we first employed unsupervised learning (K-means) on groups that were selected according to criteria presented in Section 4. With it we obtained a maximum ROC-AUC score of 57% and 55% for males and females, using all features combined after feature reduction

with Random Forests. The score did not improve significantly over the versions without feature selection (54% and 53% for males and females).

In the next step we trained another linear SVM and employed feature selection according to the *SelectFromModel* heuristic which was presented above. The simple SVM without or with feature selection through Random Forests did not perform very well. The maximum ROC-AUC score obtained in this fashion was 69%, using all features for training and 100-fold cross-validation.

Table 9.3: OA early prediction maximum scores obtained by all four algorithms for males and females separately when selecting features with *SelectFromModel*. The combined scores are ROC-AUC scores as well, but computed on models trained on the complete pool of features. The sensitivity, specificity, precision and accuracy of the best-performing models are also given. MOST data set.

gender	male					female				
feature group	BSV	BVV	BEV	BCV	all	BSV	BVV	BEV	BCV	all
ROC-AUC	0.63	0.69	0.65	0.75	0.76	0.68	0.63	0.62	0.70	0.76
sensitivity	0.50	0.60	0.59	0.66	0.66	0.58	0.56	0.63	0.65	0.67
specificity	0.77	0.75	0.68	0.78	0.78	0.76	0.69	0.62	0.70	0.76
precision	0.69	0.71	0.65	0.75	0.75	0.71	0.65	0.62	0.69	0.74
accuracy	0.63	0.67	0.63	0.72	0.72	0.67	0.63	0.62	0.68	0.72

As one can observe, the highest scores are obtained by BCV features (75% and 70% for males and females respectively), while a combination of all features yields higher, but again similar results for both (76% and 76% respectively, as shown in Figure 9.9). We furthermore computed the two-sample t-statistic of the scores obtained by both classifiers (males and females) to test the null hypothesis that the distribution of the scores obtained are equal, which yielded $p > 0.05$, meaning that there is not enough evidence in the data to reject the hypothesis. In other words, we assume that there is no significant statistical difference between men and women in terms of classification scores.

We also looked at the top-ten-scoring features of the classifiers. These top-ten-scoring features were identified by counting the number of times each feature was selected in the *SelectFromModel* approach. The results are plotted in Figure 9.10. We can observe that 8 out of 10 features for the males and 6 out of ten for the females from the top ten are BCV features. This result matches the results from Tables 9.2 and 9.3, showing that BCV produces overall the best performing set of features for OA prediction purposes as well. At the same time, 9 out of 10 features for males and 5 out of ten features for females from the top ten are features measured on the medial sector of the knee, where generally the largest loads are generally registered [101]. Moreover, 3 out of 10 features for males and 4 out of 10 features for females are measured on the femur. This observation made us interested in investigating the influence of the femural features on the classification results. Thus, we applied the same pipeline on the feature pool after filtering out the femural features (ending with "RMF" or "RLF").

The ROC-AUC of the optimal features (the ones that obtained the highest scores as seen in Figure 9.9) can be seen in Figure 9.11 for males and in Figure 9.12 for females for all feature groups separately. To compute this we first selected the features that obtained maximum scores based on the model trained with a previous application of SelectFromModel and fed them into another linear SVM for training. All the obtained results are listed in Table 9.3 for both females and males. The metrics of the best model configuration that was derived from the ROC-curve are also shown.

The influence of the femur features is shown in Figure 9.13 for both males and females. We observe that there must be a significant difference in the performance of the classifier with and without femur features as expected judging by the ten best features presented in Figure 9.10. This fact is confirmed by two-sample t-tests where we compared the means of the classifier scores obtained by the classifiers in turn: with and without femural features. The tests yielded p-values < 0.05 for both women and men. In other words, the null hypothesis that the mean of the scores of the "femural" and "non-femural" classifiers are equal can be rejected.

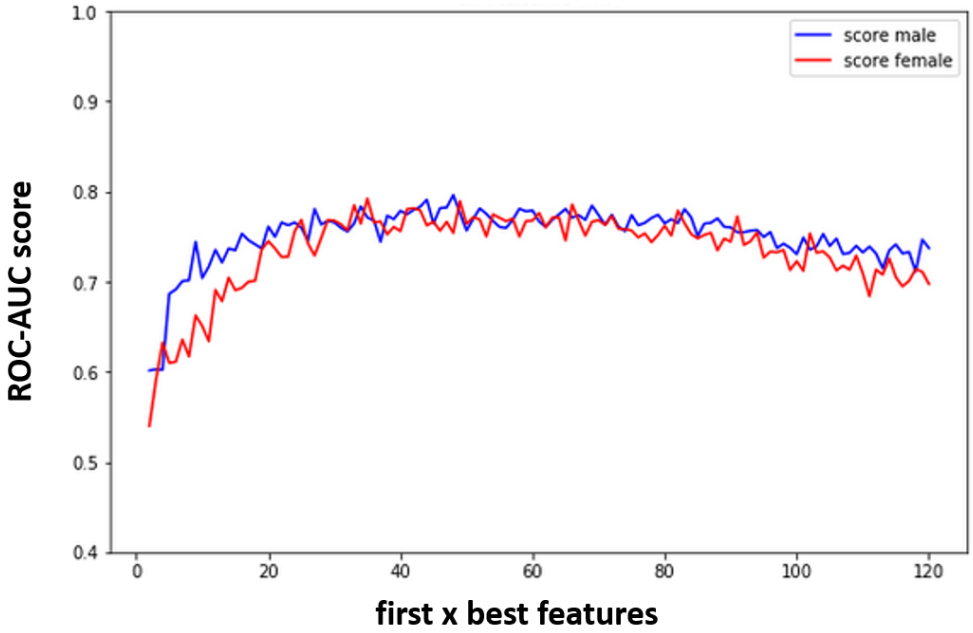
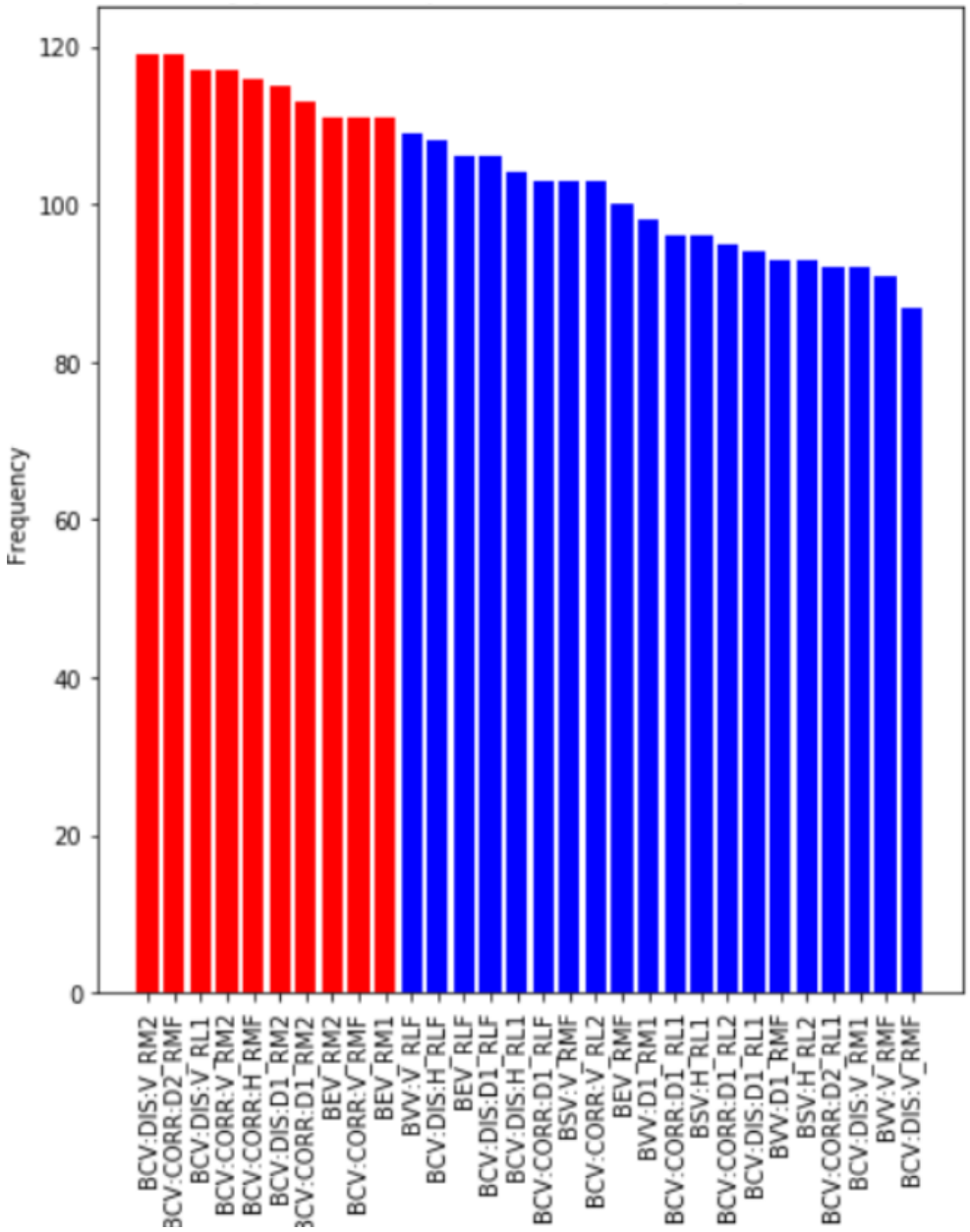
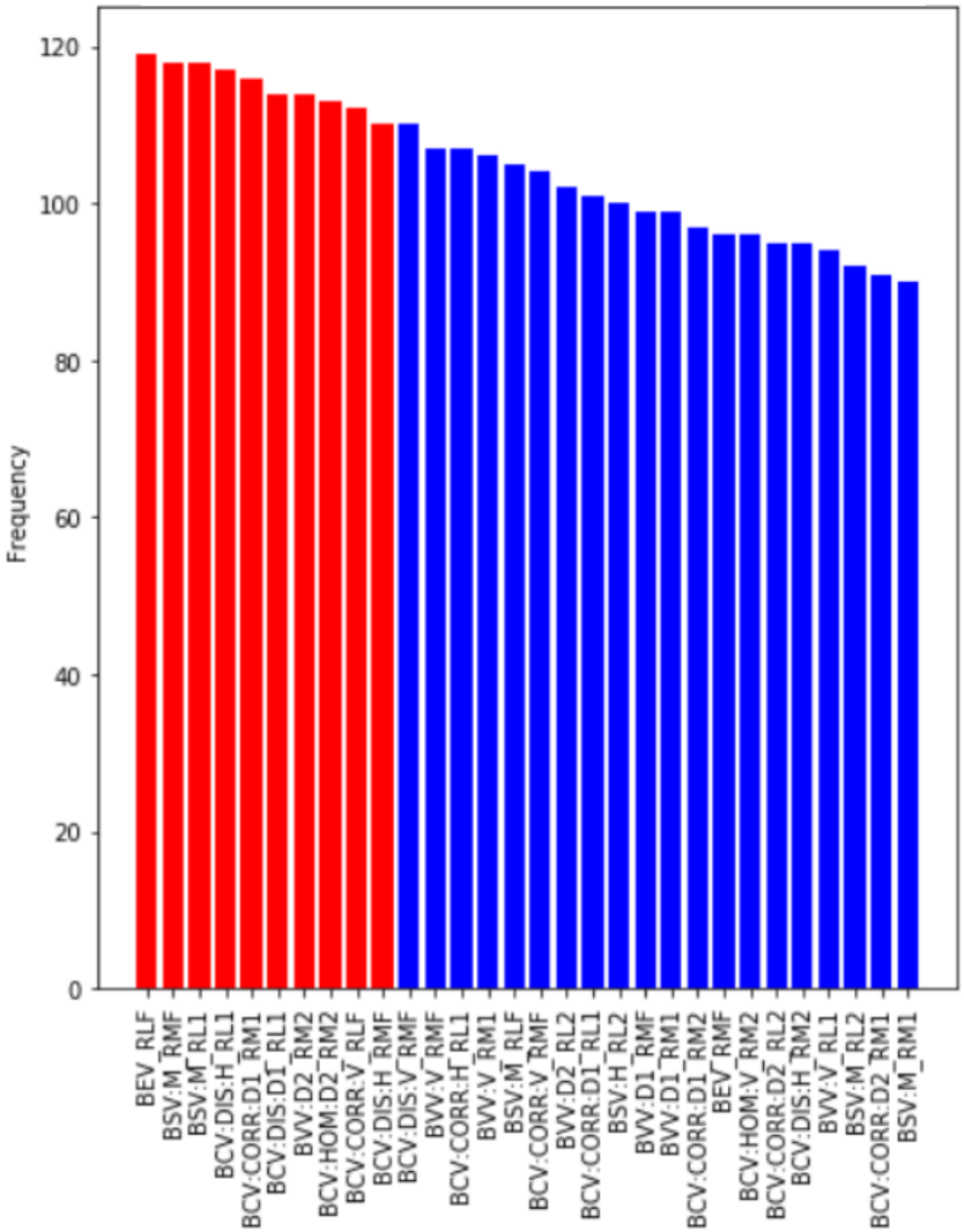


