Texture Analysis Improves the Estimate of Bone Fracture Risk from DXA Images

A thesis submitted for the degree of

Doctor of Philosophy

By

Rui-Sheng Lu

School of Computer Science, Engineering and Mathematics

Faculty of Science and Engineering

Flinders University

31, March, 2017

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AAM	Active appearance model
aBMD	Areal bone mineral density
AUC	Area under the ROC curve
ASM	Active shape model
DXA	Dual energy X-ray absorptiometry
FLDA	Fisher's linear discriminant analysis
FNF	False negative fraction
FPF	False positive fraction
FRAX	Fracture risk assessment model
HAL	Hip axis length
NSA	Neck shaft angle
PCA	Principal component analysis
PDM	Point distribution model
ROC	Receiver operating curve
ROI	Regions of interest
SVM	Support vector machine
TBS	Trabecular bone score
TNF	True negative fraction
TPF	True positive fraction
VFA	Vertebral fractures assessment

SUMMARY

Elderly people are at a significantly higher risk of suffering a bone fracture as a result of a fall than the ambient population. Accurate prediction of fracture risk allows for preventative intervention and reliable advice on lifestyle. Traditionally, dual energy X-ray absorptiometry (DXA) is used to assess fracture risk. This technique allows the calculation of areal bone mineral density (aBMD), T-score and geometric parameters commonly used to assess risk of fracture. However, these measures may not fully exploit the information content available in DXA images regarding risk of fracture as there are still several limitations to the way images are currently analysed. First, aBMD is not an accurate measurement of true bone mineral density because it measures area rather than volume of bone. Second, bone density is averaged over the entire image or over specified regions of interest (ROI) and ignores local information. Third, only a few discrete geometric measures are usually considered rather than full shape information. Finally, density and geometric information are analysed separately.

In this study, an active shape model (ASM) and an active appearance model (AAM) were used to allow a quantitative characterization of the shape and gross structure of the proximal femur. These models provide a level of risk assessment comparable to conventional risk measures such as BMD and T-score. In order to improve risk assessment, these methods were augmented with image texture analysis methods, including Gabor filters and textons applied to various ROI. Texture methods allow quantification of structure patterns that have not been considered previously in assessing risk of bone fracture. To evaluate these methods, we analysed hip DXA scans from the Osteoporosis Centre of Southampton General Hospital. The data consisted of 29 DXA scans from subjects with a history of fragility fracture and 90 DXA scans from subjects with no known fractures. Feature selection was used to determine which method, or combination of methods, was best to discriminate between the fracture and control groups. The data was separated into two, roughly equal sets, each containing similar ratios of fracture and non-fracture examples. One set was used to develop a new scheme for estimating risk of fracture and the other set was used to measure performance of the risk scheme.

Results showed that by including texture information based on Gabor filters and focusing on a specific region of the image (the whole femoral neck), better risk assessment was possible than using either aBMD or T-score alone. Thus the main conclusion of this work is that DXA scans include more information regarding fracture risk than is normally exploited and, in particular, that including texture information has the potential to improve estimates of fracture risk.

LISTS OF PUBLICATIONS AND PRESENTATIONS

Journal Paper

Rui-Sheng Lu, Elaine Dennison, Hayley Denison, Cyrus Cooper, Mark Taylor, and Murk J. Bottema, 2017, *Texture Analysis Improves the Estimate of Bone Fracture Risk from DXA Images*. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (ID: 1271726 DOI:10.1080/21681163.2016.1271726).

Conference Presentation

R. S. Lu and M. Taylor and M.J. Bottema, 2016, "Texture Analysis Improves the Estimate of Bone Fracture Risk from DXA Images, "The 2016 OARSI World Congress on Osteoarthritis-Promoting Clinical and Basis Research in Osteoarthritis", 2006:S319-S320.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Rui-Sheng Lu, Candidate

Murk J. Bottema, Principal Supervisor
Mark Taylor, Co-supervisor

ACKNOWLEDGEMENTS

I praise God, the almighty for providing me this opportunity and granting me the capability to complete my Ph.D. thesis. The completion of this thesis is due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

• First and foremost, my beloved wife Celine (潘莉玲) and baby Jerry (呂杰睿), for being there when I needed it most. I could not have completed this without both of you. All of my family, for their constant support throughout my entire candidature. Thank you especially to my beloved parents, 呂欽煌 and 余碧雲, for all their great support and love all the way through.

• My supervisor A/Prof. Murk Bottema and co-supervisor Prof. Mark Taylor for their great support and encouragement through my Ph.D. study. Importantly, they helped me improve my research skills in medical image analysis, academic writing and mathematics theories for my project.

• Dr. Gobert Lee, Dr. Mariusz Bajger, and Dr. Simon Williams, and for their support in in academic fundamental knowledge, suggests for my research. Dr. Simon Williams, Dr. Darfiana Nur, and Dr. Sherry Randhawa for giving me the opportunity of doing some mathematics, statistics, and electronics related part-time jobs in the university.

• My dearest friends, Brian and Ruth Fagan. I would like to express my heartfelt thanks to both of you for your warm encouragement and your spiritual supports during my Ph.D. study.

• A special thanks to the participants of the Hertfordshire Cohort Study and their General Practitioners for providing images for my research.

• Finally, to everyone whom I have neglected too much whilst writing this thesis, I apologise.

Rui-Sheng Lu

March 2017, Adelaide, Australia

1 INTRODUCTION

This thesis presents a study on using image analysis and machine learning methods to improve the accuracy of assessing the risk of fracture. In this chapter, information on the incidence of bone fracture and current clinical practice in estimating risk is presented in order to motivate the study and provide context.

1.1 Osteoporosis and Bone Fractures Incidence

Many elderly people, especially women with low bone density, are at a significantly higher risk of suffering fracture when compared to healthy young adults [1, 2]. In particular, fracture is the leading cause of morbidity and mortality among elderly people worldwide [3, 4]. Fractures among the elderly have become a major public health and policy problem because of their prevalence, health-care costs, and health effects [4-7]. A common cause of fracture in the elderly is osteoporosis, a condition marked by low bone mass. Osteoporosis can lead to a weakening of the bone architecture and increased susceptibility to fracture. In other words, the prevalence of an osteoporotic fracture significantly increases the risk of further osteoporotic fracture [4, 8-11]. Although osteoporosis affects the entire skeleton, osteoporotic fractures predominantly occur at the hip, the wrist, and the vertebrae [4, 8, 9].

Approximately 98% of hip fractures occur among people aged 35 years and older, and the incidence of hip fracture in most populations increases dramatically with age [12]. In Australia, the number of patients sustaining hip fractures each year is predicted to increase by 15% every five years until 2036, and then by about 10% every five years until 2051 [13]. The annual cost of treatment in Australia for atraumatic fractures occurring in people older than 60 years of age was AUD 779 million in 1992, and the direct costs for hip fracture alone is expected to double in most Western countries by 2025 [14]. Approximately 340,000 people over 65 years old suffer from hip fracture annually in the United States [15], and the cost of a hip fracture is estimated at more than USD 8.5 billion annually [16-18]. In the UK, around 80,000 patients with hip fracture are treated each year, and the estimated annual cost, including healthcare costs, is approximately GBP 2 billion [19, 20]. Worldwide, the total number of hip fractures

is expected to more than triple from 1.66 million in 1999 to 6.26 million in the year 2050 [21], and the annual cost of hip fracture is projected to reach USD 131.5 billion by 2050 [22].

With regard to other fracture sites in the body, it is estimated that 1.4 million clinical vertebral fractures and 1.7 million wrist fractures occurred globally in 2000. In addition, in the year 2000, 4.3 million out of 9 million osteoporotic fractures were at sites other than hip, spine and wrist [23]. Non-vertebral, non-hip fractures include fractures of the ribs, pelvis, humeral shaft, proximal humerus, clavicle, scapula, sternum, tibia, fibula, distal forearm, and femoral fractures other than hip. It was observed in the UK general practice research database (GPRD) study that survival of women five years after vertebral fracture was 56.5% for the general population of England and Wales during the period 1988-1998 [24].

1.2 Current Clinical Practice

1.2.1 Image Based Risk Assessment

Estimates of risk allow better advice to be provided on lifestyle and planning of care [25]. For this reason, estimating risk of fracture has received much attention. Early detection of osteoporosis or fracture can be done using aBMD measured by DXA. DXA measures the reduction in intensity of both high- and low-energy X-ray beams through the entire body or a specific region of the body. The attenuation depends on the energy of the X-rays as well as the density of the tissue (bone mineral and soft tissue). For the time being, DXA is considered to be the primary means for measuring aBMD.

In general, aBMD is obtained for any skeletal site, with each site having unique information to offer, but clinical use has concentrated on specific regions of interest (ROI), e.g., spine, proximal femur, forearm and total body [26-28]. The International Society for Clinical Densitometry (ISCD) has recommended measuring aBMD at both posterior-anterior spine and hip [29]. The ROI on the proximal femur are typically the total proximal femur, femur neck, intertrochanteric region, trochanter, and Ward's triangle [29-35]. Evidence has suggested that the femoral neck or total proximal femur are the optimum sites for predicting the risk of hip fracture as well as for predicting

the overall risk for any type of fracture [1, 2, 29, 36].

DXA scans can also be used to measure the geometric parameters of bone. DXA is the technique most widely used in current clinical practice because it is easy-to-use and economical [37]. However, different bone densitometry units will yield different aBMD results due to the differences in scanning methods, manufacturers, calibration skills, detector types, software programs, and ROI considered [38, 39].

Low aBMD is an important indicator of the risk of osteoporotic fracture [1, 2]. The diagnosis of osteoporosis is based on the assessment of aBMD and defined by T-score. The T-score is defined as the number of standard deviations (SD) of bone density above the bone density of an average healthy young adult of the same sex and ethnicity as the patient [40].

The World Health Organisation (WHO) has proposed four general diagnostic categories for assessment based on T-score. Subjects are classified as: normal and healthy if hip aBMD is greater than 1 SD below the young adult mean (T-score $\geq$ -1); osteopenic if hip aBMD is between 1 SD below the young adult mean and greater than 2.5 SD below the young adult mean (-2.5 < T-score < -1); osteoporotic if hip aBMD is 2.5 SD or more below the young adult mean (T-score $\leq$ -2·5); and severe (established) osteoporotic if hip aBMD 2.5 SD or more below the young adult mean, and the subject has previously suffered an osteoporotic fracture [40]. aBMD has been shown to be an important predictor of fracture risk and is used to describe the bone quantity. However, there is increasing evidence supporting the view that the material and structural basis of bone are also critical in determining resistance to fracture [36].

Hip axis length (HAL), the distance along the femoral neck axis from the base of the greater trochanter to the inner pelvic brim, has been reported to be predictive of hip fracture [41-43]. However, this is not recommended in clinic practice as there are no widely accepted thresholds [44, 45]. Recently, hip structural analysis (HSA), a more advanced method, has been developed to extract not only the BMD of the hip, but also structural geometry of cross-sections traversing the proximal femur at specific locations from DXA-derived images of the hip, including femoral neck cross-sectional moment of inertia (CSMI) and cross-sectional area (CSA) [46]. Therefore, the main advantages of using HSA are that BMD and bone structural geometry, both of which

contribute to bone strength, are taken into account. However, the current HSA method is of limited value in evaluating bone structure strength as its precision is strongly dependant on proximal femur positioning. Additionally, the structural parameters are highly correlated with BMD. These limitations of HSA primarily reflect the limitations imposed by the two dimensional (2D) nature of DXA [47-49].

The prevalence of an osteoporotic fracture, like vertebral fracture, substantially increases the risk for additional osteoporotic fractures. Vertebral fractures assessment (VFA), a quick, non-invasive, low radiation technique, examines lateral DXA images of the spine to screen for the presence of osteoporosis and vertebral fractures. VFA requires only a modest amount of time and cost as the images can be obtained concurrently while measuring aBMD at the same location by DXA [50, 51]. However, not all vertebrae are readable by VFA and VFA is not sensitive to mild fractures [52-54]. Thus, the effectiveness of VFA is limited.

In summary, DXA scans are readily available and valuable for the diagnosis of bone density abnormalities and osteoporosis. However, it seems that there are still widely acknowledged limitations and disadvantages with DXA. In particular, the measurement of aBMD alone using DXA scans may not tell the whole story and, therefore, it is important to include more information regarding fracture risk than is normally exploited.

1.2.2 Full Information Based Risk Assessment

To improve the prediction of fracture risk, the country specific fracture risk assessment model (FRAX) was developed by the WHO in 2008. FRAX uses several clinical risk factors in addition to aBMD to provide a prediction tool for assessing the risk and probabilities of hip and major osteoporotic fracture over the next 10 years for postmenopausal women and men aged 40 to 90 years. The risk can be calculated from clinical risk factors with or without the measurement of femoral neck aBMD. The clinical risk factors included in the FRAX model are: age, sex, weight, height, previous fracture, parental hip fracture, current smoking habits, glucocorticoids, rheumatoid arthritis, secondary osteoporosis and alcohol intake [55, 56]. There are a number of limitations to FRAX. First, FRAX does not take into account all risk variables that a

physician could reasonably be aware of such as other bone density assessments. Second, FRAX uses only yes/no answers for most clinical risk factors rather than seeking answers to open ended questions, such as the quantity of alcohol or tobacco consumption. Finally, FRAX does not take into consideration whether the subject is currently being treated, and variations of fracture rates within countries [57].

Trabecular bone score (TBS) has recently been introduced to extract further information of bone strength by analyzing the DXA lumbar spine image. The TBS is a structural parameter that quantifies the local variations in pixel grey-level in DXA images of the lumbar spine, and its variations may correlate with three dimensional (3D) bone microarchitecture which, in turn, is correlated with the mechanical strength of bone. Thus, it provides skeletal information that is not captured from the standard aBMD measurement [58-61]. The higher the TBS value, the better the microstructure of the skeletal bone and, thus, the lower risk of osteoporosis and fracture risk [62]. TBS can be used in conjunction with FRAX to adjust risk assessment and help improve the prediction of risk of fracture in clinical practice [63]. However, there are a number of limitations to TBS. For example, TBS correlates with, but does not measure, bone microarchitecture due to the effect of spondylosis. TBS is also limited by spatial resolution [64].

1.3 Objectives

The most common method of assessing fracture risk is DXA, which is used to calculate aBMD and geometric parameters. However, there are still several limitations to the way DXA images are currently analysed. In particular, only discrete measures are taken of bone density or morphology, potentially ignoring a wealth of information contained within the image [65-69].

Methods exist for assessing risk from DXA images by analysing the quantitative characterization of the shape and gross structure of the femoral neck. These are referred to as statistical shape and appearance models. These models provide a level of risk assessment comparable to conventional risk measures such as aBMD and geometric parameters [70-73]. However, little has been done to explore whether additional texture information collected as part of femoral DXA scans will improve

the prediction rate. In principle, texture analysis methods allow quantification of structure patterns that have not been previously considered in assessing risk of fracture. In addition, previous studies on risk assessment based on femoral DXA images have focussed on femoral fracture only. Ideally, analysis of DXA scans should provide information on risk of fracture anywhere in the body and not just for the region covered by the DXA scans.

The objective of this thesis is to determine if image texture information computed from DXA images of the femoral neck contributes to improved estimates of low-energy fracture risk throughout the body. A low-energy fracture is defined as a fracture resulting from minimal trauma, falling from standing height or less, rather than any other type of trauma such as motor vehicle accident [74]. In particular, the objective is to determine if such texture information in combination with aBMD and standard shape and appearance measures provides a better estimate of fracture risk than aBMD plus standard shape and appearance measures alone.

1.4 Contribution of the Thesis

The main contribution of the thesis is demonstration that texture does provide information regarding fracture risk beyond the information provided by aBMD, T-score, and standard shape and appearance measures.

The thesis also demonstrates that the information obtained from DXA images of the femoral neck applies to fracture risk at locations other than the femoral neck.

In addition, the thesis demonstrates that when comparing or reporting the performance of new risk assessment methods, it is necessary to measure the performance on a set of testing data that is independent of the dataset used to train the method. While this practice is common in many areas of machine learning and in computer aided medical image analysis, this has not been the standard in the field of fracture risk prediction based on DXA images.

1.5 Overview of the Thesis

This thesis is organised into five chapters including this current introductory chapter.

Chapter 2 provides the background knowledge required to understand this study. It contains a literature review of relevant history and work, other researchers' views, and the key technical theories and practice used in the thesis. These include statistical shape and appearance modelling, texture analysis using Gabor filters and textons, feature subset selection, classification, and ROC analysis.

An overview of the datasets used for conducting the experiments in this thesis is presented in Chapter 3. In addition, Chapter 3 describes the experimental details of implementing the methods described in Chapter 2 to arrive at a scheme for estimating fracture risk that includes shape, appearance and image texture.

Chapter 4 reports the experimental results of this work. Baseline characteristics of all subjects in terms of age, aBMD, T-score and geometry are provided as the benchmark. Then, the classification performance on different ROI using individual methods and combinations of methods are presented.

Chapter 5 draws together the findings of this study. This chapter contains a discussion in which the experimental results are examined critically in the light of the previous work. Finally, perspectives for future work are given.

2 LITERATURE REVIEW AND TECHNICAL BACKGROUND

This chapter provides relevant history and necessary technical background to understand this thesis. The structure and function of the hip and femur, and the types of fractures that occur are explained in Sections 2.1 and 2.2. Some principal assessment methods such as fracture risk prediction based on aBMD and geometry are described in Section 2.3. Section 2.4 introduces two deformable models, ASM and AAM, for characterising shape and appearance. Gabor filters and textons are methods of texture analysis used in this thesis and are reviewed in Section 2.5. Section 2.6 provides an overview of classification and Fisher's linear discriminant analysis. ROC analysis is reviewed in Section 2.7 and techniques for selecting an optimal subset of features from a large collection of features are described in Section 2.8.