Figure 9.9: Typical output of a *SelectFromModel* applied with a SVM classifier on the entire feature pool.

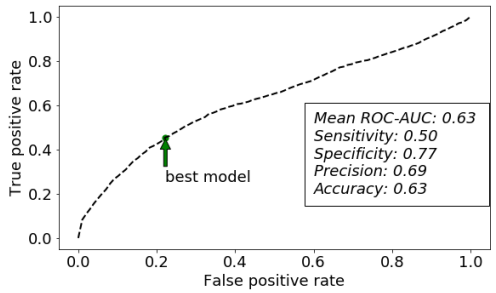

Frequency
0
20
40
60
80
100
120
BCV:DIS:V_RM2
BCV:CORR:D2_RMF
BCV:DIS:V_RL1
BCV:CORR:V_RM2
BCV:CORR:H_RMF
BCV:DIS:D1_RM2
BCV:CORR:D1_RM2
BEV_RM2
BCV:CORR:V_RMF
BEV_RM1
BVV:V_RLF
BCV:DIS:H_RLF
BEV_RLF
BCV:DIS:D1_RLF
BCV:DIS:H_RL1
BCV:CORR:D1_RLF
BSV:V_RMF
BCV:CORR:V_RL2
BEV_RMF
BVV:D1_RM1
BCV:CORR:D1_RL1
BSV:H_RL1
BCV:CORR:D1_RL2
BCV:DIS:D1_RL1
BVV:D1_RMF
BSV:H_RL2
BCV:CORR:D2_RL1
BCV:DIS:V_RM1
BVV:V_RMF
BCV:DIS:V_RMF

(a) male

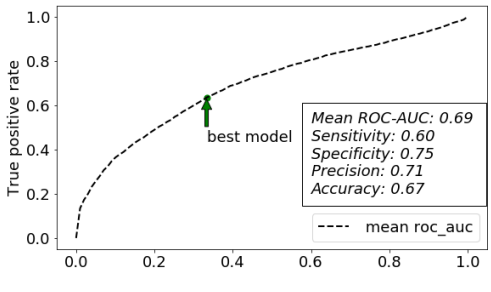


(b) female

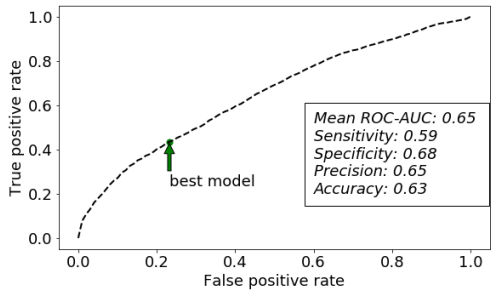
Figure 9.10: Top thirty features that were selected by the male classifier (a) and by the female classifier (b) in the early prediction task. The frequency denotes how often the feature was selected as important by the *SelectFromModel* technique. The ten best features are displayed in red.



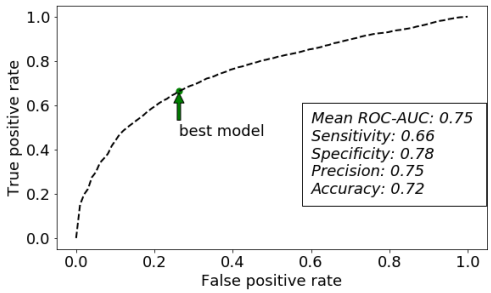
(a) BSV features



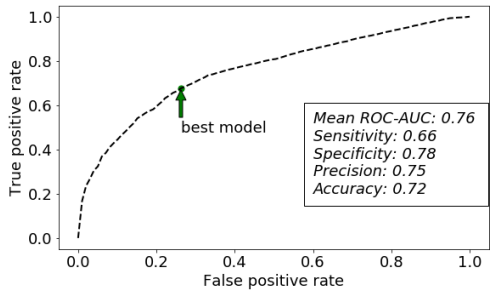
(b) BVV features



(c) BEV features

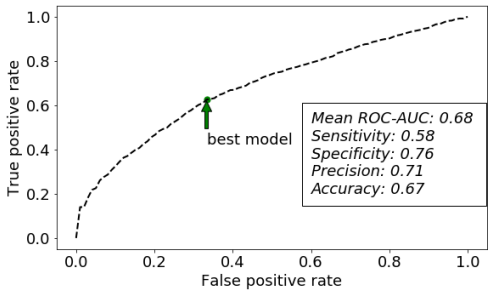


(d) BCV features

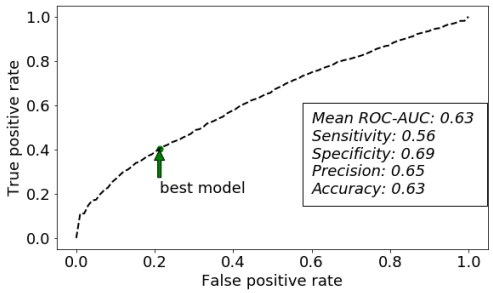


(e) all features

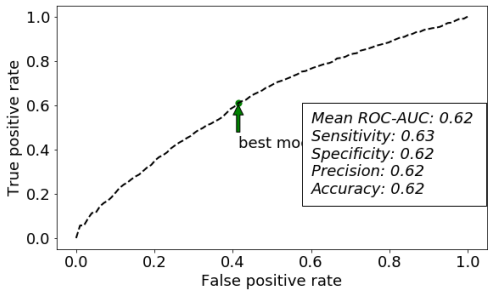
Figure 9.11: OA early prediction optimal scores obtained (a)-(d) by each algorithm independently and all algorithms combined (e) for males. The green arrow points to the best model configuration that achieves the highest model metrics. These model metrics are also displayed. MOST data set.



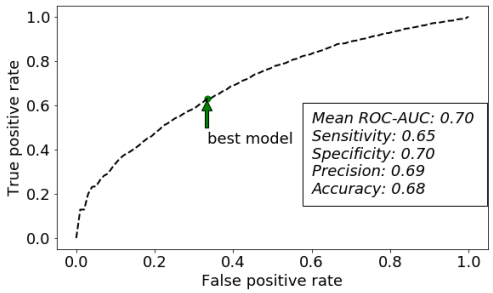
(a) BSV features



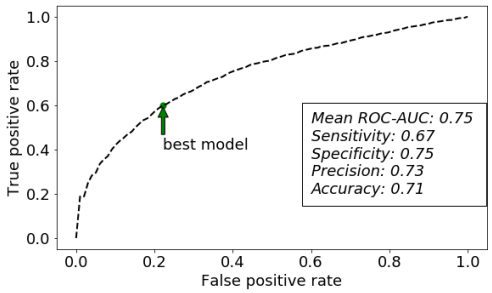
(b) BVV features



(c) BEV features

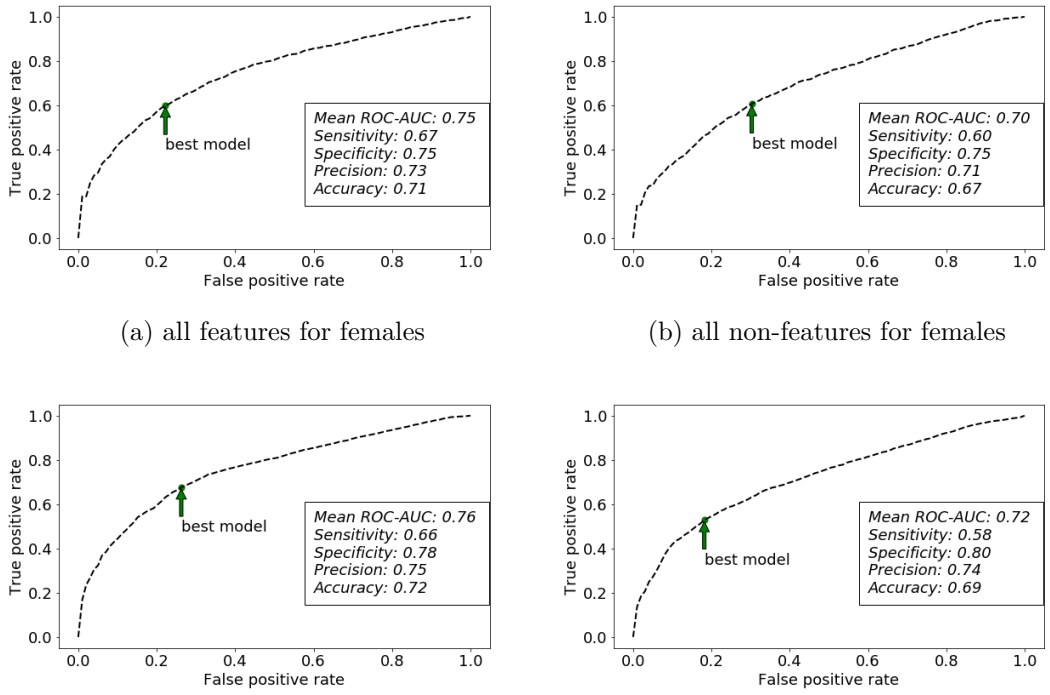


(d) BCV features



(e) all features

Figure 9.12: OA early prediction optimal scores obtained (a)-(d) by each algorithm independently and all algorithms combined (e) for females. The green arrow points to the best model configuration that achieves the highest model metrics. These model metrics are also displayed. MOST data set.



(a) all features for females

(b) all non-features for females

(c) all features for males

(d) all non-features for males

Figure 9.13: The influence of femoral features on the early prediction of OA for (a)-(b) females and (c)-(d) males. The plots on the left side show all 126 features including femoral features. The plots on the right side show the 'pruned' features, with no femural features. The green arrow points to the best model configuration that achieves the highest model metrics. These model metrics are also displayed. MOST data set.

9.2.3 Progression

To get an idea whether our algorithms could track the changes in KL scores or OA incidence, we first looked at the means of the groups we defined in Section 4.2.1 over the three visits. The results are displayed in Figures 9.14 and 9.15 for males and females respectively. We can observe that in general there is a tendency among the values to drop for the 'KL worsening' group and for the 'OA incidence' group as compared to the 'KL constant' group and 'stay healthy' group, which remain constant across the study.

To have a better confirmation of any significant differences between the visits, we employed repeated measures one-way ANOVA given the dependency of the samples across the study. The independent variable was 'time', i.e., the visit (BL, 30m and 84m) and the dependent variables are the 126 engineered features.

First, we tested the 'conditions' (visits in this case) for variance heterogeneity using O'Brien's test [102] regarding each of the 126 available features (dependent variables). The O'Brien test is based on the null hypothesis that the variances under the different conditions are equal. We thus, filtered out all the features for which we obtained a p-value > 0.05. For those features, we could not reject the null hypothesis at a significance level of 5%. In other words, we can assume that the measurements recorded at the three different visits for these filtered features come from distributions with equal variances.

Second, for the significant features with respect to the O'Brien Test, we employed repeated measures ANOVA to find possible significant differences between the means of the samples coming from different visits. This test is able to detect whether there is a significant difference somewhere among the conditions, but it is not able to identify the exact conditions that produced samples that are significantly different. Thus, this step narrowed the feature pool even further. We excluded all the features for which the ANOVA $p - value$ was bigger than 0.05. For these features we could not reject the null hypothesis that the sample means are equal. For the remaining features, we reject this assumption and consider the means different.

Third, we employed a post-hoc Tukey HSD test for multiple cross comparisons of different features under different conditions. For each comparison, all the significant features were recorded.

The above steps were applied for all the groups of patients defined: KL-constant, KL-worsening, OA-incidence, stay-healthy. We obtained the following features, per group:

1. **'KL constant':**
    a) **BL vs. 30m** : no features
    b) **30m vs. 84m** : no features
    c) **BL vs. 84m** : no features
2. **'KL worsening':**