2.1 Anatomy of Hip and Femur

The hip is a ball-and-socket joint, consisting of the ball-shaped femoral head, which rotates within the socket-shaped acetabulum (Figure 2.1). Strong ligaments and muscles, including the ligamentum teres and the transverse acetabular ligaments, surround the hip joint to help reinforce this structure and make it more stable by holding the femoral head in the socket. The hip joint moves thousands of times during daily activities. During routine activities of daily living the hip joint experiences forces which are several multiples of body weight. For instance, the hip joint experiences forces which are twice or three times body weight during slow walking and four or five times during fast walking [75]. The load increases to over seven times body weight during stair climbing [76].

The femur is nearly cylindrical over most of its length and is the longest and strongest bone in the skeleton. The femur is an important part of the skeletal structure because the femur not only transmits the load from the acetabulum to the tibia but also helps the major muscles to control and stabilise the motions of the hip and knee joints. All the forces acting on the femur cause both the internal microstructure and external geometry of the femur to change along its length [77].

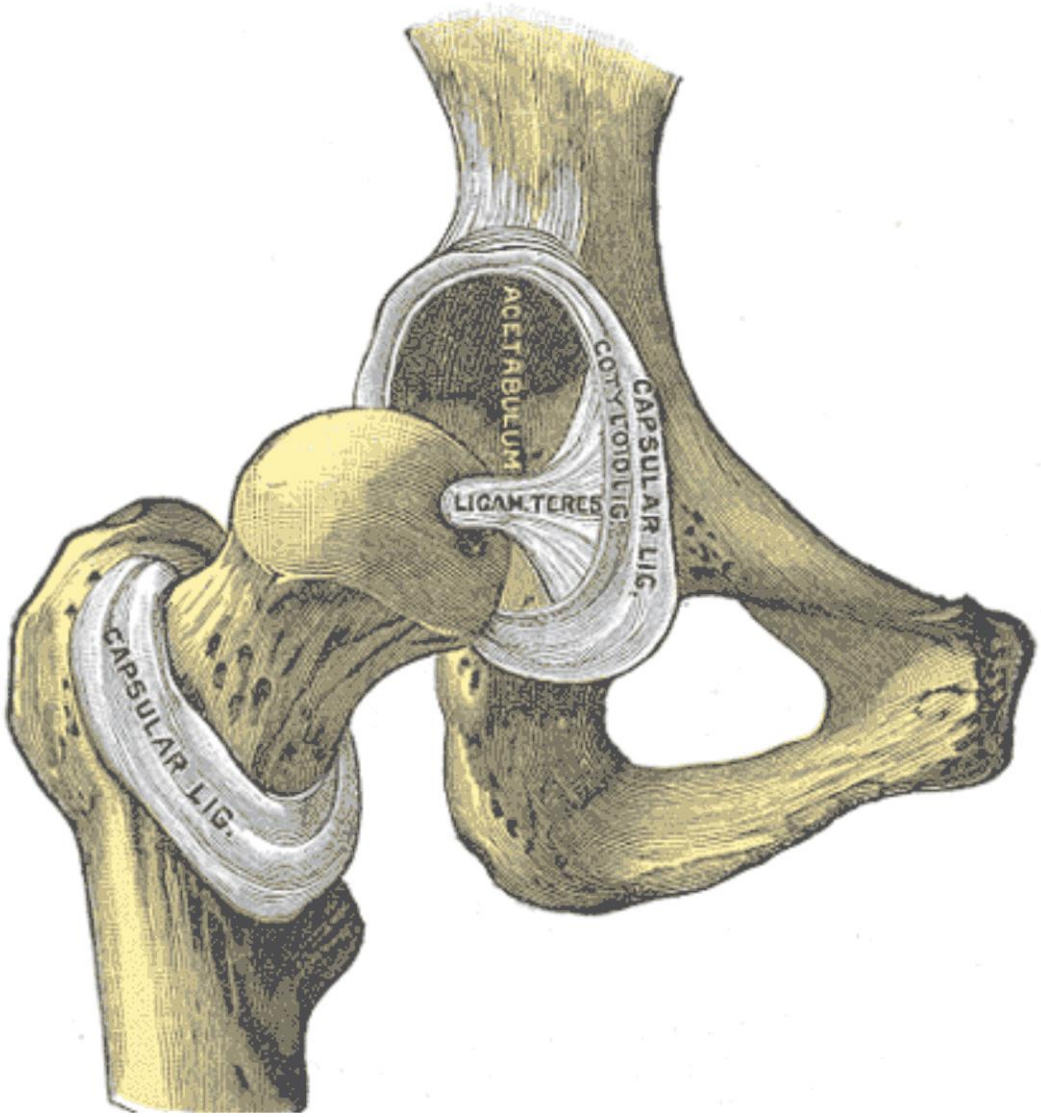


Figure 2.1 : The hip joint [77] (Reproduced with permission).

The femur consists of three parts: the femoral shaft, the proximal femur, and the distal femur. The major features of the proximal femur are the femoral head, femoral neck, and the greater and lesser trochanters (Figure 2.2). The neck joins the head to the body of the femur, and it merges with the lesser trochanter at its inferior limit and with the base of the greater trochanter at its lateral limit. The trochanters are irregularly shaped with rough surfaces and vary greatly in form from person to person. The greater trochanter in an adult is approximately 1cm lower than the head at its superior point, and the lesser trochanter is located at the lower and posterior part of the base of the neck. The trochanters are points at which hip and thigh muscles attach to support the muscles controlling rotation of the thigh. The greater and lesser trochanter are sights where major muscles attach to the femur [77, 78].

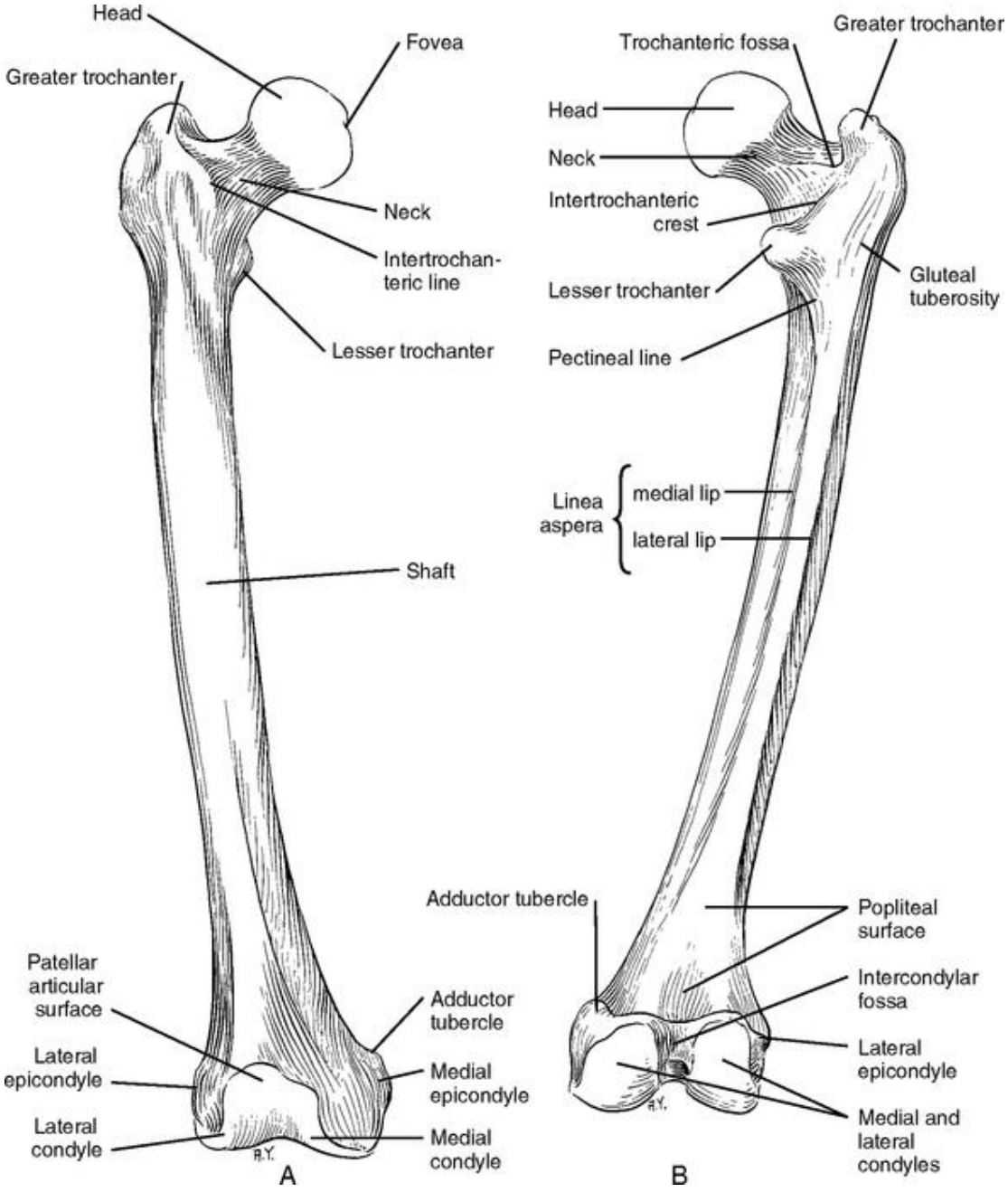


Figure 2.2 : Anterior (A) and posterior (B) views of the femur [78] (Reproduced with permission).

Figure 2.3 shows the frontal longitudinal midsection of the femur. The femur is a bone with one shaft and two ends. The ends of the femur contain spongy bone material and the epiphyseal line. Between the two ends, the shaft of the femur consists of a roughly cylindrical, hollow tube of thick cortical bone, and a central area with bone

marrow. The full structure of the bone represents a balance between weight and strength [79].

The femoral neck, which is located near the top of the femur bone, is especially susceptible to fracture because it is the weakest part of the femur [80]. It is a cylinder, contracted in the middle, and broader laterally than medially. Some investigators have examined the bone structure and histology of the femoral neck, and demonstrated that there are changes in the structural features of the femoral neck associated with fracture [81-83]. For instance, Bell et al. investigated regional changes in both cortical and cancellous bone from cross-sections of the femoral neck in cases of fracture, in comparison with a control group, and found that in intracapsular fracture of the femoral neck, loss of cortical rather than cancellous bone is the predominant feature [81].

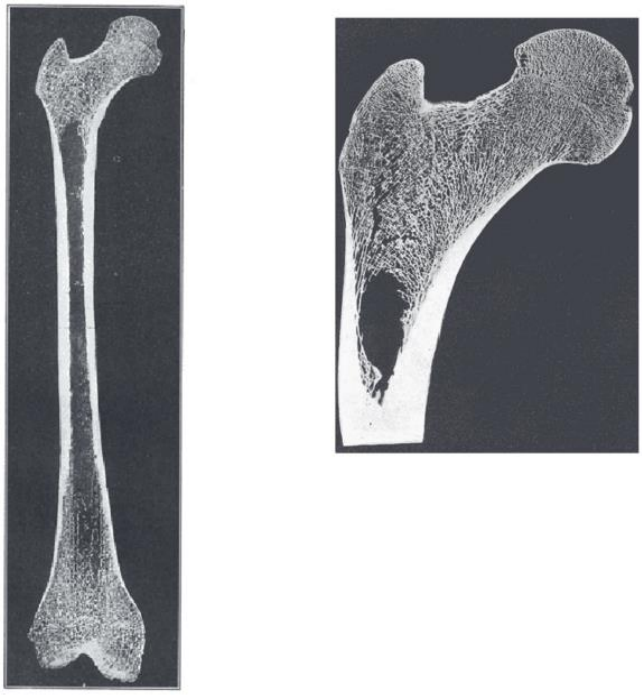


Figure 2.3 : Frontal longitudinal midsection of left femur (Left) and upper femur (right) [79] (Reproduced with permission; Copyright © 1999-2016 John Wiley & Sons, Inc. All Rights Reserved).

2.2 Bone Fracture

Almost all of the adult body's 206 bones can experience various types of bone fractures, ranging from minor inconveniences to severe, life-threatening fractures. However, in the elderly, osteoporotic population fractures are more likely to occur at certain sites. The most common osteoporotic fractures are of the proximal femur (hip), vertebrae (spine), and distal forearm (hand wrist) [4, 8, 9].

Hip fractures are usually classified onto three broad categories in terms of anatomic locations. Intracapsular fractures (subcapital or transcervical neck fracture) occur below the ball of the hip joint. Intertrochanteric fractures generally cross in the area between the greater and the lesser trochanters. Subtrochanteric fractures are located between the lesser trochanter and the femoral isthmus (Figure 2.4). In general, subtrochanteric fractures, greater trochanter fractures, and lesser trochanter fractures are less common than the others [84].

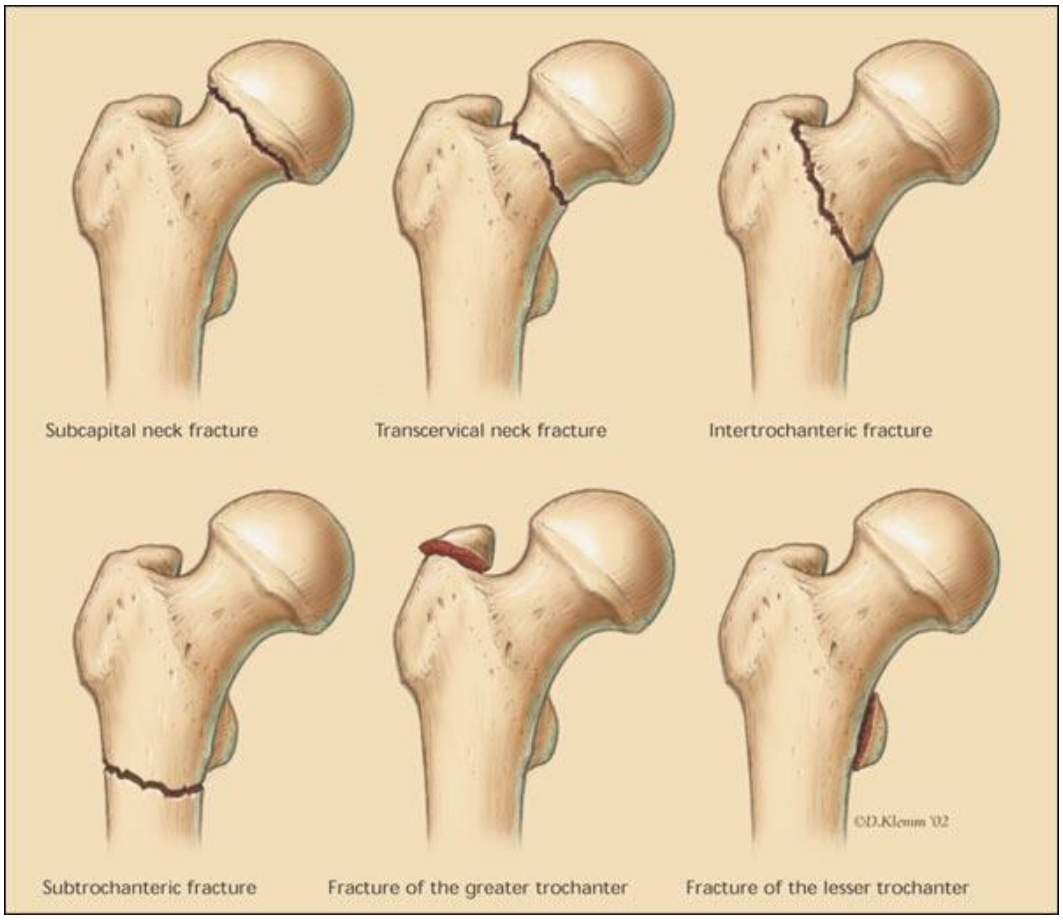


Figure 2.4 : Types of hip fractures [84] (Reproduced with permission).

Falls on an outstretched hand are the most common cause of fractures of the wrist (hyperextension), but any sufficiently strong force on the hand can break the wrist. The most common wrist fractures resulting from a fall onto an outstretched hand are scaphoid fractures, Colles fractures (“distal radius fractures”) and lunate dislocation fractures. Scaphoid fracture is the most common type of bone fracture in the wrist. These fractures occur through the scaphoid bone, a wedge-shaped bone located on the thumb side of the wrist, just where it meets the radius. Colles fracture is a fracture to the lower end of the radius (Figure 2.5) [85, 86].

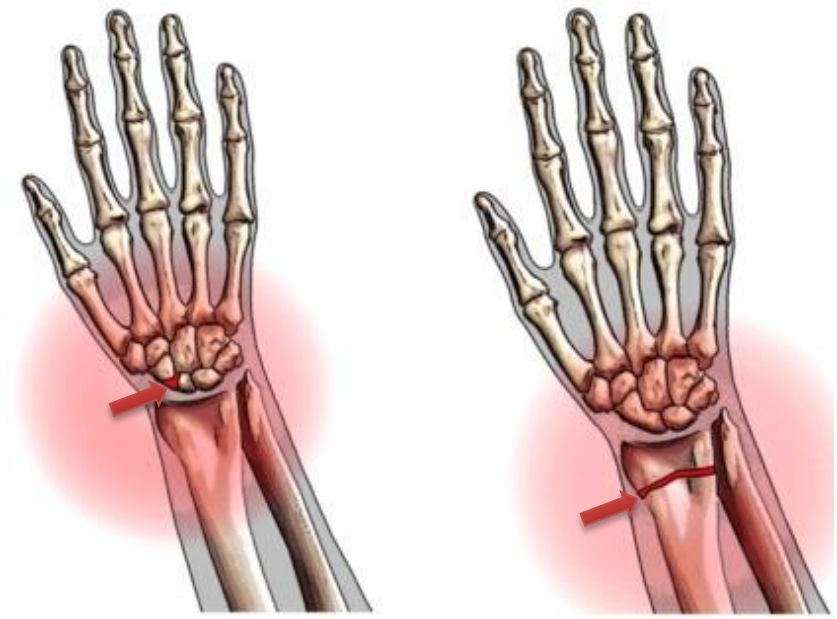


Figure 2.5 : Types of wrist fractures - Scaphoid fracture (Left) and Colles fracture (Reproduced with permission).