a) **BL vs. 30m** : no features

b) **30m vs. 84m** : no features

c) **BL vs. 84m** : no features

3. **'stay healthy':**

a) **BL vs. 30m** : no features

b) **30m vs. 84m** : no features

c) **BL vs. 84m** : no features

4. **'OA incidence':**

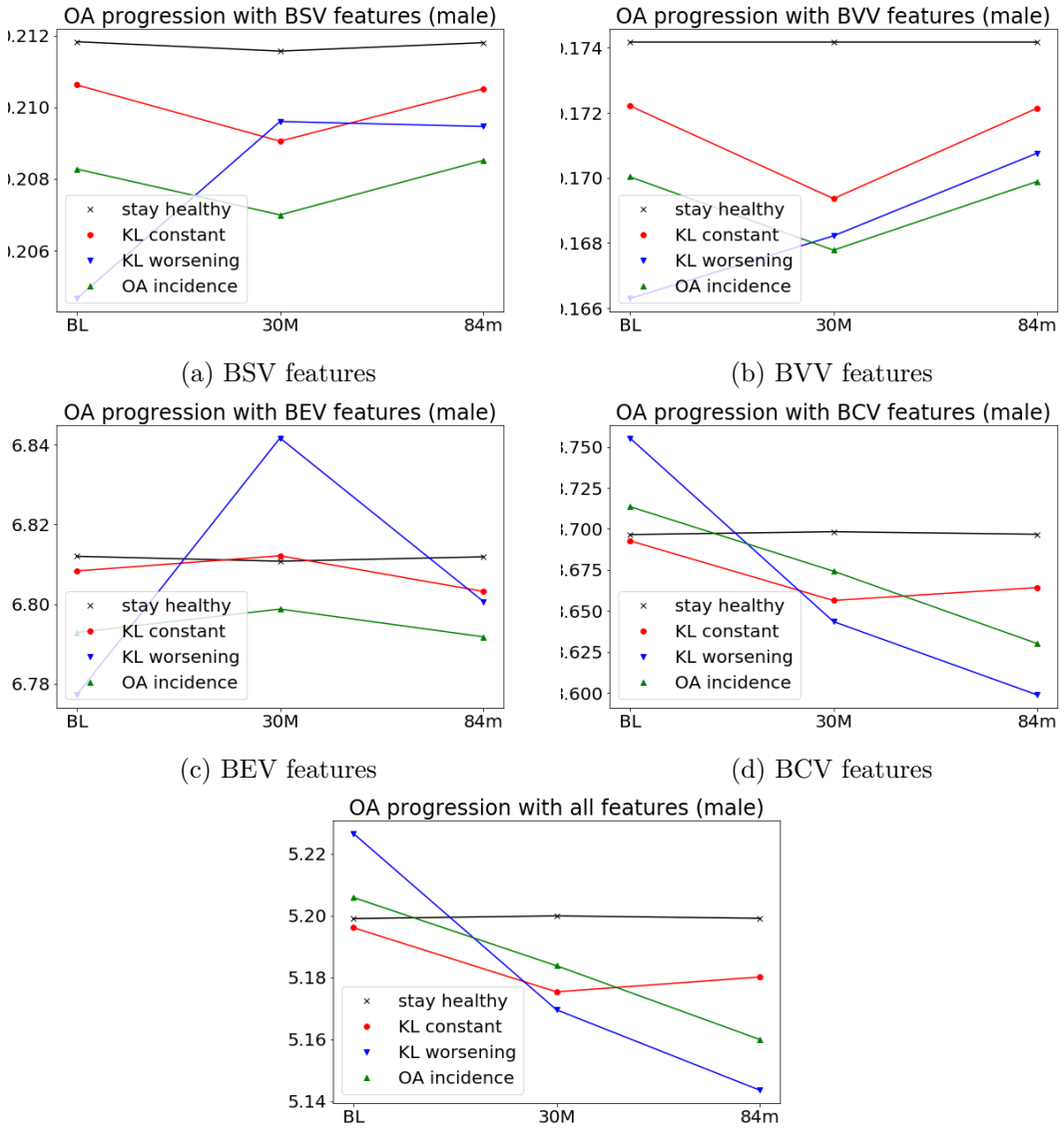
a) **BL vs. 30m** : 'BVV:V_RM1'

b) **30m vs. 84m** : no features

c) **BL vs. 84m** : 'BVV:V_RL1', 'BVV:V_RM2', 'BVV:D1_RM2', 'BVV:M_RM2', 'BVV:H_RM2', 'BVV:V_RL2', 'BVV:D1_RL2', 'BVV:M_RL2', 'BVV:H_RL2', 'BVV:V_RM1', 'BVV:D1_RM1', 'BVV:M_RM1', 'BVV:H_RM1', 'BVV:V_RMF', 'BSV:V_RM2', 'BSV:M_RM2', 'BSV:H_RL2', 'BSV:V_RL2', 'BSV:M_RL2', 'BSV:H_RM1', 'BSV:V_RM1', 'BSV:M_RM1', 'BSV:V_RMF', 'BEV_RL2', 'BEV_RM1'

We observe that the 'KL constant' and 'stay healthy' groups did not produce any features that are significantly different between the visits taken pair-wise. The 'KL worsening' group only produced a single feature that was significantly different between BL and 84m, i.e., across a larger period of time. This feature is fractal-related and was measured in the medial compartment of the TB. However, the 'OA incidence' group produced more features for the BL vs. 84m cross-comparison that are listed above. 64% of the features produced were measured at the medial compartment where the highest loads are expected [103]. For the smallest time interval only one feature with significant differences was detected with the exception of a single one, i.e., the same feature (BVV:V_RM1) that was detected in the 'KL worsening' group. Here, the difference was recorded between the first two visits, whereas for the KL-worsening group it was recorded between the last two visits. We assume that no significant differences were recorded between the visits of the 'KL worsening' group, since this group per definition also contains patients for which the KL score changed only by a single unit. This leads us to the conclusion that the algorithms are not sensitive to slight changes in the progression of the disease, but rather detect the early onset and changes over longer periods of time.

We also notice that the significant feature pool of the 'OA incidence' group between BL and 84m is dominated by fractal features. 92% of the detected features are fractal-based and only 8% are entropy-based. Out of these, 40% are features measured in the vertical direction (i.e., theoretically along the trabeculae), while the rest of 60% is equally divided among mean, horizontal, and diagonal features. Unexpectedly, no BCV features were detected as significant. Only 8% (2 features) are femur features.



(a) BSV features

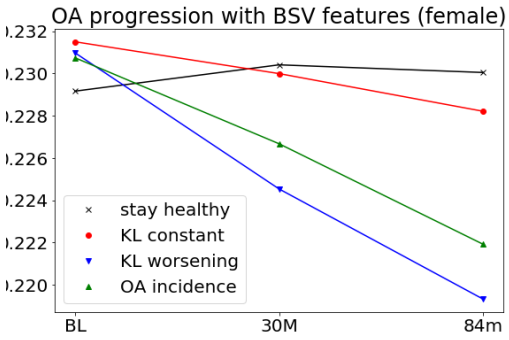
(b) BVV features

(c) BEV features

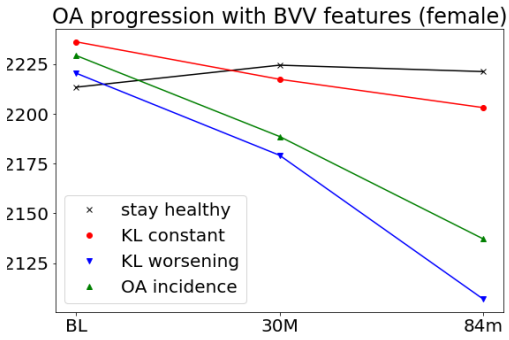
(d) BCV features

(e) all features

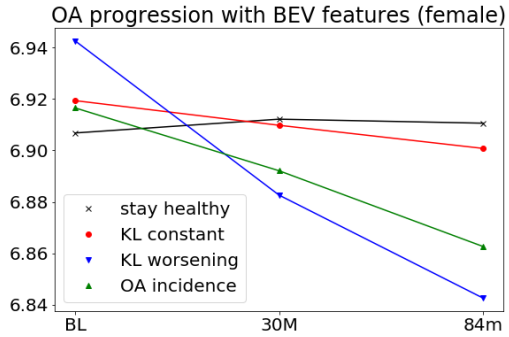
Figure 9.14: OA progression visualization of sample means. (a)-(d) each feature group independently and (e) all algorithms combined for males.



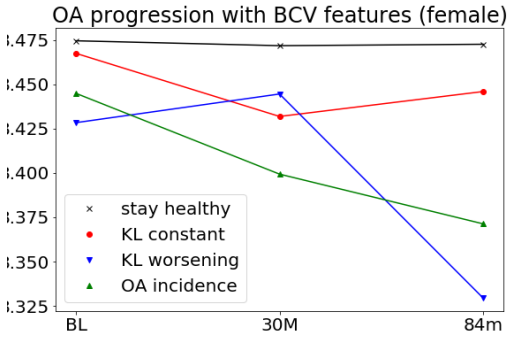
(a) BSV features



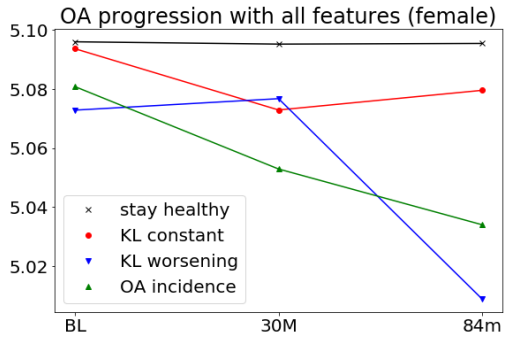
(b) BVV features



(c) BEV features



(d) BCV features



(e) all features

Figure 9.15: OA progression visualization of sample means. (a)-(d) each feature group independently and (e) all algorithms combined for females.

In this chapter we have presented the outcomes of the statistical tests employed on our texture features and the classification scores obtained by the trained models. We observe that in the case of the Portugal data, the unsupervised learning approach with the basic K-means algorithm performs very well when considering only the subspace of the most significant features: 71% in terms of ROC-AUC. With supervised learning, the best configuration of a SVM model trained on the entire feature pool produces a ROC-AUC score of 84% with 73% sensitivity, 85% specificity, 83% precision and 79% accuracy. Feature selection has a negative impact on the metrics. It is worth mentioning that the highest contributors to the correct learning of the model are the BCV medial features and removing the femural features does have a significant impact on the performance of the models. In general, the most sensitive model was found to be the one built on BSV features (73%). The model trained only on BVV features was the most specific and most precise (83% and 80%). The most accurate model was trained on the whole feature pool (79%). We note that the lowest sensitivity was achieved by the BVV-, BEV- and BCV-based models, i.e., around 65%.

In the case of the MOST data set we notice that the BCV produces the best features if considering the algorithms independently, obtaining high ROC-AUC scores for the discrimination and early prediction tasks, both for men and for women. Considering all features combined, the scores improve even further. However, for disease progression, the fractal features produced from BVV and BSV appear to be of more significance. Similar to the Portugal data set, the impact of the femoral features is crucial on the classification score, as models trained on non-femoral features only produce significantly lower scores ($p < 0.05$). For the discrimination task, we notice that the model trained on the entire feature pool showed the highest ROC-AUC scores and the highest sensitivity, specificity, precision, and accuracy, i.e., higher than 70%, for the male and female participants at BL, 30m and 84m. The sensitivity of the model trained for females at 30m is an exception, reaching only 60%. The situation is similar in the case of the early prediction models. The ROC-AUC scores of the best configurations of the models trained on all features exceed 75% ROC-AUC, while the other model metrics lie around 75% with the exception of the sensitivity which is slightly below 70%.

CHAPTER 10

Conclusion

In this work we have conducted a clinical experiment to investigate the possibility of detecting early signs of OA in the human knee TB based on engineered textural features from simple 2D knee radiographs. We first gave a motivation for this area of research based on the current worldwide reports and statistics on the negative impact on the general economy and the quality of life of individuals caused by OA. The early detection of OA might render some of earliest bone-damaging processes reversible provided that the treatment is applied in time. Also, by using conventional X-rays and automating the detection, the diagnosis costs (in time and money) can be significantly reduced.

Second, we shortly presented novel approaches from the literature that attempt at solving the early-OA problem in different ways. From the use of texture features to the use of MRI images for lesion detection and blood serum tests, the need for early detection tests have become clear lately and an increasing number of publications attempt to find a cheap and reliable way that could predict OA.

Third, we provided a biological background of OA by presenting the general knee anatomy and the microscopical architecture of the TB. We also described some of the earliest known processes to date that are known to weaken the bone and prepare it for the onset of OA. The earliest changes due to OA in the TB of the knee, that are undetectable by conventional methods, are analyzed as well. These changes happen generally a long time before the affection becomes symptomatic and the patient becomes aware of it.

Fourth, we described the image data sets that we use in our experiment in detail: the Portugal data set and the MOST data set. At the same time, based on the available data we defined our investigation tasks that would be possible for each data set, i.e., patient discrimination attempt using textural features (for both Portugal and MOST data sets), early prediction of the disease and disease progression tracking (for the MOST data set, since it is a longitudinal OA study).

Fifth, we discussed four algorithms that we use for feature engineering in detail. We have thus seen four different approaches at describing the structure complexity in an ROI: two fractal-based techniques, one entropy-based technique and the last technique based on Haralick features derived from the GLCM.

Sixth, we created artificial surfaces for each algorithm to serve as a validation basis. The artificially generated surfaces were developed with fixed, known, and theoretical parameters (fractal dimension, entropy, homogeneity, correlation, etc.) and the algorithms that we presented were employed on these surfaces to test how well they approximate the theoretical parameters through the applied heuristics.

Seventh, we introduced and explained the statistical methods and models that we have used to interpret the engineered features. We discussed different statistical tests used to search for statistically significant differences among samples of features coming from ill and healthy patients. We also presented simple neural networks (SVM) that are trained to learn the differences between these patient groups and predict the diagnosis of a new patient. All discussed models that were used in this work were introduced and mathematically presented with a view of understanding their suitability for characterizing the present data and providing answers to our research questions.

Last but not least, we presented and interpreted the outputs of the used statistical methods. For the Portugal data set we have obtained the highest classification score of 84% in terms of ROC-AUC using all 126 features combined in a linear SVM model without feature reduction. The best model produced a sensitivity of 73%, a specificity of 85%, a precision of 83% and an accuracy of 79% Most (75%) of the top features were BCV features and most (75%) were features measured at the medial compartments of the femur. We have investigated the influence of the femur regions with two separate classifiers. The scores obtained with the femur features were significantly higher than the scores obtained by the classifier without the femur features considered. If considered alone, the BCV features perform the best on the Portugal data set. In the case of the MOST study we found that for diagnosis tasks, the BCV features obtain the highest scores both for females and males at all visits (between 70% and 80% in terms of ROC-AUC) if considered alone, similar to the Portugal data set. The maxima, however, are reached when considering all the features combined (between 75% and 85%). In general, over 75% of the features that were found to be significantly different by statistical tests between ill and healthy participants were BCV features both for males and females.

In the case of the prediction task, the situation is similar. Alone, the BCV features obtain the highest classification scores (75% and 70% for males and females). The combination of features improves this even further (76% and 76%), which outperforms the current state of the art, as we manage to predict the KL score by only using textural features. There was again no significant difference according to the t-statistic between the distribution of scores obtained for men and women. 80% of the ten most significant features for males were BCV features. The influence of the features measured in the femur area have proven to be crucial. Removing them yields significantly lower scores both for men and women.

In the case of the disease progression tracking, we have learned that the fractal features are more suitable as opposed to the other tasks where BCV produced the best features. 92% of the significantly different features between BL and 84m were of fractal nature. 60% of the same features are measured at the medial compartments and the femur features do not play a significant role (only 2% of the features were measured at the femur). However, significant differences could only be found between BL and 84m, but not between BL and 30m or between 30m and 84m in the 'OA incidence' group. Also, 40% of the features were measured in the vertical direction, which indicates that the fractal algorithms are sensitive to the trabeculae orientation in the image. The other groups showed no differences between any of the visits. The BEV features appear to be unsuitable for the tasks defined in this work, as they were rarely picked as important features.

CHAPTER 11

Future Work

Even though in the present work we were able to build classifiers that are capable of detecting early signs of OA and of discriminating between healthy and ill patients with high accuracy, there are still some limitations or unanswered questions. By answering this, we might bring improvements to our models and/or also to the engineering of features.

A future research idea would be to investigate whether the algorithms are influenced by the pixel spacing or by the resolution of the images. Also the influence of the machine manufacturer is not known, but we already posses strong evidence that even though the machine parameters are set to be the same, the recording of an X-Ray image is affected by other production parameters that are not accessible during post-production.

A strong limitation of this work is that we were not yet able to define certain thresholds that are generally valid. In other words, there is no boundary per feature (group) below which (or above which) one could say that all regions that produce values in those intervals come from unhealthy (or healthy) patients. Before such a threshold can be found, we assume that we must normalize the images initially to account for all the differences that come from the machine (production) parameters and other possible influential factors.

An interesting research area regarding these algorithms might be their extension to the detection of other affections (such as OP, RA, bone cysts, implant loosening) in other anatomical regions, such as the hips, hands, spine. Other than osseous tissues, we have also shortly tested the algorithms on mammographies for the detection of tumors and on microscopic images of red blood cells for the automatic detection of their agglomeration. The results encourage further research in the areas, but more image data is needed to safely apply statistical methods and to build classification models.

Throughout this study we have noticed that the BEV is rarely chosen as a significant feature. An idea would be to actually test the importance of this feature group in the same manner that we tested the influence of the femur regions. Does the inclusion of

the BEV feature change the prediction scores of our models? Provided that the results are not significantly improved by the consideration of the BEV features, we propose an improvement to the algorithm. Instead of computing a single mean measure per chosen region, the algorithm could be adapted to calculate an entire feature image by subdividing the original ROI into subregions. The same idea could also be applied to the other procedures as well to investigate whether the scores can be improved.

The features measured in the lateral compartments were as well rarely selected as important to the model. We could investigate whether we could entirely remove these features from consideration and focus only on the medial compartments. It is rarely the case that the bone is so damaged that the lateral compartments are equally or more affected by bone deformation as the medial ones.

In this work we assumed the ideal case that the features we measure must be linearly separable and thus we employed a SVM with a linear kernel to build the models. However, the separation boundary could be in fact much more complex than a simple hyperplane, given the high number of biological and risk factors that directly or indirectly influence the bone structure.