Vertebral fracture occurs when individual vertebrae become so weak that they deform and collapse. Vertebral fractures do not usually require hospitalization. Vertebral deformities may be caused by a variety of conditions, such as osteoporosis, severe trauma, congenital deformities, Schenermann's disease, osteoarthritis, and multiple myeloma. Vertebral fractures are classified by type of deformity (wedge, biconcavity, or compression) and further by the degree of deformity (grades 1 and 2). Radiographs of the thoracic and lumbar spine are the standard tools for diagnosing vertebral fractures as most fractures occur in these two locations or at the connection between them (thoracolumbar junction) (Figure 2.6) [87-89].

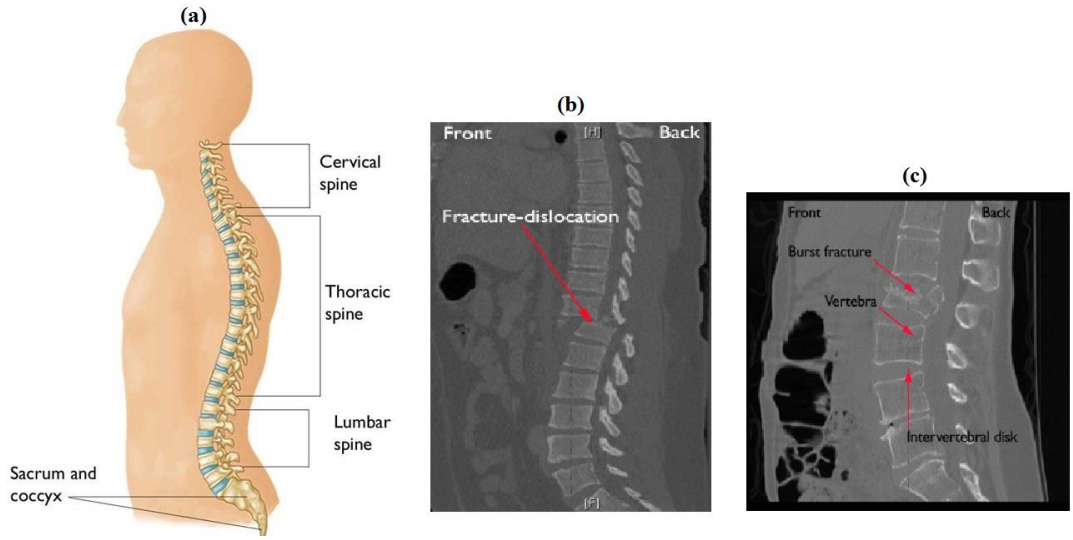


Figure 2.6 : The spine and spinal fractures. (a) The regions of the spine. (b) A fracture-dislocation in the thoracic spine. (c) A burst fracture in the lumbar spine. (Reproduced with permission from *OrthoInfo*. © American Academy of Orthopaedic Surgeons. http://orthoinfo.aaos.org).

2.3 aBMD and Geometry

The most common methods of assessing the risk of bone fracture are according to bone density and geometry captured from DXA scanned images. Calculation of aBMD from DXA is viewed by many as the current clinical gold standard for the assessment of the risk of fractures associated with osteoporosis. DXA scans use two X-ray energies to produce a two-dimensional image of the bone. As the X-ray photon passes through three types of tissue (bone mineral, lean tissue, and adipose tissue), X-ray energy is absorbed, attenuated or transmitted at different rates by different types of tissue. Thus, quantifying the degree of attenuation is to quantify the sum of tissue density over the path of the photon. aBMD allows the calculation of bone mineral content (BMC) in grams. This allows the BMD to be computed in g/cm^2 given that the two-dimensional projected area of the bone is measured in cm^2. The majority of clinical aBMD measurement performed using DXA scans are at the spine and hip. The hip regions of interest include the femoral neck, trochanter, and total hip (Figure 2.7) [33].

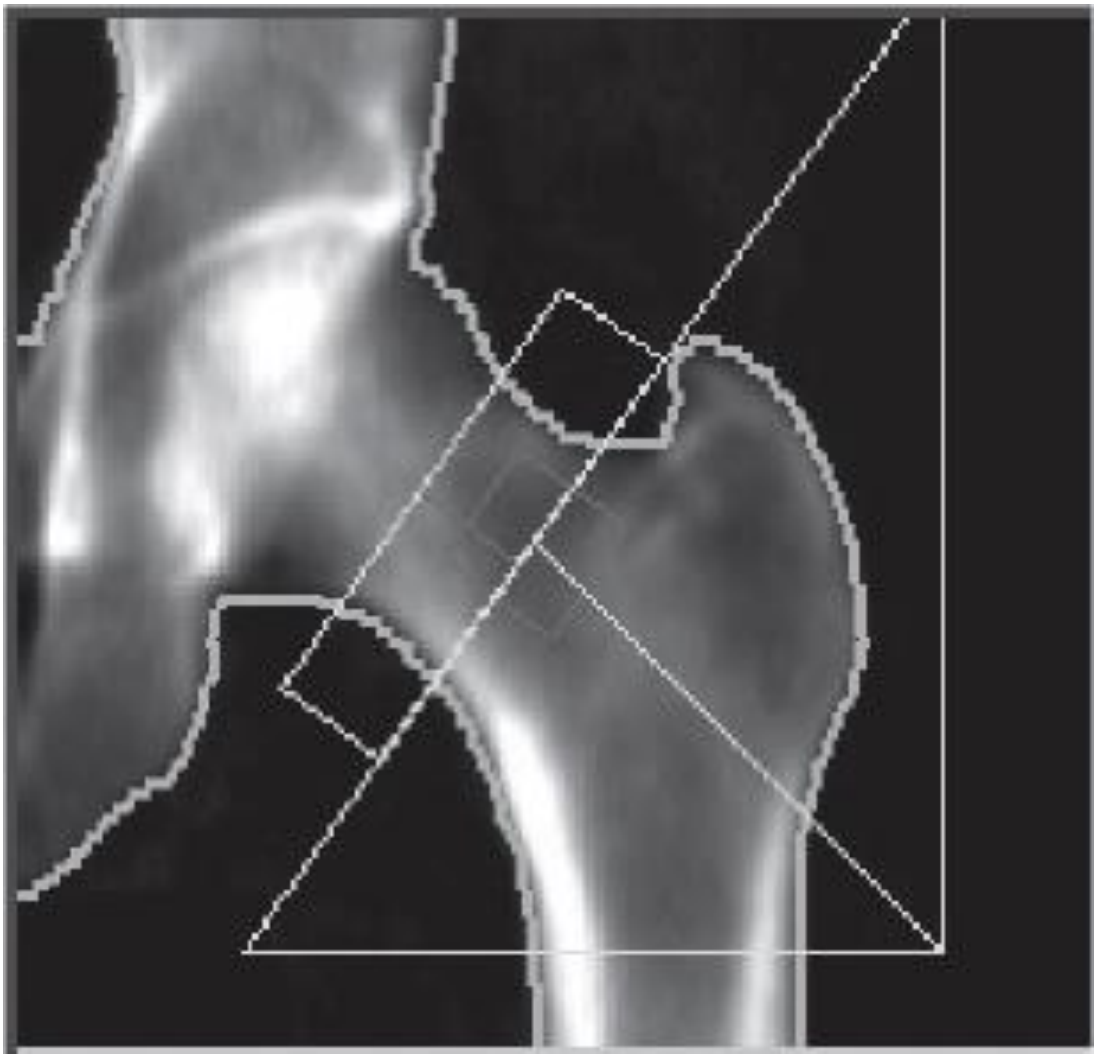


Figure 2.7: An example of DXA image of the hip from Hologic DXA scanner [33].

Previous studies have shown the ability of aBMD to predict fracture risk [1, 67, 72]. Cummings, et al. (1993) assessed the association between aBMD at various sites of hip and hip fractures, and found that the predictive capacity using aBMD was between AUC=0.75 and AUC=0.78. Gregory, et al. (2004) demonstrated a similar AUC score of 0.79 when using femoral neck aBMD, however the AUC scores were 0.63 and 0.95 for using total proximal femur aBMD and Ward's triangle aBMD, respectively. Pulkkinen, et al. (2010) studied the ability of femoral neck aBMD to discriminate cervical fracture, trochanteric fracture and any fracture. ROC analysis revealed that the AUC score was 0.94 for trochanteric hip fracture, 0.74 for cervical fracture and 0.81 for any fracture (Table 2.1).

Table 2.1: The ability of aBMD to predict fracture risk from previous studies. The results are reported as AUC

ROI	Cummings [1]	Gregory [72]	Pulkkinen [67]
Total proximal femur	0.76	0.63	n/a
Femoral neck	0.76	0.79	0.94[a] 0.74[b] 0.81[c]
Intertrochanteric	0.75	n/a	n/a
Trochanter	0.77	n/a	n/a
Ward's triangle	0.78	0.95	n/a

[a] Trochanteric hip fracture; [b] Cervical fracture; [c] Any fracture.

However, some studies have suggested that the ability of aBMD to predict fracture risk is limited [40, 67, 90-92]. More than 50% of fractures occur in people without low aBMD. This occurs, in part, because low aBMD is not the only risk factor for fractures [40, 67]. Some risk factors for osteoporosis, including body weight, exercise, and caffeine intake, appear to act independently of aBMD [93]. In addition, aBMD does not necessarily provide an accurate measure of bone mineral density since it measures the projection of density onto a plane instead of a true volumetric density. aBMD is usually computed over the entire image and averaged over all regions scanned. In particular, only discrete measures are taken of bone density or morphology, potentially ignoring the wealth of information contained within the image. Finally, aBMD and geometric information are analysed separately.

aBMD alone may not exploit the full information available in DXA images regarding fracture risk. The geometry of the femur is thought to provide extra information either on its own or in combination with aBMD [67], and several researchers have explored the influence of other geometric parameters on the risk of fracture of the femur. Pulkkinen et al. (2009) proposed that an aBMD T-score ≤ -2.5 discriminates the risk of trochanteric fractures, whereas geometric risk factors are able to discriminate cervical fracture cases from control cases with similar aBMD [67]. Hip

axial length, femoral length, femoral neck width, and neck shaft angle are the most examined geometric parameters with respect to fracture risk, but findings have been inconsistent [94]. For instance, some studies observed that fracture risk is associated with hip axial length [41, 68], but some did not [95, 96]. Michelotti, et al. (1999) reported that fracture cases had thinner femoral cortices, larger femoral heads, and larger femoral neck diameters than controls [97]. Dinçel, et al. (2008) observed that there was a significant increase in the ratio between femoral neck width and femoral length in the fracture group [94]. A wider femoral neck and shaft, and a larger neck-shaft angle were observed in the fracture group by Gregory et al. (2008) where both male and female fracture subjects were compared with controls [98]. Some of these geometrical parameters, such as hip axis length, neck-shaft angle and femoral neck, are highly correlated [68]. Some studies have observed that femur bone dimensions continue to change with age so it is difficult to fit these geometric parameters to individual DXA images [69, 99].

Discrete geometric measures such as the width, length and angle of the femoral neck are likely to be correlated among themselves and with other characteristics including aBMD. Therefore, coupling between characteristics limits the ability of discrete geometric measures to improve on aBMD for predicting fracture risk. Instead, shape models and appearance models represent the shape variation and appearance variation in terms of de-correlated modes and so allow a clearer picture as to the contribution of individual measurements. The details of the shape models and appearance models are presented in the following sections.

2.4 Active Shape Models and Active Appearance Models

Two methods, active shape models (ASM) and active appearance models (AAM), were developed by Cootes et al [70, 100, 101]. The important difference between ASM and discrete geometric measures such as the width of the femoral neck, the orientation of the femoral neck, etc., is that for discrete geometric measures, one has to guess what might be important aspects of a shape and then test them. Instead, the correct parameters for characterising shape and appearance are determined automatically as part of the ASM or AAM implementation. AAM, as well as ASM, has been shown to

be as accurate as human observers in measuring shape [102], locating hip fractures [103-106], and other medical imaging tasks [70, 107, 108].

In ASM, the shape of the entire object is specified by a large number of points that determine the boundary of the object. Then, a number of principal modes for optimally representing the shape are found automatically using well-established statistical methods [109]. Each principal mode is an independent descriptor of the shape of the object [100]. These distinct descriptors have been used to describe morphometric features to identify subjects at high risk of hip fracture [71]. Gregory, et al. (2004) demonstrated that using ASM to predict the risk of hip fracture is more effective than aBMD and discrete geometric measurements. The study also showed that the combination of Ward's triangle aBMD and ASM can be used to improve the accuracy of training AUC=0.96 compared with training AUC=0.95 using Ward's triangle aBMD alone [72].

While ASM describes the shape of the proximal femur, it does not provide any information regarding variation of grey-level within the image, which, in turn, is directly correlated with the mechanical properties of the bone. Active appearance models (AAM), on the other hand, match a model to the shape and grey-level distribution in the image of the proximal femur [110, 111]. Goodyear, et al. (2013) demonstrated that the features derived from ASM and AAM gave a prediction of fracture comparable to aBMD. The study also showed that the combination of ASM, AAM, and aBMD gave an improvement in the prediction of hip fracture (AUC=0.65) compared with using aBMD alone (AUC=0.62). This improvement could predict an additional 2000 hip fracture cases and potentially save more than GBP 20 million per year in the UK [111]. Waarsing, et al. (2010) analysed the DXA appearance of the proximal femur with respect to osteoarthritis. The study demonstrated that the statistical appearance models captured the total variation in both shape and density of the proximal femur and this was shown to be predictive of osteoarthritis progression [73].

There are two major components in ASM and AAM, the average shape or appearance and the principal modes of shape or appearance that capture variation from the average values. The shape or appearance distribution of bones within a certain population of individuals, either fractured or non-fractured, are described by adding

the linear combination of the principal modes to the average values.

In general, there are two steps needed to implement the ASM or AAM for image interpretation. The first step is to create a parameterized shape model or appearance model based on a set of training images. The second step is to interpret objects in previously unseen images by fitting them with the models established in the first step.

The following two sections provide detailed descriptions of ASM and AAM. In principle, these methods apply to data of any dimension. For instance, a statistical shape and intensity model based on 3D volumetric CT scans that incorporates both shape and grey-scale properties has been developed [112, 113]. However, the formulation here is described for two-dimensional arrays since this study concerns DXA images.

2.4.1 Active Shape Models

2.4.1.1 Labelling the Training Set

The principle idea is to capture the possible shape variation within a population by using a sufficiently large number of examples collected from the population as a training set. The first step is to assign landmark points to the boundary of the shape. To illustrate this step, consider the task of describing the shape of hands. Given a sample set of 16 hand shapes, for instance, it is possible to build a shape model (Figure 2.8a). Each hand in the training set is described by a set of labelled landmark points, which are placed (manually or automatically) at the same relative locations around the boundary. These labels must be consistent from one hand shape to the next. For example, point 38 always corresponds to the tip of the thumb, and point 36 always corresponds to the purlicue (Figure 2.8b). Landmark points across multiple shapes are used to examine and measure shape variation. These landmark points can be divided into three general types in terms of their usefulness [100]. The first type of landmark are those points with particular application dependent significance, such as the tips of fingers. Points 8, 15, 23, 31, and 38 are of the first type. The second type of landmark are those points with application independent significance, such as the curvature of a hand or fingers, and so points 2, 6, 32, 35, and 37 are of this second type. The third type of landmark are points that can be interpolated from points of the first and second

type 1 and 2; for instance, points 3, 4 and 5 are equally spaced along the boundary between points 2 and 6, and so are of the third type.

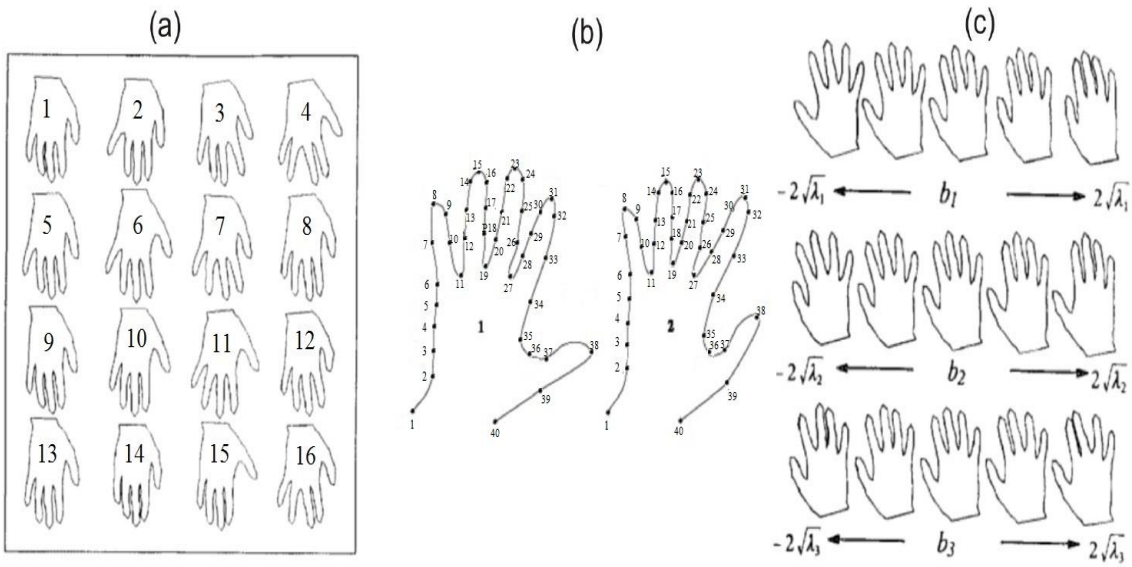


Figure 2.8 : Hand shapes as an example to illustrate a deformable model. (a) The training set of hand shapes (numbered 1, 2, …, 16). (b) Landmarks assigned to two example hands (each hand is labelled with 40 points). (c) Effects of the first three principal modes of shape variation, b_s (s=1, 2, 3) within limits $\pm 2\sqrt{\lambda_s}$, where λ_s is the s largest eigenvalue [100]. By varying the first three parameters of the shape vector, b_s, one at a time, we can demonstrate some of the principal modes of variation allowed by the model. Thus, in this example, the first mode corresponds to the size of the opening between the thumb and the rest of the hand, the second mode corresponds to the overall spread of the fingers and thumb and the third mode corresponds to the orientation of the middle finger.