Last bun not least, the applicability of the algorithms on other modalities, such as CT or MRI data, must also be investigated. This is strongly related to the idea of a possible extension of the algorithms to work in 3D spaces as well. Also, in this work we restricted the experiments to more or less homogeneous data, i.e., the models were built on image data coming from the same study. It would be interesting to investigate whether the combination of data sets would improve or negatively impact the scores.

List of Figures

List of Tables

Glossary

bone condyle a rounded protuberance at the end of some bones, forming an articulation with another bone.. 25

continuous-time a continuous-time stochastic process is one in which the random variable takes a contiguous set of values. 43

Gaussian process the Gaussian process is a stochastic process in which every random variable is normally distributed. In addition, a

nite set of those variables is multivariately normally distributed.. 43

Golden Section Search The GSS is an iterative technique which allows one to

ne a minimum or maximum for a unimodal function by repeatedly narrowing the intervals in which the respective extremum is known to lie. At each step, the intervals are narrowed by the golden ratio.. 45

homogeneous function a function is homogeneous of order n if it satisfies $f(tx) = t^n f(x)$. In other words if the argument of the function is multiplied by a factor, then the result will come out multiplied by the n^{th} power of that specific factor.. 43

Hurst exponent statistically, the Hurst coefficient is a measure of long-range dependency of time series. This means that the Hurst exponent is a global property if a signal. Given the fact that fractal signals are self-similar, 'the local properties are reflected in the global ones' [70, p. 1]. 35, 36, 39, 40, 42, 43

long-range dependency also known as long-range memory, long memory or long-range persitency — a term that arises when studying the decay of statistical dependency (autocorrelation) of two or more measurments with increasing time between the measurments. 113

osteophyte a rounded protuberance at the end of some bones, forming an articulation with another bone.. 2

stationary process a stationary process is a stochastic process whose combined probability distribution does not change when shifted in time. Consequently, the mean and variance are also stationary.. 43

Acronyms

AI Artificial Intelligence. 4, 9, 25

ANOVA analysis of variance. xiv, 65–69, 109

BCV Bone Coocurrence Value. ix, xi, xiii, 4, 26, 33, 47, 50, 56, 57, 78, 80, 89–91, 93, 94, 97–100, 102, 108

BEV Bone Entropy Value. ix, xi, xiv, 4, 26, 33, 45, 49, 57, 58, 78, 89, 93, 94, 98–100, 103, 108

BMI body mass index. 3, 4, 19, 27, 29, 77, 78, 88, 109

BSV Bone Score Value. ix, xi, xiii, 4, 26, 33, 42, 43, 51–55, 78, 80, 88, 89, 93, 94, 98–100, 108, 109

BVV Bone Variance Value. ix, xi, xiii, 4, 26, 33, 39, 40, 51–55, 78, 89, 93, 94, 98–100, 108

CM Co-occurrence Matrix. 47–50

DXA dual-energy X-ray absorptiometry. 27, 28

fBf fractal Brownian function. 42

fBm fractal Brownian motion. 42–44

FD fractal dimension. 28, 29, 34, 35, 42, 51, 52, 55, 57

fGn fractal Gaussian noise. 43, 44

GLCM Gray-Level Co-occurrence Matrix. 47, 57, 102, 108

GLRLM grey level run length matrix. 8

H Hurst coefficient. 35, 39, 40, 44, 45, 51–55, 108

IB Lab Image Biopsy Lab G.m.b.H.. 4, 25

IID independent and identically distributed. 73

JSN Joint Space Narrowing. 4, 23, 24, 107

KL Kellgren-Lawrence grade. 4, 7, 23, 25, 28, 29

LBP low back pain. 26

MLR multi-linear regression. 77, 78

MOAKS MRI OA knee score. 7

MOST Multicenter Osteoarthritis Study. vii, xiii, 28, 29, 31, 88–90, 100, 102, 111

MRI magnetic resonance imaging. xiii, 7, 28, 101

NCP non-collagenous proteins. 12, 16, 17

OA osteoarthritis. xi, xiii, xiv, 1–5, 7–10, 12, 14, 16–30, 51, 59, 63, 71–74, 77, 79, 80, 86, 87, 89–91, 93–95, 98, 99, 101, 105, 107, 109–111

OARSI Osteoarthritis Research Society International. 2, 23

OP osteoporosis. 27, 105

PCA Principal Component Analysis. xiv, 75, 81, 109

PDF probability density function. 43, 44

QP Quadratic Programming. 72

RA rheumathoid arthritis. 8, 26, 105

RMANOVA repeated-measures analysis of variance. 68

RMD Rheumatic and musculoskeletal disease. 26, 27

ROC-AUC Area Under the Receiver Operating Characteristic Curve. x, xi, 4, 79, 84–91, 100, 109, 111

ROI region of interest. 5, 8, 25, 27, 29, 32, 41, 43, 77, 78, 90, 102, 106–108

SPR Portuguese Society of Rheumatology. 26

SVM Support Vector Machine. xiv, 5, 72, 73, 79, 85, 87–91, 100, 102, 106, 109

TB trabecular bone. xi, 2, 3, 11, 12, 22, 33, 35, 37, 39, 40, 43, 51, 59, 101, 107

VOT Variance Orientation Transform. 40, 41, 108

Bibliography

[1] C. R. Chu, A. A. Williams, C. H. Coyle, and M. E. Bowers, "Early diagnosis to enable early treatment of pre-osteoarthritis," *Arthritis Research & Therapy*, vol. 14, no. 3, p. 212, 2012.

[2] F. Bronner, M. Farach-Carson, and H. I. Roach, *Bone and Developement*, 2010.

[3] C. Helmick, "The Burden of Musculoskeletal Diseases in the United States," 2017. [Online]. Available: www.boneandjointburden.org

[4] M. Favero, R. Ramonda, M. B. Goldring, S. R. Goldring, and L. Punzi, "Early knee osteoarthritis," in *RMD Open*, vol. 1, 2015.

[5] G. Lester, J. McGowan, and J. Panagis, "Handout on Health: Osteoarthritis," 2016. [Online]. Available: www.niams.nih.gov

[6] J. Dequeker and F. P. Luyten, "The history of osteoarthritis-osteoarthrosis," *Annals of the Rheumatic Diseases*, vol. 67, no. 1, pp. 5–10, 2008. [Online]. Available: http://ard.bmj.com/cgi/doi/10.1136/ard.2007.079764

[7] R. S. Karsh and J. D. McCarthy, "Archeology and Arthritis," *A.M.A Archives of Internal Medicine*, vol. 105, no. 4, pp. 640–644, 1960.

[8] R. F. Loeser, S. R. Goldring, C. R. Scanzello, and M. B. Goldring, "Osteoarthritis: A disease of the joint as an organ," *Arthritis and Rheumatism*, vol. 64, no. 6, pp. 1697–1707, 2012.

[9] "Medical Expenditures Panel Survey (MEPS)." Agency for Healthcare Research and Quality. U.S. Department of Health and Human Services, Tech. Rep. [Online]. Available: http://meps.ahrq.gov/mepsweb

[10] K. P. H. Pritzker, S. Gay, S. A. Jimenez, K. Ostergaard, J. P. Pelletier, K. Revell, D. Salter, and W. B. van den Berg, "Osteoarthritis cartilage histopathology: Grading and staging," *Osteoarthritis and Cartilage*, vol. 14, no. 1, pp. 13–29, 2006.

[11] C. Buckland-Wright, "Subchondral bone changes in hand and knee osteoarthritis detected by radiography," *Osteoarthritis and Cartilage*, vol. 12, pp. 10–19, 2004.

[12] F. Bronner and M. C. Farach-Carson, *Bone and Development*. Sptinger, 2010.

[13] D. B. Burr and M. a. Gallant, "Bone remodelling in osteoarthritis," *Nature Reviews Rheumatology*, vol. 8, no. 11, pp. 665–673, 2012. [Online]. Available: http://dx.doi.org/10.1038/nrrheum.2012.130

[14] T. Neogi, "Clinical Significance of Bone Changes in Osteoarthritis," in *Osteorheumatology 2011 : International Congress on Bone Involvement in Arthritis*, vol. 14, no. S2, 2012, p. 2.

[15] A. Chang-Miller, *Osteoarthritis*, 2016. [Online]. Available: http://www.mayoclinic.org/diseases-conditions/osteoarthritis/diagnosis-treatment/diagnosis/dxc-20198270

[16] N. L. Fazzalari and I. H. Parkinson, "Fractal Dimension and Architecture of Trabecular Bone," *Journal of Pathology*, vol. 178, no. 1, pp. 100–105, 1996.

[17] 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." *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine*, vol. 223, no. 2, pp. 211–236, 2009.