2.4.1.2 Aligning the Training Set

In the previous example for characterising the shape of the hand, an assumption was made that the size, orientation and positioning of the hand in the images were comparable. In many applications, this assumption does not hold and a method is needed for aligning points prior to applying ASM.

In general, to compare shapes, all landmark points of the examples in the training set are aligned into a common coordinate frame using scaling, rotation, and translation

to eliminate non-shape variations between examples. Procrustes analysis is a means used in ASM to remove scaling, rotation, and translation between two shapes [114]. Thus, the coordinates of the aligned points for a single example boundary are recorded relative to a common coordinate frame rather than absolutely. Procrustes analysis has many variations and forms. Generalized orthogonal Procrustes analysis (GPA) [114] is a common method used to align image shape and a modification of GPA was used for alignment in this study.

First we consider the problem of aligning an object to a standard or reference object. The shape of the reference object is determined by n landmark points of the form $(x'_{0,k}, y'_{0,k})$, with $k = 1, 2, \ldots, n$. The shape of an object i to be aligned to the reference object is determined by an analogous set of n points $(x'_{i,k}, y'_{i,k})$. The reference object is represented by the vector of length $2n$ given by

$$X'_0 = (x'_{0,1}, y'_{0,1}, x'_{0,2}, y'_{0,2}, \ldots, x'_{0,n}, y'_{0,n})^T, \tag{2.1}$$

where T denotes the transpose. Similarly, the object to be aligned is represented by the vector

$$X'_i = (x'_{i,1}, y'_{i,1}, x'_{i,2}, y'_{i,2}, \ldots, x'_{i,n}, y'_{i,n})^T. \tag{2.2}$$

The objective is to translate, rotate and scale the object to be aligned as well as possible with the reference object. For a single point $P' = (p', q')$ in the plane, the point $P'' = (p'', q'')$ is obtained by a horizontal shift of x_t, a vertical shift of y_t, scaling by s and counterclockwise rotation by θ. This is described by

$$\begin{pmatrix} p'' \\ q'' \end{pmatrix} = T_{x_t, y_t, s, \theta} \begin{pmatrix} p' \\ q' \end{pmatrix}, \tag{2.3}$$

where $T_{x_t, y_t, s, \theta}$ is the operator given by

$$T_{x_t, y_t, s, \theta} \begin{pmatrix} p' \\ q' \end{pmatrix} = s \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} p' - x_t \\ q' - y_t \end{pmatrix}. \tag{2.4}$$

Applying T to the vector X'_i results in the transformed vector

$$X''_i = (x''_{i,1}, y''_{i,1}, x''_{i,2}, y''_{i,2}, \ldots, x''_{i,n}, y''_{i,n})^T, \tag{2.5}$$

Where

$$\begin{pmatrix} x''_{i,k} \\ y''_{i,k} \end{pmatrix} = T_{x_t,y_t,s,\theta} \begin{pmatrix} x'_{i,k} \\ y'_{i,k} \end{pmatrix}. \tag{2.6}$$

A natural criterion for determining the best shifts, scale and rotation is to minimize the square of the Euclidean distance between the reference shape vector X'_0 and the transformation of the object shape vector X'_i given by

$$\begin{aligned} E &= \left\| X'_0 - T_{x_t,y_t,s,\theta}(X'_i) \right\|^2 \\ &= (X'_0 - T_{x_t,y_t,s,\theta}(X'_i))^T (X'_0 - T_{x_t,y_t,s,\theta}(X'_i)) \\ &= \sum_{k=1}^{n} (x'_{0,k} - x''_{i,k})^2 + (y'_{0,k} - y''_{i,k})^2. \end{aligned} \tag{2.7}$$

over the parameters x_t, y_t, s and θ. However, in many applications, not all landmark points are equally reliable. Accordingly, a more practical criterion for determining the best parameters is given by

$$E_W = (X'_0 - T_{x_t,y_t,s,\theta}(X'_i))^T W (X'_0 - T_{x_t,y_t,s,\theta}(X'_i)), \tag{2.8}$$

where the $n \times n$ diagonal matrix W provides a weighting for each landmark point. In principle, W could represent prior information regarding the process of assigning landmark points or information about the data set in question. A general and automatic method for determining the matrix W based on the data itself is as follows [100].

Consider a set of m objects (say, in a training set) with shape vectors X'_i, i = 1, 2, ..., m. Let R_{ikl} be the difference between the landmark point k and landmark point l for an object i with shape vector X'_i. Thus,

$$R_{ikl} = \sqrt[2]{\left(x'_{i,k} - x'_{i,l}\right)^2 + \left(y'_{i,k} - y'_{i,l}\right)^2}. \tag{2.9}$$

The average distance between the landmark point k and landmark point l over all the shapes in the set of m objects is

$$\bar{R}_{ikl} = \frac{1}{m}\sum_{i=1}^{m} R_{ikl}, \tag{2.10}$$

and the variance is

$$V_{kl} = \frac{1}{m}\sum_{i=1}^{m}(R_{ikl} - \bar{R}_{ikl})^2. \tag{2.11}$$

For a particular landmark point *k*, if $\sum_{l=1}^{m} V_{kl}$ is small, then the landmark point is consistent and may be viewed as reliable. If $\sum_{l=1}^{m} V_{kl}$ is large, the landmark point may be viewed as less reliable. Hence, a reasonable definition for the k^{th} diagonal element of the matrix *W* is

$$w_k = (\sum_{l=1}^{m} V_{kl})^{-1}. \tag{2.12}$$

Once the matrix *W* is determined, finding the best parameters for x_t, y_t, *s* and θ is a matter of minimizing the value of E_W in Equation 2.8. Taking partial derivatives of E_W with respect to x_t, y_t, *s* and θ and setting these to zero yields four equations in four unknowns that may be solved by standard methods [100].

We then use the following algorithm to align the shape of all training examples as follows:

1. Select the shape of the first training example object to be the initial mean shape.
2. Align the shapes of the remaining training examples to the mean shape.
3. Repeat until the process converges
    - Calculate the mean shape from the aligned shapes.
    - Normalize the orientation, scale, and origin of the current mean shape.
    - Realign every shape with current mean shape

Once the alignment process is complete, the aligned version of the shape vector X_i' is denoted by X_i.

2.4.1.3 Constructing a Point Distribution Model

From these collections of aligned landmark points, a point distribution model (PDM) [115] is constructed to model the variation of the points within equivalent landmarks. The procedure is as follows. The coordinates of the aligned landmark points for the

object in image i are denoted by $(x_{i,1}, y_{i,1})$, $(x_{i,2}, y_{i,2})$, ..., $(x_{i,n}, y_{i,n})$. The aligned landmark points are stacked in vectors, and the shape of the object in image i is represented by a $2n$ element vector $X_i = (x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2}, ..., x_{i,n}, y_{i,n})^T$, where n is the number of labelled landmark points**.**

For the collection of shape vectors from the training set, $\{X_i, i = 1, 2, ..., N\}$, the mean vector is calculated as:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{2.13}$$

and the covariance matrix, S, is the $2n \times 2n$ matrix

$$S = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})\,(X_i - \bar{X})^T. \tag{2.14}$$

Principal component analysis (PCA) consists of finding the eigenvalues and associated normalised eigenvectors, Φ_s (s=1, ..., $2n$), of s, listed in order of decreasing eigenvalue. The eigenvalues, λ_s (s=1,..., $2n$), indicate the proportion of the information content encoded in the corresponding eigenvector [116]. When there are fewer examples in the training set, N, than point coordinates, $2n$, the eigenvectors $\widehat{\Phi}_s$ (s=1, ..., N) and corresponding eigenvalues λ_s (s=1,..., N) of the $2n \times 2n$ covariance matrix, S, can be calculated from the eigenvectors of a smaller $N \times N$ matrix derived from the same data [100]. PCA is a general method but when applied in this context, the eigenvectors are called principal modes. The principal modes are mutually orthogonal ($\Phi_s\Phi_t^T = 0$ if $s \neq t$) and so PCA separates the information content of the vectors X_i into uncorrelated principal modes Φ_s.

We then apply PCA to the data. PCA is a statistical procedure concerned with elucidating the covariance structure of variables in a dataset. The reason for using PCA is that it allows us to convert a set of possibly correlated variables into a set of linearly uncorrelated variables or modes, and, thus, it is used to de-correlate the shape information in preparation for feature selection. A second benefit is that each of the de-correlated modes comes with a measure based on the associated eigenvalue of the information content represented by the mode. By selecting the modes associated with the most information, it is possible to remove insignificant variations in shape due to, for example, inconstancies in the placement of boundary points and redundant shape

information [117].

Each principal mode describes a portion of overall shape variations and, collectively, all principal modes account for 100% of overall shape variations within the original training data set. Each principal mode is ordered according to the amount of variation explained. The lower numbered principal modes explain a larger percentage of variations. Therefore, the object of interest in the set of training images within the population can now be represented by a vector, X_i, as the mean plus a perturbation specific to the individual by

$$X_i = \bar{X} + \sum_{s=1}^{2n} b_s \Phi_s. \quad (2.15a)$$

Here $2n$ is the total number of principal modes. The coefficients b_s are computed by

$$b_s = (X_i - \bar{X})\Phi_s^T. \quad (2.15b)$$

For a shape vector X from an example object outside the training set

$$X \approx \bar{X} + \sum_{s=1}^{2n} b_s \Phi_s. \quad (2.15c)$$

Only principal modes corresponding to the top t eigenvalues are chosen so as to explain a certain proportion of the variances in the training shapes, usually ranging from 90% to 99.5%. Accordingly, the final vector representing the example shape is denoted by $\hat{X}$ and is defined as

$$\hat{X} = \bar{X} + \sum_{s=1}^{t} b_s \Phi_s. \quad (2.15d)$$

Here t is the number of principal modes used. The number of principal modes depends on the desired balance between detail and redundancy in the description of the example shape.

This shape model can be used to fit new examples of shapes by varying the parameters b_s within the range of $\pm m\sqrt{\lambda_s}$, where m usually has a value between two and three. For instance, the effects of varying the first three parameters of the hand shape vector, b_s (s=1, 2, 3), within limits $\pm 2\sqrt{\lambda_s}$ are demonstrated in Figure 2.8c.

2.4.1.4 Using Point Distribution Models to Search Objects

The point distribution model may be used to automatically determine the shape parameters for an object in an image that was not part of the training set used to create the model. This involves finding shape parameters b_s and a set of pose parameters (x_t, y_t, s, θ) that cause the model to coincide with landmarks of objects in the image. The pose parameters (x_t, y_t, s, θ) comprise the Euclidean transformation $T_{(x_t,y_t,s,\theta)}$ as presented in Equation 2.4.

An iterative search approach is used in ASM to fit the model, $\hat{X}$, to the target object in an image. The process is as follows [100]:

0. Give a rough starting approximation of the positions of a set of model points that represent the boundaries of the object in an image, and initialize the shape parameters b_s to zero.
1. Generate the shape model according to Equation 2.15d.
2. Generate X_{image} in the image coordinate frame from $\hat{X}$ by the transformation $T_{(x_t,y_t,s,\theta)}$ (Equation 2.4).
3. Examine the region along a normal to each model point of X_{image} toward each of the nearest edge points on the object in the image to search for the best object shape Y_{image}, and calculate the adjustments, dX_i, in the image coordinate frame required to make $Y_{\text{image}} = X_{\text{image}} + dX_i$.
4. Project Y_{image} to the model coordinate frame by inverting the transformation T, that is, $\hat{Y} = T^{-1}{}_{(x_t,y_t,s,\theta)} Y_{\text{image}}$, to update the pose parameters (x_t, y_t, s, θ)
5. Update shape parameter $b_s' = \Phi^T(\hat{Y} - \bar{X})$
6. Apply a constraint to the shape parameter b' (e.g. constrain to the range of $\pm m\sqrt{\lambda_s}$) to ensure a plausible shape.
7. If dX_i is greater than a pre-set threshold value, assign $b_s = b_s'$ and return to step 1. Else, $b_s = b_s'$ are the final shape parameters.

This final set of shape parameters b_s represents the shape of the target object.

2.4.2 Active Appearance Model

While the active shape model represents the example dataset in terms of the shape variation from a mean shape, it does not contain any information regarding the grey-scale value distribution of the image over the region corresponding to the example. Two femurs with identical profiles in DXA images may have very different risks of fracture due to the amount and distribution of bone tissue within the boundary of the femur. Such information manifests in the grey-scale values of the pixels representing the femur. An active appearance model is used to simultaneously explain shape and grey-scale variation. Such a model is generated by combining a model of shape variation as well as a model of grey-scale variation in a shape-free frame, which is referred to as the appearance.

To build a model of grey-scale variation, it is necessary to warp each training image so that its boundary points match the corresponding points of the mean shape, obtaining a shape-free patch (using Delaunay triangulation algorithm [118]). Within the region covered by the mean shape, the intensities are sampled. Because shape-free patches are used, the locations of the pixels in one image closely match the locations of the pixels in the other images. Thus, the intensity values at these pixels record spatial intensity patterns that may be compared between images.

For image i, the numbers $g_{i,1}$, $g_{i,2}$, ..., $g_{i,m}$ represent the grey-scale values at pixels 1, 2, ..., m within the region of the shape-free image covered by the mean shape. The grey-scale vector associated with image i is G_i given by

$$G_i = \left(g_{i,1}, g_{i,2}, \dots, g_{i,m}\right)^T, \tag{2.16a}$$

and the average grey-scale vector for the training set of N images is

$$\bar{G} = \frac{1}{N}\sum_{i=1}^{N} G_i. \tag{2.16b}$$

The same steps used for ASM are used to construct a grey-scale model for the objects. The grey-scale model is combined with the shape to create a single model. Since there may be correlation between the shape and grey-scale models, PCA is

applied to the combined shape and grey-scale representations to generate the appearance model. Thus, for any appearance vector (shape and grey-scale) from the training or testing set, the appearance of the objects are represented by $\tilde{X}$ and $\tilde{G}$, and are defined as

$$\tilde{X} = \bar{X} + \sum_{a=1}^{z} c_a \, \Phi_a, \tag{2.17a}$$

$$\tilde{G} = \bar{G} + \sum_{a=1}^{z} c_a \, \Psi_a, \tag{2.17b}$$

Where Φ_a and Ψ_a are the principal modes of shape and grey-scale variation, respectively. Here z is the number of principal modes used. The parameters c_a control the shape $\tilde{X}$ as well as grey-scale $\tilde{G}$ and, thus, the set c_a, a = 1, 2, …, z is used as the set of appearance features for each object.

The AAM is used to automatically determine the appearance parameters for an object in an image that was not part of the training set used to create the model. The search for these appearance parameters could be treated as an optimisation problem in which the difference between a new image and one generated by the AAM is minimised by varying the appearance model parameters c_a. Similar to ASM, an iterative search approach is used in AAM as follows [101]:

0. Given a current estimate of appearance model parameters, c_a, and the sample of the grey level information G_s from the shape-free image at this current estimate.
1. Evaluate the error vector $\delta_{g0} = G_s - G_m$ and the current error $E_0 = \left|\delta_{g0}\right|^2$, where G_m is the vector of grey-scale values for the current model parameters and generated using Equation 2.17b.
2. Compute the predicted displacement $\delta_c = A\delta_{g0}$. Where A is a scaling parameter for the grey levels learned by applying multiple multivariate linear regressions on the training data.
3. Update the appearance parameters $c_a = c_a - k\,\delta_c$, where initially k = 1.
4. Sample the image at this new prediction, and calculate a new error vector, δ_{g1}.
5. If $\left|\delta_{g_1}\right|^2 < E_0$ then accept the new estimate, c_a, otherwise decrease the value of k and go to step 4.

2.4.3 Comparison between ASM and AAM

In this thesis, we use two deformable models, the ASM and the AAM to represent the variations in shape and/or appearance of the target objects. The former searches around the current locations of each model point along profiles and updates the current estimate of the shape of the target objects while the latter seeks to match both the locations of each point and the grey-levels of the object to the image. Thus, there are three main differences between ASM and AAM. Firstly, the result search for ASM may be less reliable than AAM as it uses only the data in a small region around each landmark point, whereas the AAM takes advantage of the appearance information of the whole region within objects. Secondly, the ASM tends to need a relatively larger number of landmark points around the boundary than the AAM so as to provide sufficient direction information for the search. It is a considerable task to label a great quantity of images. Thirdly, the AAM is more robust as it gives a better match to the object, whereas the ASM is faster and performs more accurately in shape localization. However, a quicker AAM algorithm could be achieved if the search area is placed only near significant boundaries or corners as this will require less image sampling during the search. [119].

2.5 Texture Analysis

Image texture, the information of the spatial variation in pixel intensities in an image or selected region of an image, has been recognized as an important attribute for quantifying the perceived appearance of an image [120]. In this study, Gabor filters and textons were applied to extract the texture features of proximal femur images and are discussed in the following sections.

2.5.1 Texture Features Based on Gabor Filters

Texture analysis based on Gabor filters is thought to be similar to perception in the human visual system. The Gabor function can be implemented as a multichannel filter that mimics characteristics of the human visual system [121, 122]. Gabor filters have been successfully employed in computer vision and a variety of image analysis applications such as classification and segmentation [123, 124], image coding and compression [125], face recognition [126], and motion analysis [127]. Gabor filters

can serve as local bandpass filters with excellent joint localization properties in both spatial and frequency domains [122, 128] and, thus, the texture features based on Gabor filters are robust to variances in rotation, scale and illumination in images.