[18] T. Lundahl, W. J. Ohley, S. M. Kay, and R. Siffert, "Fractional brownian motion: a maximum likelihood estimator and its application to image texture." *IEEE Transactions on Medical Imaging*, vol. 5, no. 3, pp. 152–61, 1986. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18244001

[19] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," pp. 610–621, 1973. [Online]. Available: http://ieeexplore.ieee.org/document/4309314/

[20] T. Janvier, R. Jennane, H. Toumi, and E. Lespessailles, "Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the Osteoarthritis Initiative," *Osteoarthritis and Cartilage*, vol. 25, no. 12, pp. 2047–2054, 2017.

[21] V. B. Kraus, S. Feng, S. C. Wang, S. White, M. Ainslie, A. Brett, A. Holmes, and H. C. Charles, "Trabecular morphometry by fractal signature analysis is a novel marker of osteoarthritis progression," *Arthritis and Rheumatism*, vol. 60, no. 12, pp. 3711–3722, 2009.

[22] T. Woloszynski, P. Podsiadlo, G. W. Stachowiak, M. Kurzynski, L. S. Lohmander, and M. Englund, "Prediction of progression of radiographic knee osteoarthritis using tibial trabecular bone texture," *Arthritis and Rheumatism*, vol. 64, no. 3, pp. 688–695, 2012.

[23] S. Oancea, "Variance Orientation Transform," Bachelor's Thesis, TU Wien, 2016.

[24] L. Sharma, J. S. Chmiel, O. Almagor, D. Dunlop, A. Guermazi, J. M. Bathon, C. B. Eaton, M. C. Hochberg, R. D. Jackson, C. K. Kwoh, W. J. Mysiw, M. D. Crema, F. W. Roemer, and M. C. Nevitt, "Significance of preradiographic magnetic resonance imaging lesions in persons at increased risk of knee osteoarthritis," *Arthritis and Rheumatology*, vol. 66, no. 7, pp. 1811–1819, 2014.

[25] D. J. Hunter, A. Guermazi, G. H. Lo, A. J. Grainger, P. G. Conaghan, R. M. Boudreau, and F. W. Roemer, "Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score)," *Osteoarthritis and Cartilage*, vol. 19, no. 8, pp. 990–1002, 2011.

[26] U. Ahmed, A. Anwar, R. S. Savage, P. J. Thornalley, and N. Rabbani, "Protein oxidation, nitration and glycation biomarkers for early-stage diagnosis of osteoarthritis of the knee and typing and progression of arthritic disease," *Arthritis Research & Therapy*, vol. 18, no. 1, p. 250, 2016. [Online]. Available: http://arthritis-research.biomedcentral.com/articles/10.1186/s13075-016-1154-3

[27] L. Shamir, S. M. Ling, W. Scott, A. Bos, N. Orlov, T. J. MacUra, D. M. Eckley, L. Ferrucci, and I. G. Goldberg, "Knee X-ray image analysis method for automated detection of osteoarthritis," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 407–415, 2009.

[28] I. Boniatis, L. Costaridou, D. Cavouras, I. Kalatzis, E. Panagiotopoulos, and G. Panayiotakis, "Osteoarthritis severity of the hip by computer-aided grading of radiographic images," *Medical and Biological Engineering and Computing*, vol. 44, no. 9, pp. 793–803, 2006.

[29] M. C. Corporation, *Mammal Anatomy: An Illustrated Guide*. Marshall Cavendish, 2010. [Online]. Available: http://books.google.com/books?id=mTPI{_}d9fyLAC{&}pgis=1

[30] K. S. Saladin, C. A. Gan, and H. N. Cushman, *Anatomy & Physiology*. New York: McGraw-Hill Education, 2017.

[31] P. J. Thurner, *Tissue Biomechanics: TU Wien Lecture Notes*, 2017.

[32] A. Unsworth, D. Dowson, and V. Wright, "The Functional Behavior of Human Synovial Joints-Part I," *Journal of Lubrication Technology*, pp. 369–376, 1975.

[33] "Engineering ABC. Coefficient of friction, rolling resistance and aerodynamics." [Online]. Available: http://www.tribology-abc.com/abc/cof.htm

[34] J. Wolff, *The Law of Bone Remodelling*, 1987, vol. 155.

[35] J. G. Betts, P. Desaix, E. Johnson, J. E. Johnson, O. Koral, D. Kruse, B. Poe, J. A. Wise, M. Womble, and K. A. Young, *Anatomy & Physiology*, 2013. [Online]. Available: https://openstax.org/details/books/anatomy-and-physiology

[36] R. Wilson, "Bone," 2008. [Online]. Available: https://commons.wikimedia.org/wiki/User:Pbroks13

[37] D. Richfield, "Medical gallery of Blausen Medical 2014," *WikiJournal of Medicine*, vol. 1, no. 2, pp. 9–11, 2014. [Online]. Available: https://en.wikiversity.org/wiki/WikiJournal{_}of{_}Medicine/Medical{_}gallery{_}of{_}Blausen{_}Medical{_}2014

[38] D. B. Burr, M. B. Schaffler, and R. G. Frederickson, "Composition of the cement line and its possible mechanical role as a local interface in human compact bone," *Journal of Biomechanics*, vol. 21, no. 11, 1988.

[39] SEER, "Anatomy & Physiology. Skeletal System. Structure of Bone Tissue." 2011. [Online]. Available: https://training.seer.cancer.gov/anatomy/skeletal/tissue.html

[40] P. Fratzl, H. S. Gupta, E. P. Paschalis, and P. Roschger, "Structure and mechanical quality of the collagen–mineral nano-composite in bone," *J. Mater. Chem.*, vol. 14, no. 14, pp. 2115–2123, 2004. [Online]. Available: http://xlink.rsc.org/?DOI=B402005G

[41] A. Ascenzi and E. Bonucci, "The compressive properties of single osteons," *The Anatomical Record*, vol. 161, no. 3, pp. 377–391, 1968.

[42] P. J. Thurner, "Atomic force microscopy and indentation force measurement of bone," *WIREs Nanomedicine and Nanobiotechnology*, vol. 1, no. December, pp. 624–649, 2009.

[43] O. L. Katsamenis, T. Jenkins, and P. J. Thurner, "Toughness and damage susceptibility in human cortical bone is proportional to mechanical inhomogeneity at the osteonal-level," *Bone*, vol. 76, pp. 158–168, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.bone.2015.03.020

[44] N. Arden, F. Blanco, C. Cooper, A. Guermazi, D. Hayashi, D. Hunter, M. K. Javaid, F. Rannou, F. W. Roemer, and J.-Y. Reginster, *Atlas of Osteoarthritis*, 2014. [Online]. Available: https://books.google.com/books?id=qT1FBgAAQBAJ{&}pgis=1

[45] M. A. MacConaill, "Joint," 2017. [Online]. Available: https://www.britannica.com/science/joint-skeleton

[46] R. Wittenauer, L. Smith, and K. Aden, "Priority Medicines for Europe and the World " A Public Health Approach to Innovation " Update on 2004 Background Paper 6.12 Osteoarthritis," *World Health Organisation*, pp. 1–31, 2013. [Online]. Available: http://www.who.int/medicines/areas/priority{_}medicines/BP6{_}12Osteo.pdf

[47] J. W. J. Bijlsma, F. Berenbaum, and F. P. J. G. Lafeber, "Osteoarthritis: An update with relevance for clinical practice," *The Lancet*, vol. 377, no. 9783, pp. 2115–2126, 2011.

[48] J. H. Klippel, *Primer on the Rheumatic Diseases*, 13th ed., J. H. Stone, L. J. Crofford, and P. H. White, Eds. New York: Springer Science+Business Media, 2008.

[49] D. J. Hunter and T. D. Spector, "The role of bone metabolism in osteoarthritis." *Current rheumatology reports*, vol. 5, no. 1, pp. 15–9, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12590880

[50] P. G. Conaghan, H. Vanharanta, and P. A. Dieppe, "Is progressive osteoarthritis an atheromatous vascular disease?" *Annals of the rheumatic diseases*, vol. 64, no. 11, pp. 1539–41, 2005.

[51] D. D. Kumarasinghe, E. Perilli, H. Tsangari, L. Truong, J. S. Kuliwaba, B. Hopwood, G. J. Atkins, and N. L. Fazzalari, "Critical molecular regulators, histomorphometric indices and their correlations in the trabecular bone in primary hip osteoarthritis," *Osteoarthritis and Cartilage*, vol. 18, no. 10, pp. 1337–1344, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.joca.2010.07.005

[52] C. J. Menkes and N. E. Lane, "Are osteophytes good or bad?" *Osteoarthritis and Cartilage*, vol. 12, no. SUPLL., pp. 53–54, 2004.

[53] H. J. Braun and G. E. Gold, "Diagnosis of Osteoarthritis: Imaging," *Bone*, vol. 51, no. 2, pp. 278–288, 2012.