Several texture analysis methods based on Gabor filters have been developed to detect fractures and signs of osteoporosis. Yap, et al. (2004) proposed an adaptive sampling method that begins with the sampled locations in different images corresponding to consistent locations. Then, a set of Gabor filters are applied to each sampled region to extract texture features. Results from this method indicated improved overall performance of fracture detection, especially when combined with neck-shaft angle measurement [129]. Lim, et al. (2004) included neck-shaft angle measurements, Gabor filters, Markov Random Field texture, and intensity gradient into their analysis. Bayesian and support vector machine (SVM) classifiers were used to classify the test samples. The results suggested that the classification accuracy, the number of correctly classified samples over the total number of samples, was improved significantly to 98.2% by combining these methods, compared with 93.5% when using neck-shaft angle alone [130]. Similarly, Pramudito, et al. (2007) used three different texture analysis methods including Gabor filters, wavelet transforms and fractal dimension to extract features that represent the structural change in trabecular pattern, and compared these methods with the corresponding Singh index grading system [131]. Their results revealed that the features extracted using Gabor filters in the form of energy are significantly correlated with the Singh indexes determined by a physician and, thus, contribute to improving early osteoporosis detection.

The Gabor feature space consists of responses produced by convolving the original image with a bank of 2D, real Gabor kernel functions at several different scales (spatial frequencies) and orientations [128, 132]. The Gabor kernel function $g_{\lambda,\theta}(x, y: \varphi, \sigma, \gamma)$ is the product of a Gaussian and a cosine function.

$$g_{\lambda,\theta}(x, y: \varphi, \sigma, \gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right) \tag{2.18}$$

$$x' = \quad x\cos\theta + y\sin\theta$$

$$y' = -x\sin\theta + y\cos\theta,$$

where $\gamma = 0.5$ is a constant, called the spatial aspect ratio that determines the ellipticity of the receptive field, λ represents the scale (so $1/\lambda$ is the spatial frequency), $\theta \in [0, \pi)$ represents the orientation, and σ is the standard deviation and determines the size of the receptive field. $\varphi \in (-\pi, \pi]$, is the phase offset that determines the symmetry of $g_{\lambda,\theta}$ with respect to the origin; the phase offset values $\varphi = 0$ and $\varphi = \pi$ correspond to symmetric functions (also called even), while $\varphi = -\pi/2$ and $\varphi = \pi/2$ correspond to anti-symmetric functions (or also called odd). The value of the standard deviation σ cannot be specified directly and it can only be set through the half response spatial frequency bandwidth, b. The value of b is related to the ratio σ/λ.

$$b = \log_2 \frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{\ln 2}{2}}}, \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\,\frac{2^b+1}{2^b-1}. \tag{2.19}$$

Let $I(x, y)$ denote an image and $g_{\lambda,\theta}(x, y: \varphi, \sigma, \gamma)$ denote the Gabor kernel function with scale value λ and orientation value θ. The Gabor filter response, at this scale and orientation, to this image is expressed as

$$G_{\lambda,\theta}(x, y) = I(x, y) * g_{\lambda,\theta}(x, y: \varphi, \sigma, \gamma) \tag{2.20}$$

The output of the Gabor filter $G_{\lambda,\theta}(x, y)$ is a number that indicates how much the intensity pattern at position (x, y) in the image resembles a line segment at orientation θ and of width and length characterised by the other parameters. By recording these filter outputs at locations (x, y) within the femur (or a subregion of interest), a summary of oriented structure is obtained. This oriented structure is a measure of texture. This terminology is standard in image analysis but it is important to note that this is an image texture resulting from patterns in tissue structure and is not a physical texture of the bone.

Thus, Gabor filters may be tuned to extract particular characteristics arising in the images by appropriately selecting each of the Gabor function parameters. There are no general methods for the selection of Gabor filter parameters, which is often a vague and application dependent task. A straightforward approach to select the parameters would be to uniformly sample the orientations and spatial frequencies while other parameters are given a fixed value by referring to literature [128, 132-134]. Some examples of Gabor filters functions are illustrated in Figure 2.9. Gabor filters exhibit

strong characteristics of spatial frequency and orientation selectivity.

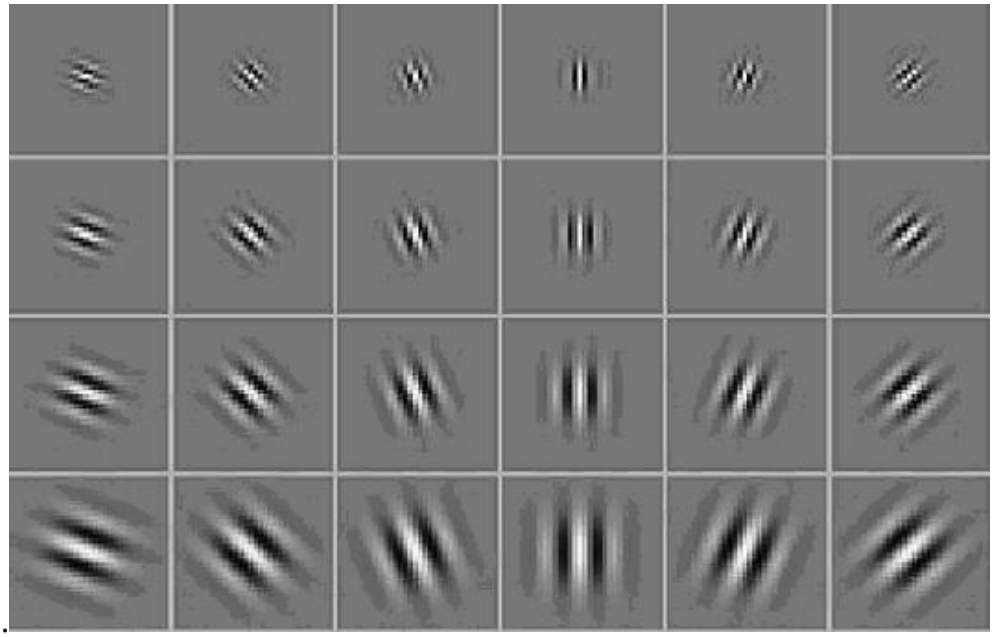


Figure 2.9 : The real part of 24 Gabor filters with four different spatial frequencies (1/16, 1/18, 1/26 and 1/32 from top to bottom) and six different orientations.

Gabor feature vectors can be used directly as input for classification or they can be transformed into new feature vectors to be used as such input. In this thesis, for example, Gabor energy features were computed. More specifically, the response outputs of a symmetric (φ = 0) and an anti-symmetric (φ = -π/2) filter with the same scale λ and orientation θ at each image pixel within the ROI were combined to find the mean value of all the pixels within the ROI. Thus, for a ROI of size M x N pixels, the Gabor energy for a particular choice of λ and θ is given by

$$E_{\lambda,\theta}(G) = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left\lceil G_{\lambda,\theta}(x,y)\right\rceil^2, \qquad (2.21)$$

where $\left\lceil G_{\lambda,\theta}(x,y)\right\rceil$ is the magnitude of the Gabor filtered image $G_{\lambda,\theta}(x,y)$.

By using Gabor energy, the dimensionality of the multichannel Gabor feature space is greatly reduced. It has been shown that the Gabor energy is closely related to the local power spectrum of the image [128].

2.5.2 Texture Features Based on Textons

Julesz proposed the term "textons" in the context of cognitive science and stated that "textons are the putative units of pre-attentive human texture perception" [135]. In image analysis applications, local texture descriptors are mapped to a feature space, the dimension of which is the number of local texture descriptors used. Clusters in this feature space are called textons and represent commonly occurring local texture patterns. Each pixel in the image is then mapped to the texton closest to it in the feature space representation of the pixel. The image is represented by the histogram of texton occurrences [136]. Texton-based approaches are simple to implement and often achieve good performance for texture image categorization and segmentation [137, 138]. For medical imaging applications, Petroudi et al. (2003) used texton-based features generated from a filter bank to classify mammograms into the four breast imaging reporting and data system (BI-RADS) classes [139]. More recently, T. Jiang et al. (2010) proposed a texton-based classification system based on raw pixel representation along with an SVM with radial basis function kernels for the classification of emphysema in computed tomography images of the lung. Classification accuracy achieved 96.43% [140].

The framework of texton generation for feature extraction and classification consists of five steps (Figure 2.10): (1) extracting local feature vectors from the images collected, (2) aggregating all local feature vectors to construct a filter response space and clustering into textons, (3) creating texton maps, (4) generating histograms of textons for each image (5) classifying the images based on the texton histograms.

In the first step, the local feature vectors are usually filter responses generated by convolving the image with a filter bank or pixel values from N x N neighbourhoods. The value $N = 3$ is well established in the literature as suitable [141-144]. For example, Varma and Zisserman classified over 2800 images of all 61 textures present in the Columbia-Utrecht database (a database of real world surface textures) [145] using local N x N neighbourhoods based textons with values of $N = 3, 5, 7, \ldots, 19$. It was demonstrated that the classification performance using $N = 7$ was optimal but only slightly better than using $N = 3$, and at the expense of a much higher computational cost [141]. Li et al. used $N = 3, 5, 7$ to estimate breast cancer risk and the best classification performance was achieved with $N = 3$ [144]. From a theoretical point of

view, a 3x3 neighbourhood contains sufficient information to assign the best local quadratic approximation of the image at the central point if the image is taken to be a discretization of a differentiable surface. The reason is that the nine points in the 3x3 neighbourhood suffice to calculate the first and second partial derivatives at the central point and these suffice to classify all quadratic surfaces.

The local feature vectors of these images are aggregated in the second step, and a number of cluster centres are learnt by applying a clustering method, such as *K*-means clustering, on these aggregated feature vectors. These cluster centres are then collected into a single dictionary, called the texton dictionary. Having learnt a texton dictionary, the next step is to construct the corresponding texton map for each image by labelling each pixel with the texton that lies closest to it in the filter response space. The histogram of each texton map, i.e., the frequency with which each texton occurs in the labelling generated in step four, represents the texture features for the corresponding image. Hence, the histogram of texton occurrences for an image is the feature vector that is used to classify the image.

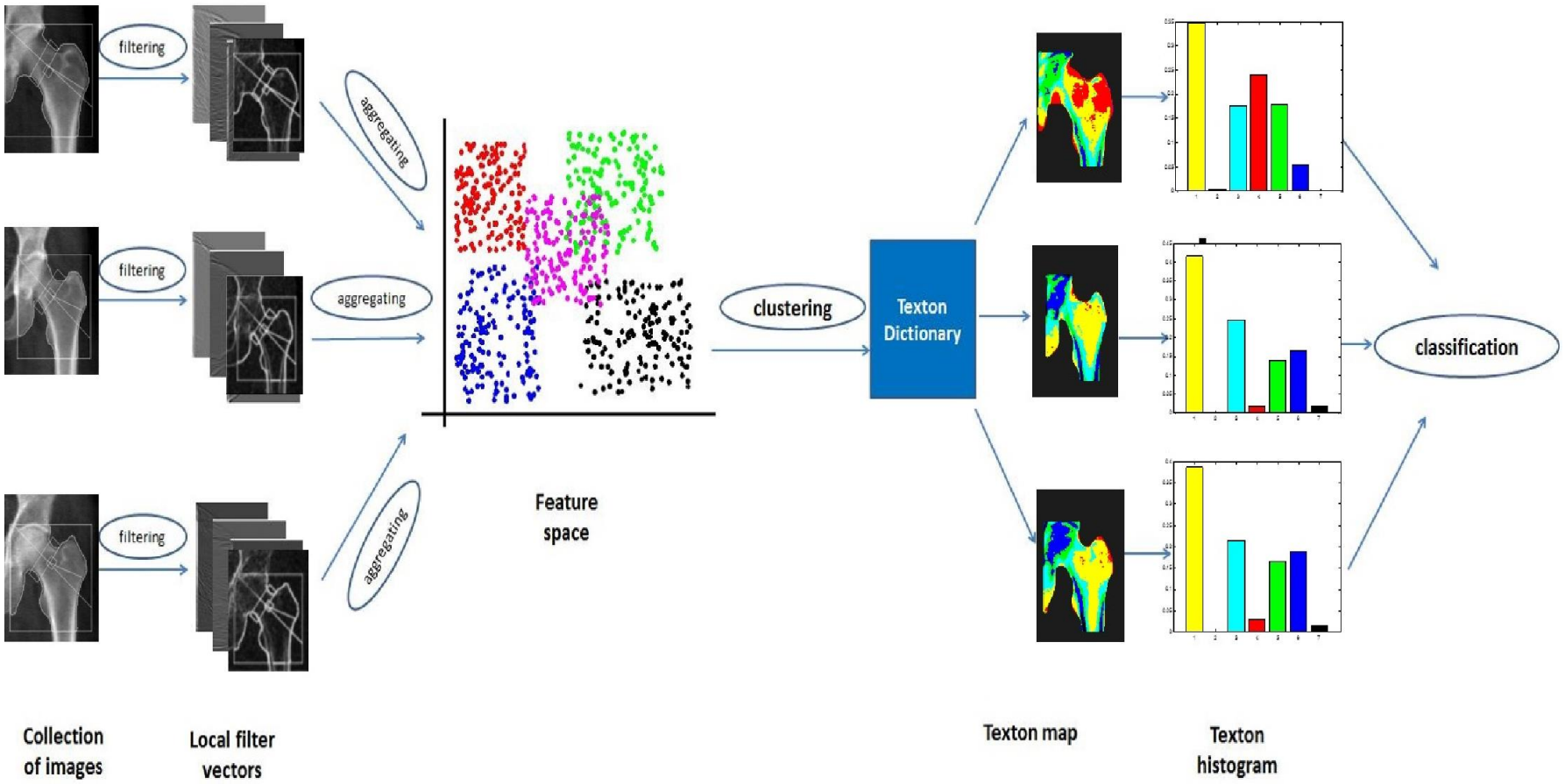


Figure 2.10 : Framework for classification based on textons described in five steps: (1) Extracting local feature vectors from the collected images. (2) Aggregating all local feature vectors to construct a filter response space and clustering into textons. (3) Creating texton maps. (4) Generating histograms of textons for each image. (5) Classifying the images based on the texton histograms.

2.6 Classification and Linear Discriminant Analysis

Classification is the procedure of assigning an individual, or a subset, of the data to one of several known classes (or groups) based on observations. Usually, the observations take the form of a list of numbers arranged in a vector that is viewed as a point in a feature space. A classification scheme determines, directly or indirectly, a set of hypersurfaces in the feature space that separates the points representing the various groups. If the hypersurfaces are hyperplanes, then the classification process is called linear discriminant analysis (LDA). Non-linear classifiers often appear to produce better results than LDA but this can be deceiving especially if the dimension of the feature space is large and the number of examples used to train the classifier is small. Since the total number of features considered in this work is around 133 and the number of examples used to train classifiers is around 60, only LDA classifiers will be implemented in this thesis.

Fisher's linear discriminant analysis (FLDA) or, in short, LDA, is one of the most commonly used techniques. The reason for using Fisher is that if the feature vectors from each group are normally distributed, then Fisher's classifier is the optimal classifier among all (linear and nonlinear) classifiers. In addition, experience shows that even if the distributions are not quite normal but somewhat close to normal, then Fisher still gives close to optimal results. It was originally developed by R. A. Fisher in 1936 and has been used successfully in many pattern recognition problems [146]. Other examples of the use of LDA include earth science [147], biomedical studies [148], and marketing research applications [149]. LDA is often preferable to non-linear discriminant analysis which suffers from the problem of over fitting if the number of training examples is small.

2.6.1 Optimal Classification

Optimal classification is aimed at minimizing the misclassification risk and, thus, minimisation of total probability of misclassification is frequently used as a criterion for optimal classification [150]. Accordingly, the total probability of misclassification, for a good classifier, should be kept as small as possible even though it may produce a few misclassifications. In some applications, misclassification cost, that is, the cost incurred when an object of class I is incorrectly classified as class J ($I \neq J$), is considered to reflect the seriousness of the errors of classification. For instance, the consequence

of classifying a cancer case as non-cancer is usually more severe than classifying a non-cancer case as cancer. Therefore, the misclassification cost for the former is higher than the latter. In this situation, minimisation of the total misclassification cost is a more reasonable criterion for the optimal classification. The Bayesian approach is another appropriate criterion for performing optimal classification. In this approach, an object is assigned to the class with maximal posterior probability. These three criteria are described in detail in the following sections.

2.6.1.1 Minimising the Total Probability of Misclassification

Let p_1 and $f_1(x)$ be the prior probability and the normal density function of group 1, and let p_2 and $f_2(x)$ be the prior probability and the density function of group 2. Assume $f_1(x)$ and $f_2(x)$ intersect once at x_1 as illustrated in Figure 2.11. The total probability of misclassification $M(x_i)$ is the sum of the probability of assigning an object x from group 2 to group 1 (region a) and the probability of assigning an object x from group 1 to group 2 (region b and c), computed by

$$M(x_i) = \int_{-\infty}^{x_i} f_2(x)p_2 dx + \int_{x_i}^{\infty} f_1(x)p_1 dx. \tag{2.22}$$

Since

$$\int_{-\infty}^{\infty} f_1(x)dx = \int_{-\infty}^{x_i} f_1(x)dx + \int_{x_i}^{\infty} f_1(x)dx = 1, \tag{2.23}$$

Equation 2.22 can be rewritten as

$$\begin{aligned} M(x_i) &= \int_{-\infty}^{x_i} f_2(x)p_2 dx + p_1\left(1 - \int_{-\infty}^{x_i} f_1(x)dx\right) \\ &= p_1 + \int_{-\infty}^{x_i}(f_2(x)p_2 - f_1(x)p_1)dx. \end{aligned} \tag{2.24}$$

Differenting the Equation above gives

$$M'(x_i) = f_2(x)p_2 - f_1(x)p_1. \tag{2.25}$$

And since $M(x_i)$ is the minimum when $M'(x_i) = 0$, that is,

$$f_2(x)p_2 = f_1(x)p_1, \tag{2.26}$$

we obtain the classification criteria,

$$\begin{cases} Group1\ ,\ if\ \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \\ Group2,\ if\ \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}. \end{cases} \tag{2.27}$$

If the prior probability of the two groups are equal, $p_1 = p_2$, the classification criteria can be rewritten as

$$\begin{cases} Group1\ ,\ if\ \frac{f_1(x)}{f_2(x)} \geq 1 \\ Group2,\ if\ \frac{f_1(x)}{f_2(x)} < 1. \end{cases} \tag{2.28}$$

It can be seen in Figure 2.11 that the total probability of misclassification is the minimum (region *a*, *b* and *c*) when $x = x_1$. The total probability of misclassification for any other value of x is relatively higher. For example the total probability of misclassification is increased by an amount equal to region d if x shifts from x_1 to x_2.