[54] J. H. Kellgren and J. S. Lawrence, "Radiological assessment of osteo-arthrosis." *Annals of the rheumatic diseases*, vol. 16, no. 4, pp. 494–502, 1957.

[55] W. Waldstein, G. Perino, S. L. Gilbert, S. A. Maher, R. Windhager, and F. Boettner, "OARSI osteoarthritis cartilage histopathology assessment system: A biomechanical evaluation in the human knee," *Journal of Orthopaedic Research*, vol. 34, no. 1, pp. 135–140, 2016.

[56] L. Shamir, S. M. Ling, W. Scott, M. Hochberg, L. Ferrucci, and I. G. Goldberg, "Early detection of radiographic knee osteoarthritis using computer-aided analysis," *Osteoarthritis and Cartilage*, vol. 17, no. 10, pp. 1307–1312, 2009.

[57] A. M. Rodrigues, N. Gouveia, L. P. da Costa, M. Eusébio, S. Ramiro, P. Machado, A. F. Mourão, I. Silva, P. Laires, A. Sepriano, F. Araújo, P. S. Coelho, S. Gonçalves, A. Zhao, J. E. Fonseca, J. M. de Almeida, V. Tavares, J. A. P. da Silva, H. Barros, J. Cerol, J. Mendes, L. Carmona, H. Canhão, and J. C. Branco, "EpiReumaPt- the study of rheumatic and musculoskeletal diseases in Portugal: a detailed view of the methodology," *Acta reumatologica portuguesa*, vol. 40, no. 2, pp. 110–124, 2015.

[58] R. Altman, E. Asch, D. Bloch, G. Bole, D. Borenstein, K. Brandt, W. Christy, T. D. Cooke, R. Greenwald, M. Hochberg, D. Howell, D. Kaplan, W. Koopman, S. Longley, H. Mankin, D. J. McShane, T. Medsger, R. Meenan, W. Mikkelsen, R. Moskowitz, W. Murphy, B. Rothschild, M. Segal, L. Sokoloff, and F. Wolfe, "Development of criteria for the classification and reporting of osteoarthritis: Classification of osteoarthritis of the knee," *Arthritis & Rheumatism*, vol. 29, no. 8, pp. 1039–1049, 1986.

[59] N. A. Segal, M. C. Nevitt, K. D. Gross, J. Hietpas, N. A. Glass, C. E. Lewis, and J. C. Torner, "The multicenter osteoarthritis study: Opportunities for rehabilitation research," *PM and R*, vol. 5, no. 8, pp. 647–654, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.pmrj.2013.04.014

[60] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[61] J. F. Veenland, J. L. Grashius, F. van Der Meer, A. L. Beckers, and E. S. Gelsema, "Estimation of fractal dimension in radiographs." *Medical Physics*, vol. 23, no. 4, pp. 585–94, 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8860906

[62] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[63] A. Pentland, "Fractal based description of natural scenes," *IEEE Transactions on Pattern analisys and Machine Intelligence*, vol. 6, no. 6, pp. 661–672, 1984.

[64] P. Campbell and S. Abhyankar, *Fractals, form, chance and dimension.* Freeman. San Francisco, 1978, vol. 1, no. 1. [Online]. Available: http://link.springer.com/10.1007/BF03023043

[65] B. B. Mandelbrot, *The Fractal Geometry of Nature.* W. H. Freeman and Company, 1983.

[66] K. Falconer, *Fractal Geometry.* Wiley, 2004. [Online]. Available: http://doi.wiley.com/10.1002/0470013850

[67] B. B. Mandelbrot, "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," *Science*, vol. 156, no. 3775, pp. 636–638, 1967. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.156.3775.636

[68] H. E. Hurst, "Long-term storage capacity of reservoirs," *Trans American Society of Civil Engineers*, vol. 116, pp. 770–799, 1951.

[69] H. E. Hurst, R. P. Black, and Y. M. Simaika, *Long-term storage: An experimental study.* Constable, 1965.

[70] T. Gneiting and M. Schlather, "Stochastic models which separate fractal dimension and Hurst effect," *SIAM Review*, vol. 46, no. 2, pp. 269–282, 2001. [Online]. Available: http://arxiv.org/abs/physics/0109031

[71] B. Mandelbrot and R. L. Hudson, *The (Mis)behavior of Markets: A Fractal View of Financial Turbulence*, annotated ed. Hachette UK, 2007.

[72] IDBHD, "Mandelbrot, Fractal Geometry, & The Language of Creation," 2018. [Online]. Available: http://idontbuthedoes.com/mandelbrot-fractal-geometry-the-language-of-creation/

[73] Science Photo Library, "Fractals in Nature." [Online]. Available: www.sciencephoto.com

[74] F. Foundation, "Fractal Dimension." [Online]. Available: https://fractalfoundation.org

[75] G. Shevchenko, "Fractional Brownian motion in a nutshell," no. 1, pp. 1–14, 2014. [Online]. Available: http://arxiv.org/abs/1406.1956

[76] I. H. Parkinson and N. L. Fazzalari, "Methodological principles for fractal analysis of trabecular bone," *Journal of Microscopy*, vol. 198, no. 2, pp. 134–142, 2000.

[77] 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*, vol. 18, no. 5, pp. 684–690, 2010.

[78] P. Moerters and Y. Peres, *Brownian motion*. Cambridge University Press, 2010.

[79] R. Harba, H. Douzi, and M. El Hajiji, "Lecture Notes in Computer Science: Maximum Likelihood Estimation, Interpolation and Prediction for Fractional Brownian Motion," in *ICISP*. Springer-Verlag Berlin Heidelberg, 2012.

[80] T. Carter, "An Introduction to Information Theory and Entropy," California State University Stanislaus, Tech. Rep., 2007.

[81] S. Bino, "Grey Level Co-Occurrence Matrices: Generalisation and Some New Features," *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 2, pp. 151–157, 2012.

[82] M. Hall-Beyer, "Glcm Texture: a Tutorial," no. February, p. 75, 2017. [Online]. Available: https://prism.ucalgary.ca/bitstream/1880/51900/1/texturetutorialv3{_}0170329.pdf

[83] J. C. Russ, *Fractal Surfaces*, 1st ed. Springer US, 1994.

[84] S. image development team, "GLCM Texture Features." [Online]. Available: http://scikit-image.org/docs/dev/auto{_}examples/features{_}detection/plot{_}glcm.html

[85] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences.* Elsevier Inc., 1977.

[86] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[87] N. M. Razali and Y. B. Wah, "Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[88] G. W. A. Snedecor and W. G. A. Cochran, *Statistical Methods.* Iowa State University Press, 1967.

[89] H. Levene, "Robust tests for equality of variances." in *Contributions to probability and statistics.* Stanford Univ. Press, 1960, pp. 278–292.

[90] J. Fisher Box, "Guinness, Gossett, Fisher, and small samples," *Statistical Science*, vol. 2, pp. 45–52, 1987.

[91] H. Lohninger, "Kombination mehrerer Einzelverteilungen," 2012. [Online]. Available: http://www.statistics4u.info/fundstat{_}germ/dd{_}distributions{_}combi.html

[92] "Repeated Measures ANOVA," 2018. [Online]. Available: https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

[93] J. W. Tukey, "Comparing Individual Means in the Analysis of Variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.

[94] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[95] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[96] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *Journal of Multivariate Analysis*, vol. 100, no. 1, pp. 175–194, 2009.

[97] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, 1995. [Online]. Available: http://ieeexplore.ieee.org/document/598994/

[98] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space,” *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14786440109462720

[99] M. Blagojevic, C. Jinks, A. Jeffery, and K. P. Jordan, “Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis,” *Osteoarthritis Cartilage*, vol. 18, no. 1, pp. 24–33, 2010.

[100] R. K. Chaganti and N. E. Lane, “Risk factors for incident osteoarthritis of the hip and knee,” *Curr Rev Musculoskelet Med*, vol. 4, no. 3, pp. 99–104, 2011. [Online]. Available: http://link.springer.com/article/10.1007/s12178-011-9088-5

[101] D. Kumar, K. T. Manal, and K. S. Rudolph, “Knee joint loading during gait in healthy controls and individuals with knee osteoarthritis,” *Osteoarthritis and Cartilage*, vol. 21, no. 2, pp. 298–305, 2013.

[102] P. C. O’Brien and T. R. Fleming, “A Multiple Testing Procedure for Clinical Trials,” *Biometrics*, vol. 35, no. 3, p. 549, 1979. [Online]. Available: http://www.jstor.org/stable/2530245?origin=crossref

[103] D. T. Felson, J. Nui, A. Guermazi, B. Sack, and P. Aliabadi, “Defining radiographic incidence and progression of knee osteoarthritis: suggested modifications of the Kellgren and Lawrence scale,” *Annals of the Rheumatic Diseases*, vol. 70, pp. 1884–1886, 2011.