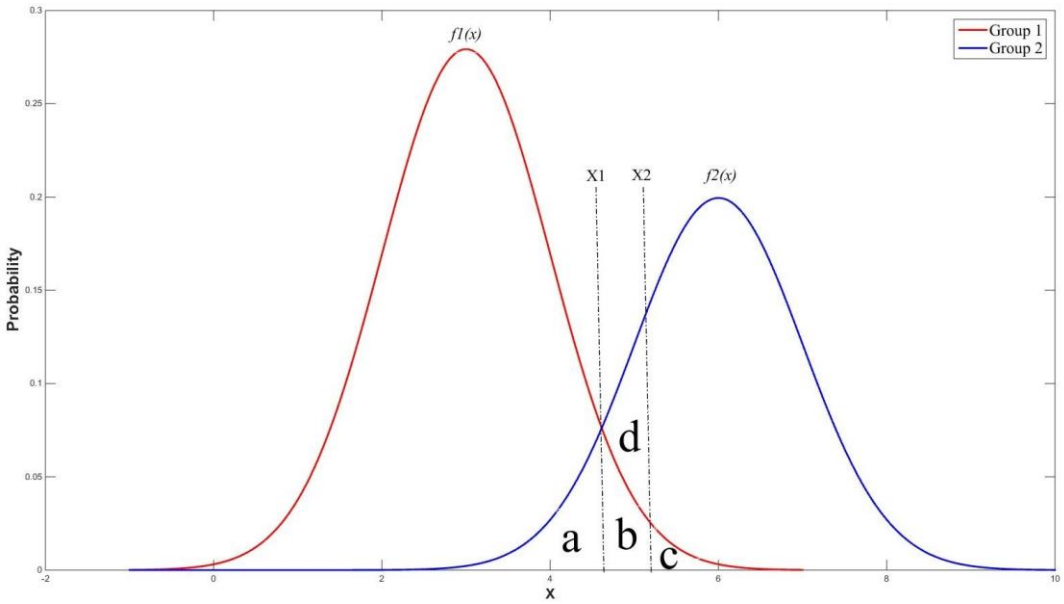


Figure 2.11: Classification rule based on minimising the total probability of misclassification.

2.6.1.2 Minimising the Total Cost of Misclassification

Let C_1 be the cost incurred when assigning an object x from group 2 to group 1, and C_2 be the cost incurred when assigning an object x from group 1 to group 2. The total cost of misclassification C is then the sum of the cost of wrongly assigning an object x to group 1 and the cost of wrongly assigning an object x to group 2, which is defined as

$$C(x_1) = C_1 \int_{-\infty}^{x_1} f_2(x) p_2 dx + C_2 \int_{x_1}^{\infty} f_1(x) p_1 dx. \tag{2.29}$$

Differenting the Equation above gives

$$C'(x_1) = C_1 f_2(x_1) p_2 - C_2 f_1(x_1) p_1. \tag{2.30}$$

$C(x_1)$ is the minimum when $C'(x_i) = 0$, that is,

$$C_1 f_2(x_1) p_2 = C_2 f_1(x_1) p_1. \tag{2.31}$$

Therefore, we obtain the following classification criteria, which is equivalent to the Bayes decision rule.

$$\begin{cases} Group1\,,\ if\ \frac{f_1(x)}{f_2(x)} \geq \frac{p_2 C_1}{p_1 C_2} \\ Group2,\ if\ \frac{f_1(x)}{f_2(x)} < \frac{p_2 C_1}{p_1 C_2}. \end{cases} \tag{2.32}$$

If the two cost weightings C_1 and C_2 are equal, then the classification criteria above can be rewritten to be the same as Equation 2.27.

2.6.1.3 Maximising the Posterior Probability

According to Bayes' theorem, an object x is assigned to group i with the maximum posterior probability $P(i|x)$. That is, x is assigned to

$$\begin{cases} Group1\,,\ if\ \ P(1|x) \geq P(2|x) \\ Group2,\ if\ P(1|x) < P(2|x). \end{cases} \tag{2.33}$$

Also, Bayes' theorem defines the posterior probability of an object *x* belonging to group *i* as

$$P(i|x) = \frac{p_i f_i(x)}{\sum p_i f_i(x)}, \tag{2.34}$$

where p_i is the prior probability of group *i.*

Thus, Equation 2.33 can be rewritten as

$$\begin{cases} Group1\ ,\ if\ \ \frac{p_1 f_1(x)}{p_1 f_1(x)+p_2 f_2(x)} \geq \frac{p_2 f_2(x)}{p_1 f_1(x)+p_2 f_2(x)} \\ Group2,\ if\ \frac{p_1 f_1(x)}{p_1 f_1(x)+p_2 f_2(x)} < \frac{p_2 f_2(x)}{p_1 f_1(x)+p_2 f_2(x)}. \end{cases} \tag{2.35}$$

Or, simplified and rearranged, as

$$\begin{cases} Group1\ ,\ if\ \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \\ Group2,\ if\ \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}. \end{cases} \tag{2.36}$$

From this it is seen that the classification criteria of maximizing the posterior probability is the same as the classification criteria of minimizing the total misclassification probability.

2.6.2 Fisher's Linear Discriminant Analysis (FLDA)

The primary purpose of FLDA is to separate samples of distinct groups as much as possible by transforming the multivariate observations to univariate observations that are optimal for distinguishing between the classes [146].

Suppose we have two populations. Let $X_1, X_2, \cdots, X_{n_1}$ be the n_1 observations from population π_1 and let $X_{n_1+1}, X_{n_1+2}, \cdots, X_{n_1+n_2}$ be n_2 observations from population π_2. The first step in FLDA is to project these $p \times 1$ vectors to a scalar output via a linear function

$$Y(X) = V^t X, \tag{2.37}$$

where V^t is a $1 \times p$ vector of coefficients.

Then, find the vector $\hat{V}$ that maximises the separation function $S(V)$, which is the ratio of the squared distance between the transformed means of the two groups relative to within group variance

$$S(V) = \frac{(\overline{Y_1} - \overline{Y_2})^2}{{S_Y}^2}, \tag{2.38}$$

where $\overline{Y_1}$ and $\overline{Y_2}$ are the univariate means of group 1 and group 2, respectively, and S_Y is the sample standard deviation.

$\hat{V}$ is found by solving the equation based on the first derivative of $S(V)$ [151] and it is of the form

$$\hat{V} = cS_{pooled}^{-1}(\overline{X_1} - \overline{X_2}), \tag{2.39}$$

where c is some non-zero constant, $\overline{X_1}$ and $\overline{X_2}$ are multivariate means of group 1 and group 2, respectively, S_{pooled} is the pooled sample covariance matrix derived from covariance matrices S_1 and S_2 of group 1 and group 2, respectively.

Thus, Fisher's discriminant function is obtained as below

$$Y(X) = \hat{V}^{-1}X = c(\overline{X_1} - \overline{X_2})^{-1}S_{pooled}^{-1}X. \tag{2.40}$$

The scalar output Y from the function above can be referred to as a discriminant score. A threshold value for the discriminant score is decided and often uses the midpoint between the two means, $\overline{y_1}$ and $\overline{y_2}$. So, the final step in FLDA is to perform classification using an allocation rule by comparing the discriminant score to the threshold value (Figure 2.12).

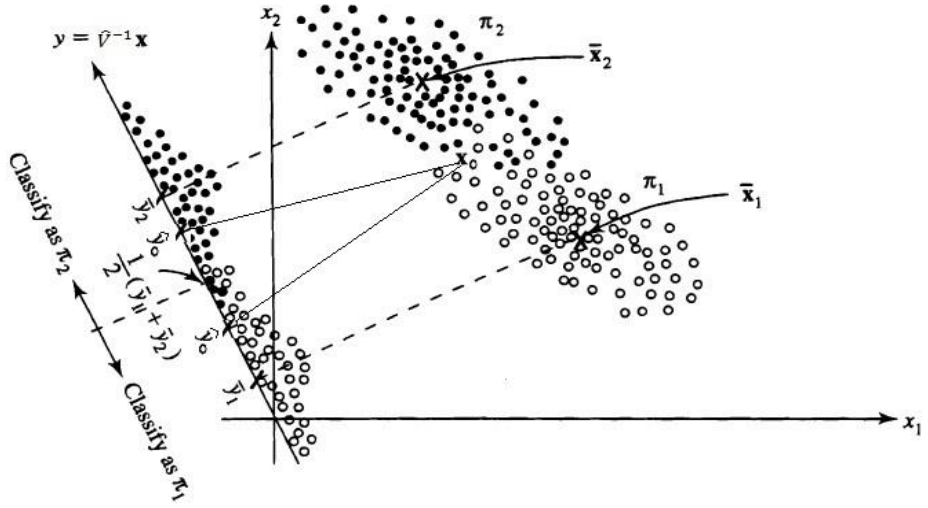


Figure 2.12: Classification based on the Fisher's discriminant function.

For example, suppose we have two populations π_1 and π_2 with common covariance matrices, we can then classify an observation x_0 to some class based on Fisher's discriminant function with the constant c set to unity:

$$\begin{cases} Population\ \pi_1, \quad if\ (\overline{x_1} - \overline{x_2})^{-1} S_{pooled}^{-1} x_0 \geq \frac{1}{2}(\overline{y_1} + \overline{y_2}) \equiv \frac{1}{2}(\overline{x_1} - \overline{x_2})^{-1} S_{pooled}^{-1}(\overline{x_1} + \overline{x_2}) \\ \qquad\qquad \text{That is, } \hat{y}_0 \text{ is on the right hand side of } \frac{1}{2}(\overline{y_1} + \overline{y_2}) \text{ (closer to } \overline{y_1}), \\ Population\ \pi_2, \quad if\ (\overline{x_1} - \overline{x_2})^{-1} S_{pooled}^{-1} x_0 < \frac{1}{2}(\overline{y_1} + \overline{y_2}) \equiv \frac{1}{2}(\overline{x_1} - \overline{x_2})^{-1} S_{pooled}^{-1}(\overline{x_1} + \overline{x_2}) \\ \qquad\qquad \text{That is, } \hat{y}_0 \text{ is on the left hand side of } \frac{1}{2}(\overline{y_1} + \overline{y_2}) \text{ (closer to } \overline{y_2}), \end{cases}$$

(2.41)

2.7 ROC Analysis

Receiver operating characteristic (ROC) analysis is used to evaluate the performance of a binary classifier system. An ROC curve is created by plotting the sensitivity against the specificity at various discrimination thresholds [152, 153]. ROC analysis has been used extensively in medical application [152-154]. ROC analysis has been used extensively in medical applications [155-157].

2.7.1 Accuracy, Sensitivity and Specificity

Accuracy, sensitivity, and specificity are the terms that are commonly associated with the statistical measure of the performance of a binary classification test. Accuracy provides a single number for classification performance. It is defined as the fraction of correctly classified instances divided by the total number of instances. In general, it is accepted that the higher the accuracy, the better the classification performance. However, accuracy is too simple in some cases. Using accuracy to evaluate the performance of binary classification in medical decision making is of limited usefulness as a conclusion based on accuracy alone is highly unreliable [152, 153]. For example, in screening for cancer, if a subject without cancer is misclassified (misdiagnosed) as having cancer, then further examination will usually reveal the error. On the other hand, if a subject with cancer is misclassified as being healthy, the cancer is likely to progress unchecked leading to severe complications or even death. In addition, using accuracy is sometimes a problem if the classes are hugely imbalanced. For example, cancer is generally rare, say, with occurrence of 1 in 100 screenings. If a classifier is adjusted to return a negative finding every time, it would still have an accuracy of 99%.

Accordingly, additional performance measures, such as sensitivity and specificity, were introduced to report classifier performance in situations where the consequences of different classification errors are not equal. By definition, sensitivity, or true positive fraction (TPF), specifies the accuracy of identifying the positive subject correctly, and is calculated as the number of positive subjects correctly assigned (TP) divided by the total number of actually positive subjects (P). Similarly, specificity, or true negative fraction (TNF), specifies the accuracy of identifying the negative subject

correctly and is calculated as the number of negative subjects correctly assigned (TN) divided by the total number of actually negative subjects (N). Both sensitivity and specificity are values in the range 0 to 1.

2.7.2 ROC Curve

An ROC curve plots the trade-off between sensitivity and specificity for several discrimination thresholds. It is a simple, yet meaningful, description of the performance of a binary classifier. In addition to TPF and TNF, there are two more possible outcomes, false positive fraction (FPF) and false negative fraction (FNF) (Table 2.2). The relationship between these four fractional quantities is

$$\begin{aligned} TPF+FNF=1 \\ TNF+FPF=1. \end{aligned} \qquad (2.42)$$

Table 2.2: The four test outcomes of a binary classification.

	Condition positive	Condition negative
Test outcome		
positive	True positive fraction TPF = TP/P	False positive fraction FPF = FP/N
negative	False negative fraction FNF = FN/P	True negative fraction TNF = TN/N

FP: the number of negative subjects incorrectly assigned (False positive); FN: the number of positive subjects incorrectly assigned (False negative).

Figure 2.13 illustrates the four possible outcomes of classification. Increasing the threshold results in increasing TNF and FNF but TPF and FPF will be reduced. Accordingly, optimal trade-off between sensitivity and specificity is necessary by selecting an appropriate threshold.

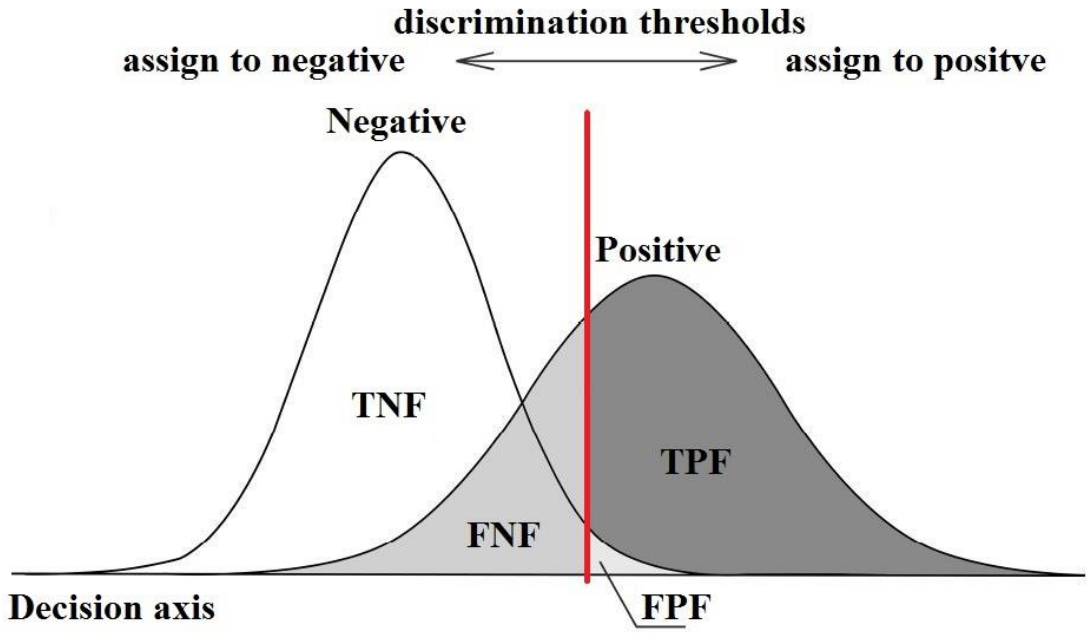


Figure 2.13: Illustration of the four possible test results (TNF, FNF, TPF, FPF) defined by a discrimination threshold (the red vertical line). The left curve represents the true negative group and the right one represents the true positive group. Different performance results are defined by moving the discrimination threshold along the decision axis.

Figure 2.14 illustrates the process of generating an ROC curve. Each point on the ROC curve in Figure 2.14b represents an FPF (horizontal axis), TPF (vertical axis) pair corresponding to a particular decision threshold in Figure 2.14a. As a result, an ROC curve is a plot of FPF versus TPF obtained by moving the discrimination threshold along the decision axis. Here we take four operating points (P_1, P_2, P_3, P_4) on the ROC curve in Figure 2.14b. These four points are generated by plotting each pair of TPF and FPF corresponding to the four different decision thresholds (D_1, D_2, D_3, D_4) respectively in Figure 2.14a. An ROC curve is obtained by plotting all possible FPF TPF pairs.

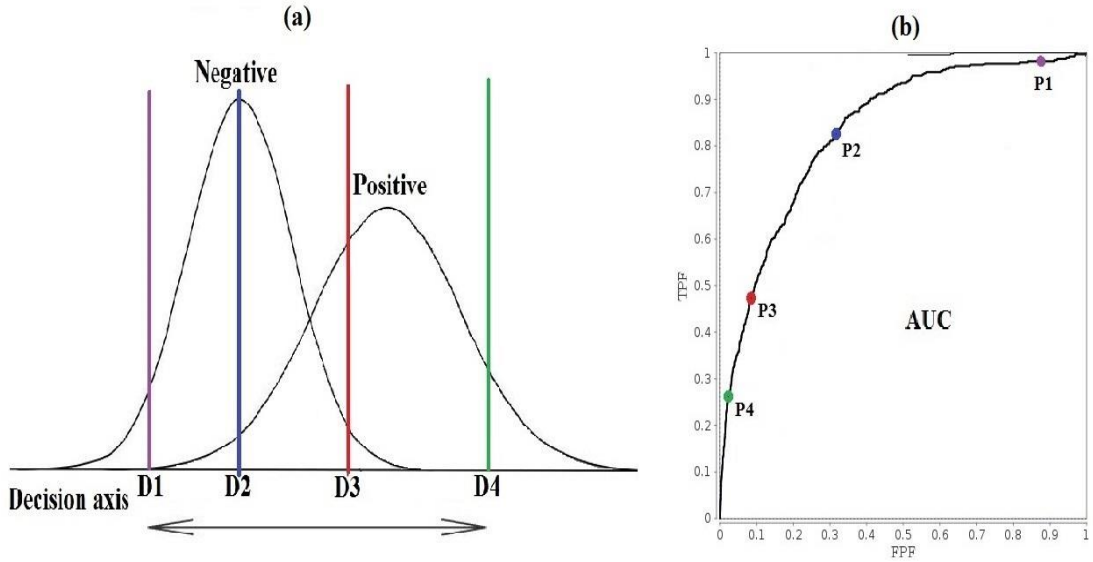


Figure 2.14: The process of generating an ROC curve. (a) Shows the four decision fractions for each of four different decision thresholds (D_1, D_2, D_3, D_4). (b) Shows four operating points (P_1, P_2, P_3, P_4) on the ROC curves corresponding to the four different decision thresholds in (a): P_1 corresponding to D_1, P_2 corresponding to D_2, P_3 corresponding to D_3, P_4 corresponding to D_4.

An example of three different ROC curves is shown in Figure 2.15. High sensitivity corresponds to a larger TPF value on the ROC curve, and high specificity corresponds to a smaller FPF value on the ROC curve. Naively, the optimal decision threshold corresponds to a point on the ROC curve nearest to the upper left corner of the ROC graph since this seems to correspond to a sensible balance between high TPF and low FPF. However, this is not always true. In some screening applications, detecting abnormal cases successfully is more important than maximizing specificity. In this case, the ideal operating point on the ROC curve will move from the upper left corner toward to the upper right corner. In contrast, maximizing specificity is of more concern in prostate cancer screening than maximizing sensitivity because benign enlargement of the prostate can cause high prostate specific antigen (PSA) values, and false positives are quite common in this case [154].

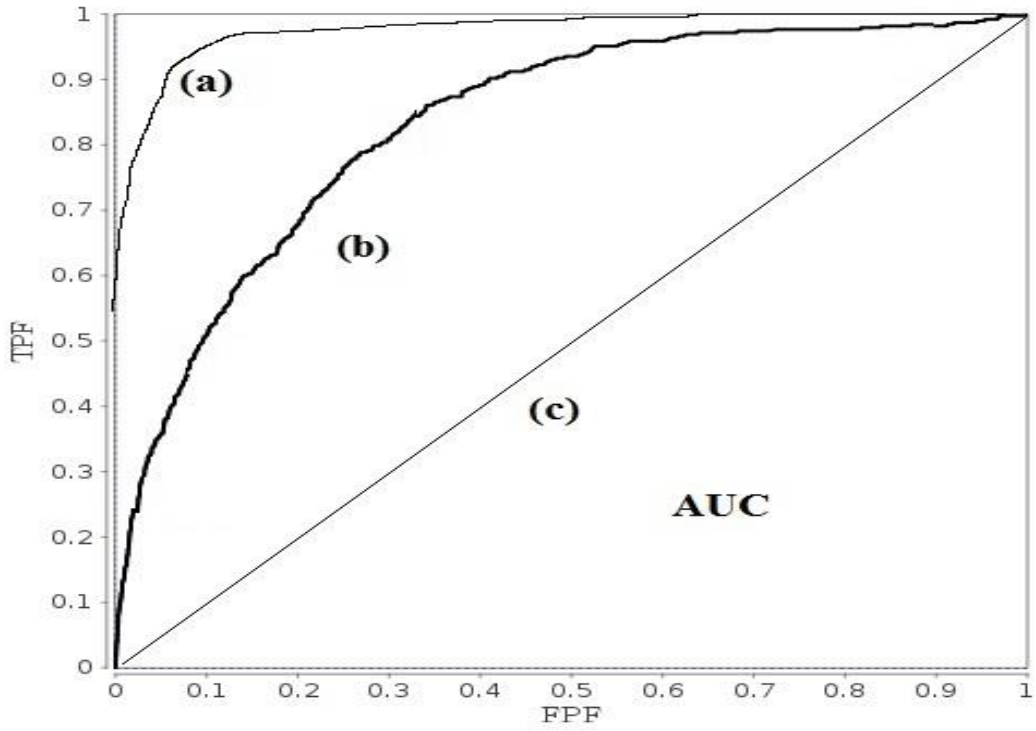


Figure 2.15: An example of three different ROC curves. ROC curve (a) has the best discriminating performance among the three. ROC curve (b) has the second best performance. The ROC curve (c) along the diagonal line from (0, 0) to (1, 1) indicates no discriminant power.

2.7.3 Area under an ROC Curve

A useful measure of classification performance in many medical contexts is the area under the ROC curve (AUC), ranging from 0.0 to 1.0. In general, a higher AUC score indicates better classification performance. Classification is perfect if the AUC is 1.0 as both the sensitivity and specificity are 1.0 so there are no false positives and no false negatives. If the ROC curve is diagonal, the AUC score is 0.5 meaning that the classification cannot discriminate between the two groups. The training AUC for an ROC curve must be between 0.5 and 1.0 as shown in Figure 2.15. However, the testing AUC can be less than 0.5 since, in this case, the discriminant surface is fixed ahead of time and the points to be classified may, in principle, fall anywhere. The AUC is widely used in machine learning to describe and compare the performance of a learning scheme.

2.8 Feature Selection

2.8.1 Introduction to Feature Selection

The size of a training set should increase exponentially as the number of dimensional features increases [158]. In practice, however, the number of possible features is often naturally large and the amount of data is often limited. Also, it is found that the classification performance goes up with the number of features to a point but then decreases or fluctuates after that. Feature selection techniques are commonly used to cope with this problem by effectively reducing the dimension of the feature space while still retaining a high level of discriminating power.

Feature selection is the process of selecting a subset of useful and relevant features in the training set and using only this subset as features in classification. Keeping irrelevant features in the dataset may result in over fitting. Over fitting occurs when a model is too complex with many parameters relative to the number of observations, and tends to capture the noise of the dataset.

Feature selection serves four main purposes. First, fewer features are desirable as it means shorter runtime during execution and less computational cost. Second, feature selection helps to reduce the complexity of the model and, therefore, facilitates understanding and explanation [159]. Third, feature selection increases classification accuracy by eliminating noise features that are unneeded, irrelevant and redundant and, therefore, can be ignored without incurring much loss of information. Finally, feature selection improves generalization by reducing over fitting (reduction of variance) [160, 161]. It has been shown that classification can still over fit if it is trained on high dimensional feature spaces even if the training set is large. This leads to poor classification performance on an unseen testing set [162, 163]. Feature selection is often essential in attaining high quality classification results.

Given a feature set $X = \{x_i | i = 1, \dots, M\}$, feature selection constitutes searching for a subset $Y = \{x_j | j = 1, \dots, N\}$, with $N < M$, that optimizes an objective function. Accordingly, feature selection requires a search strategy to select candidate subsets with the objective function evaluating these candidates and feeding results back to the search strategy to allow it to select new candidates.

There have been several search strategies developed to explore a feature space in an efficient fashion. These search strategies can be typically divided into three categories: exponential algorithms, sequential algorithms, and randomized algorithms [164]. Exponential algorithms, such as exhaustive search and Branch and Bound (B&B) evaluate a number of subsets that grow exponentially with the number of features and thus these algorithms have high complexity $O(2^n)$ and are frequently too expensive to use [164]. Sequential algorithms, such as sequential forward selection and sequential backward selection, add or remove features sequentially until some termination criterion is met. This algorithm has relatively low computational complexity but does not examine all possible subsets and so it is not guaranteed to find the optimal subset. In addition, this algorithm tends to fall into local minima, which is caused by a so called nesting effect, resulting in there being no possibility of discarding a feature once it has been selected [165]. Finally, randomized algorithms, such as genetic algorithms, incorporate randomness as part of the feature selection process to try to avoid the problem of the nesting effect.

Objective functions can be broken up into two broad categories: filters and wrappers [166] [167]. Filters evaluate feature subsets based on the characteristics of training data such as interclass distance, statistical dependence or information-theoretic values without any learning algorithm or classifiers involved. Wrappers evaluate and determine the feature subsets by the performance of the predetermined learning algorithm by using statistical resampling or cross-validation [166]. Wrappers generally tend to achieve better learning performance as better feature subsets are selected, but it also tends to have higher computational cost and less generality than filters. In contrast, filters have the advantage of computational efficiency and are preferable to wrappers when there are a large number of features. In addition, filters can execute much faster and exhibit more generality than wrappers as filters evaluate the intrinsic properties of the data non-iteratively only, rather than interacting with a particular classifier [168].

2.8.2 Sequential Feature Selection

In sequential feature selection, only one feature among all successors is selected at each selection stage. This method gives completeness, but does not guarantee global

optimality of the selected subset of features, since local optimality is possible. In addition, sequential feature selection suffers from the so-called nesting effect because a feature that has been selected or removed cannot be removed or selected at a later stage. There are several types of sequential feature selection methods discussed in the pattern recognition literature including sequential forward feature selection [169], sequential backward feature selection [170], bidirectional feature selection [171], and sequential floating feature selection [172].

Sequential forward feature selection starts with the empty feature set and repeatedly includes the most significant feature from the features not yet selected until no further improvement of the objective function can be achieved. The most significant feature is the feature that results in the best value of the objective function among the remaining features when used along with the previously selected features. In contrast, sequential backward feature selection starts with the full feature set and repeatedly eliminates the feature that contributes the least to the objective function until the removal of further features does not improve the objective function. Sequential forward feature selection works well if the number of features in the optimal subset is small; sequential backward feature selection is preferred if this is not the case [173]. A problem with these search techniques is that when a feature becomes obsolete after the inclusion of other features, it is still retained during forward feature selection, and cannot be discarded, while a feature eliminated in sequential backward feature selection cannot be re-evaluated for usefulness once removed.

In bidirectional feature selection, both sequential forward and sequential backward feature selection are performed [174]. Sequential forward feature selection is performed from the empty feature set at the same time that sequential backward feature selection is performed from the full feature set. Convergence of the feature selection from both directions is ensured by not adding features eliminated and not eliminating features added. One variant of bidirectional feature selection is plus-L minus-R feature selection. This is a version of sequential forward feature selection and sequential backward feature selection that allows some backtracking during the selection process [174]. Plus-L minus-R feature selection is a bottom-up procedure if $L > R$. It starts with an empty feature set and repeatedly adds L features using sequential forward feature selection and then removes the worst R features via sequential backward

feature selection. In contrast, plus-L minus-R feature selection is a top-down procedure if L < R, starting with the full feature set and repeatedly removing R features followed by L additions until the required number is achieved. Plus-L minus-R feature selection attempts to compensate for the weaknesses of sequential forward and backward feature selection. However, the lack of a theoretical way of predicting the optimal values of L and R to achieve the best feature subset is its main limitation.

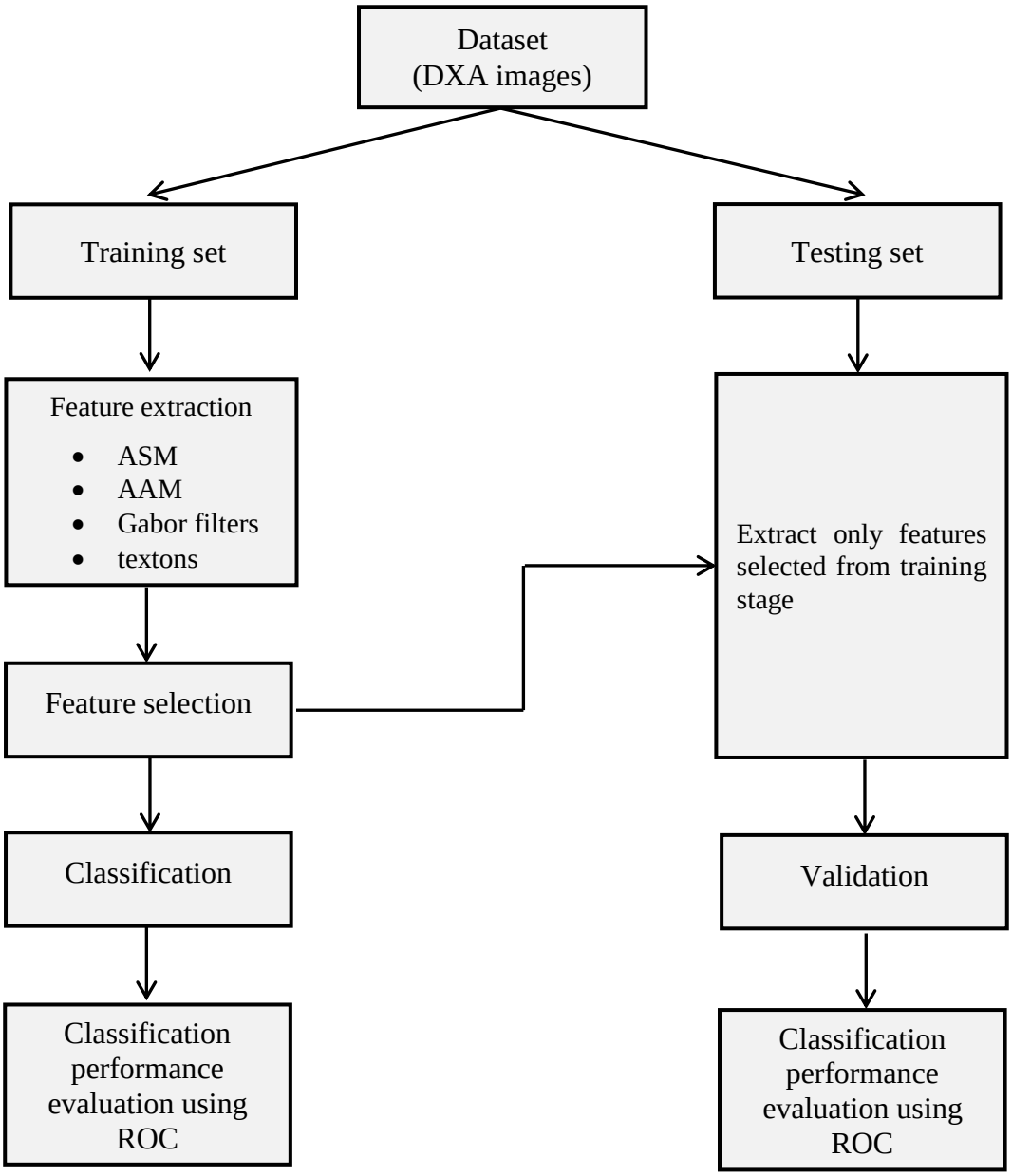
Alternatively, sequential floating feature selection is an extension to the plus-L minus-R feature selection with flexible backtracking capabilities. Instead of fixing the values of L and R during plus-L minus-R feature selection, sequential floating feature selection allows those values to be flexibly changed, i.e., floating up and down during the search so as to approximate the best feature subset as much as possible. There are two types of sequential floating feature selection designed and implemented based on the dominant direction of the search. The search in the forward direction is referred to as sequential floating forward selection, while in the opposite direction it is called sequential floating backward selection. The former starts from the empty feature set and performs, after each forward step, a number of backward steps as long as the value of the objective function of the corresponding feature subset is better than the previous ones. In contrast, sequential floating backward selection starts from the full feature set and performs, after each backward step, a number of forward steps as long as the objective function improves. Even so, there is no guarantee that the optimal subset of features will be found.

2.8.3 Exhaustive Search Feature Selection

As the name suggests, exhaustive search feature selection methods evaluate all possible combinations of the input features in order to determine the best subset of features that optimizes the objective function. Such a feature selection method is an optimized search that guarantees the best subset. There are $2^n - 1$ possible feature subsets for a feature space of dimension *n*. Hence, this method is feasible only for feature sets of low dimension. However, if the number of features selected is limited beforehand to *k* features, then the total number of subsets to consider is reduced to $\binom{n}{k}$. Hence exhaustive search becomes practical even for fairly large *n* as long as *k* is small.

3 MATERIAL AND METHODS

This chapter describes the dataset and methods used in this thesis to estimate the risk of fracture based on DXA images. These methods analysed the quantitative characterization of the shape and gross structure of the proximal femur, i.e., statistical shape and appearance models for the proximal femur were constructed. In addition, this study also captured image texture distribution in order to improve risk assessment. The image texture methods used are based on Gabor filters and textons. The process steps between image acquisition and fracture risk analysis are presented in the following diagram.



3.1 DXA Dataset

Data for this study comprised DXA images from a set of 119 individuals who participated in the Hertfordshire cohort study [175].

3.1.1 The Hertfordshire Study

This cohort study was set up to evaluate the interaction between the genome, the intrauterine and early postnatal environment, adult diet and lifestyle and the risk of cardiovascular disease in later life [175]. 3000 men and women born in Hertfordshire between 1931 and 1939 were recruited and information about the early environment of individuals, adult diet and lifestyle, and their health outcomes 60 years later were recorded. The entire cohort was also followed up through primary care and hospital records over a 10-year period (1998–2007). Of the 3000 subjects who attended a clinic in Hertfordshire, 966 (498 males, 468 females) returned for DXA bone scan and knee radiography using a Hologic QDR4500 scanner. The results of the Hertfordshire study showed that foetal and post-natal growth is associated with adult disease [175].

3.1.2 Data Subset Used in This Study

In this study, images from 119 subjects (33 males, 76 females) from the Hertfordshire study were collected. All subjects in the dataset were from the same ethnic background. These images were taken around the individuals' 66^{th} birthday (range 59–74 years old). All 119 subjects underwent a scan of the left hip by DXA. Several basic parameters were measured and recorded: aBMD for the femoral neck and total hip, T-score for the femoral neck and total hip, hip axis length (HAL), and neck shaft angle (NSA). The pixel resolution of digitized proximal femur radiographs is 250 x 300 pixels with an effective depth of 8 bits.

The subjects in this dataset included a fracture subgroup comprising 29 white subjects with reported low-energy fractures (10 males, 19 females), and a control group comprising 90 white subjects without a history of low-energy fractures (33 males, 57 females). The type of fracture was not restricted to hip fracture but also included wrist, hip, lower limb, spine fractures, etc. The whole data set was divided

into two folds with almost equal control-to-fracture ratio, fold A with 60 subjects (15 fracture subjects, 45 control subjects) and fold B with 59 subjects (14 fracture subjects, 45 control subjects) (Figure 3.1). Owing to the small sample size, 2-fold cross-validation was conducted (especially due to the small number of fracture subjects). Fold A was used as the training set to develop the methods and train the classifiers and fold B was reserved as a testing set. Then the roles were reversed so that fold B was used as the training set and fold A was used as the testing set.

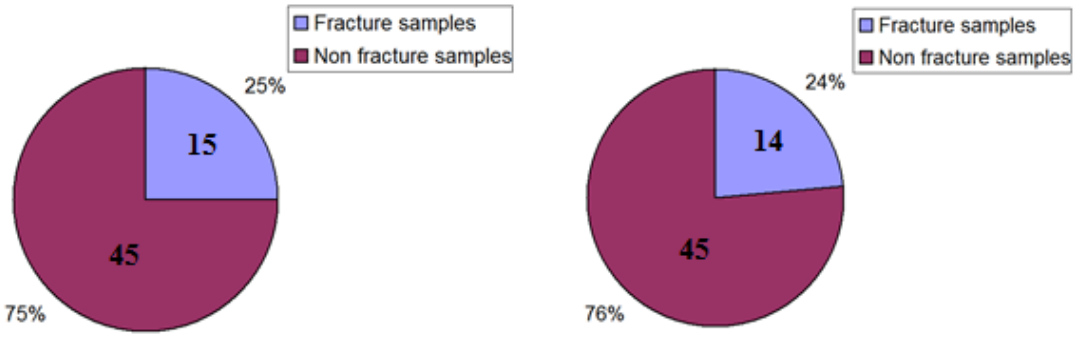


Figure 3.1: Profile of data set used in this study. (Left) Fold A with 60 subjects (15 fracture subjects, 45 control subjects); (Right) Fold B with 59 subjects (14 fracture subjects, 45 control subjects).

3.1.3 Note on Image Data Collection

The results appearing in this thesis are estimates based on data collected from a small fraction (119 images) of the total subjects in the Hertfordshire study. A brief history of this study explains why only a small number of DXA images were used.

This study was undertaken with the understanding that hundreds of DXA images from the large longitudinal study (the Hertfordshire cohort study) conducted in Southampton in the UK would be available [175]. The group at Southampton sent some preliminary images that were used to confirm that the techniques planned for this study could be applied. The images were of low resolution since this preliminary version was taken from the radiologists' reports rather than copies of the original DXA

images, but they demonstrated the feasibility of this study.

Researchers associated with the Hertfordshire study explained that staffing limitations and problems with assessing large amounts of data stored on very old machines would result in delays in providing a more extensive data set. The project proceeded since this time could be spent exploring and tuning methods for assessing risk based on the initial data set of 119 low resolution images. Much later it transpired that the difficulties in providing full resolution data were not just a matter of manpower or time. Full resolution images only existed in the proprietary format of Hologic. While it is possible to display these images, mark them up and apply simple image processing steps within the Hologic framework, the images cannot be exported in a format (e.g., DICOM, JPEG, TIFF) that allows the application of novel image analysis methods via non-Hologic platforms. A considerable amount of time was spent attempting to convert the files. Advice from experts around the world was sought and paid for but converting the images has not been possible. We note that it is possible to reformat some modern Hologic images, but not the older DXA images in the Hertfordshire study, which were collected in the late 1990s.

All we were able to access were the report images as described in the thesis. The group at Southampton reports that extracting these files is labour intensive since they are stored on very old machines—a consequence of the longitudinal aspect of the study for which this data was collected. This has limited the number of images that could be extracted. We also sought other sources of data including, for example, data from the Study of Osteoporotic Fractures (SOF) at the California Pacific Medical Center in 2015. We applied for, paid for, and received access to this data only to discover that these images were also encoded in the Hologic proprietary format using old machines (QDR 1000) and again we could not decode these images despite seeking advice far and wide.

All this took an enormous amount of time while work on the project using the original 119 images continued in parallel. Eventually, it was necessary to resign to the fact that this study would be conducted using a much smaller data set of much lower resolution images than originally planned. Subsequently, the methods used in this study were carefully chosen to mitigate the shortcoming of the dataset.

3.2 Implementing Active Shape and Appearance Models

In this study, ASM and AAM of the proximal femur were built by processing the information captured from the training set of 2D DXA images based on the methods of Cootes et al [70, 100, 101](Section 2.4).

3.2.1 Implementing Active Shape Model

As introduced in Section 2.4.1, the first step in building the ASM model was to manually identify enough landmark points on the boundary of the femur in each image in the training set. To place landmark points along the boundary, the digital DXA image was displayed on a computer screen and the author used mouse clicks to identify boundary points. In-house software written in MATLAB® was used to capture mouse-click locations and store the point locations as (x, y) coordinates. In this study, 44 landmark points were used to outline the proximal femur in each DXA image. Care was taken to place the landmark points consistently across all the images. To do this, closely spaced landmark points were placed along significant morphological features of the femur, including the femoral neck and greater trochanter (Figure 3.2). However, the lesser trochanter and the femoral head were not included among the main shape features since the lesser trochanter did not appear with sufficient clarity in many images. The appearance of the lesser trochanter depends heavily on its orientation. The femoral head was not included because it is often masked, partly or in full, by the pelvis and this would result in uncertain assignment of its boundary. The same number of landmark points was used across all the images for particular significant features. An additional number of landmark points was placed on the boundary of the femur, roughly evenly spaced, between the significant features (Section 2.4.1.1).

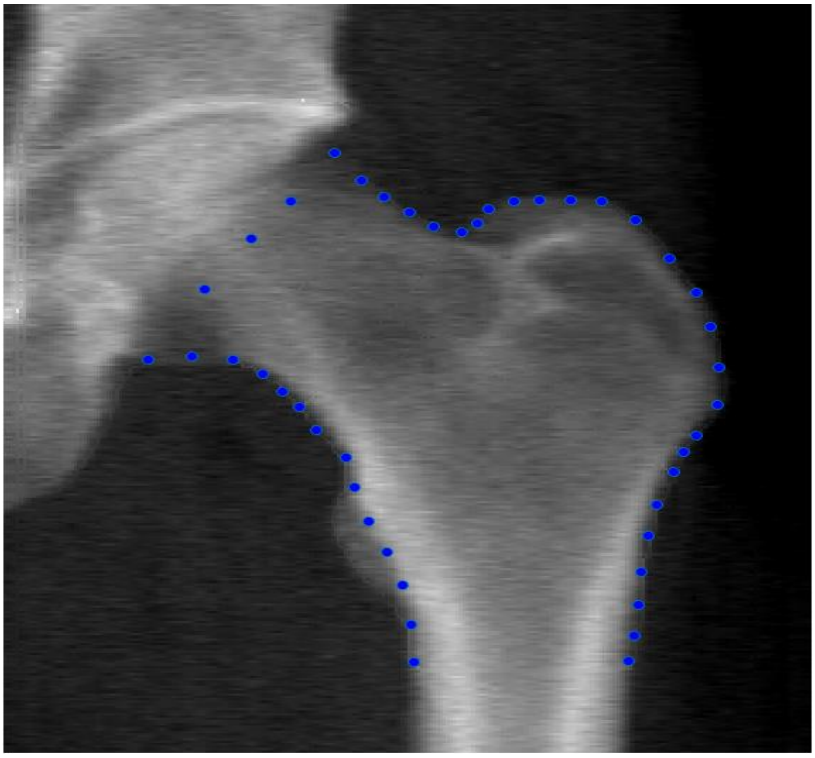


Figure 3.2: An example of 44 manually selected contour points of the boundary of the proximal femur.

The number of these were also consistent over all the images (Figure 3.3). It is important to note that the landmark points were deliberately spaced non-uniformly along the boundary in order to concentrate shape information at highly curved parts of the femur outline. The exact locations of the boundary points do not matter because there are enough of them and because care is taken to be somewhat (though not necessarily exactly) consistent between subjects. The literature reports that the exact location of the boundary points does not significantly impact the final model as long as sufficient care is taken to be somewhat consistent between subjects [176]. One reason for the robustness of these models against the exact placement of landmark points is that in the next step, PCA (Section 2.4.1.3) removes fine shape anomalies while retaining significant shape information.

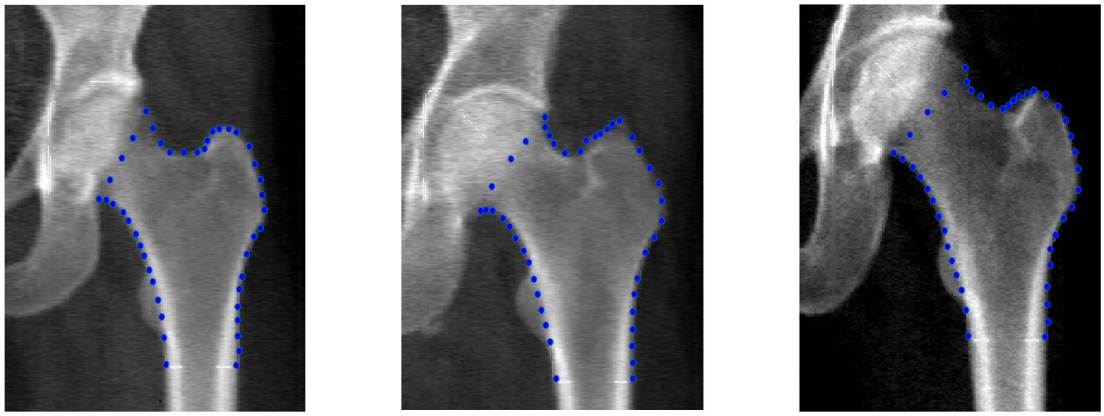


Figure 3.3: An example of 44 selected contour points consistent over three subjects.

In order to compare shapes, the landmark points in the training set of images were aligned into a common coordinate frame using Procrustes analysis including scaling, rotating, and translating to eliminate differences between femur size, orientation and position within the image as introduced in Section 2.4.1.2 (Figure 3.4).

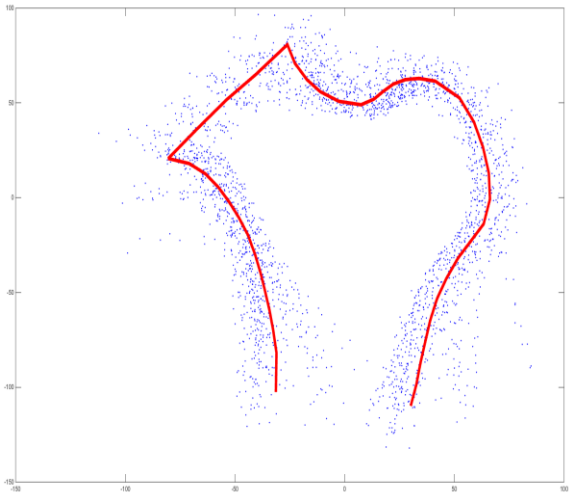


Figure 3.4: An example of alignment contour points and the corresponding mean shape (red line).

From these aligned landmark points, a point distribution model was constructed as introduced in Section 2.4.1.3. Each boundary point represents two items of information (x and y coordinate) that contribute to the total description of the shape of the femur. However, since the boundary of a femur in a DXA image is fairly smooth compared to the spacing of the boundary points, information represented by the 44 boundary points of a femur contains redundancies that may obscure further processing.

Accordingly, PCA was used to decompose the shape information into 60 principal modes (eigenvectors) listed in decreasing order of information content according to the value of the associated eigenvalue. Only the first 12 principal modes were used in further analysis in the ASM as they were found to represent 99% of the overall variations in shape within the training dataset, and therefore an example, $\hat{X}$, can be approximated using

$$\hat{X} = \bar{X} + \sum_{s=1}^{12} b_s \Phi_s, \tag{3.1a}$$

$$\bar{X} = \frac{1}{60} \sum_{i=1}^{60} X_i, \tag{3.1b}$$

where $\bar{X}$ is the mean shape vector over all aligned 60 training subjects, Φ_s is principal mode number s. The 12 weighting values of the principal modes (b_1, ..., b_{12}) were used as shape features for each training subject and as input for later feature selection.

This point distribution model generated from the training process was used to interpret unseen examples of shapes by varying the parameters b_s within suitable limits (as exemplified by the hand shape demonstration in Figure 2.8c). Having generated this model, we use it to fit femurs from the testing dataset in an iterative way, starting from the mean shape. This involves finding the shape and pose parameters that make the model coincide with the structures of the testing images using the iterative search approach introduced in Section 2.4.1.4. Accordingly, the 12 updated values of the principal modes (b_1, ..., b_{12}) as derived from the model and fitted for each testing subject were also used as shape features for each testing subject. Examples of fitting the model to three proximal femurs from the testing set are presented (Figure 3.5).

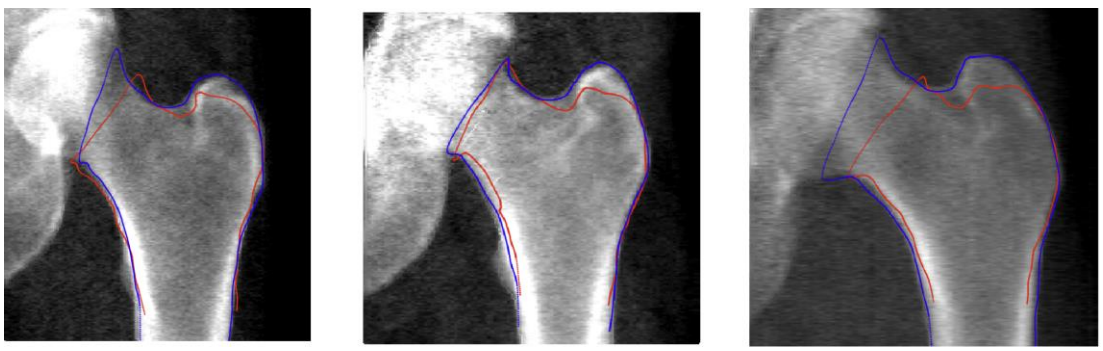


Figure 3.5: Results (blue lines) of fitting ASM model to three proximal femurs from the testing data set. The red lines are initial contours for iterative searches.

3.2.2 Implementing Active Appearance Model

In AAM (Section 2.4.2), a model of shape variation and a model of grey-scale variation in a shape-free frame were combined to simultaneously explain shape and grey-scale (referred to as the appearance). The same training set was used with the 44 boundary points already defined for ASM. Then, Procrustes analysis was used to align the shapes of the training images and PCA was applied to generate a model of shape variation as introduced in Section 2.4.1.3.

To access this information in the grey-scale variation, a procedure similar to the ASM was performed on the grey-scale values of pixels within the region of the femur in the DXA images. To do this, we used a Delaunay triangulation algorithm [118] to warp each training image so that its boundary points matched the mean shape, obtaining a shape-free patch. Then, the grey-scale values at 82,830 evenly spaced pixels from the shape-free image over the region covered by the mean shape were recorded (Figure 3.6). Because shape-free patches were used, the locations of the pixels were reasonably consistent between femurs in different images (Section 2.4.2),

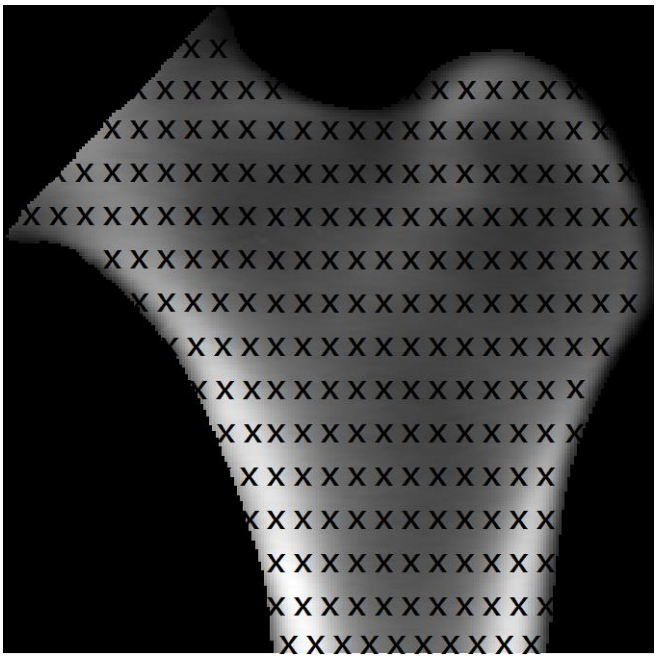


Figure 3.6: An example of normalized proximal femur image showing locations of sample points for characterising intensity patterns. The locations shown are indicative only of the regular sampling pattern used. The actual number of pixels sampled (82,830) is too large to display on the image.

By applying PCA, a model of grey-scale variation was obtained. The shape and grey-scale variation from the same image can be combined. Finally, a further PCA was applied to the data to generate a combined appearance model with a series of principal modes of shape and grey-scale variation. Only the first 57 principal modes were used in further analysis in this study as they represent 99% of the overall variations in appearance within the training data set. Thus, the appearance of the proximal femur within the population, including shape $\tilde{X}$ and grey-scale $\tilde{G}$ can be described as

$$\tilde{X} = \bar{X} + \sum_{a=1}^{57} c_a \, \Phi_a \qquad \bar{X} = \frac{1}{60} \sum_{i=1}^{60} X_i \qquad \text{(3.2a)}$$

$$\tilde{G} = \bar{G} + \sum_{a=1}^{57} c_a \, \Psi_a \qquad \bar{G} = \frac{1}{60} \sum_{i=1}^{60} G_i, \qquad \text{(3.2b)}$$

where $\bar{X}$ is the mean shape of all 60 aligned training subjects, $\bar{G}$ is the mean grey-scale of all 60 training subjects in a shape-free frame, and Φ_a and Ψ_a are the principal modes of shape and grey-scale variation, respectively. The 57 weighting values of the principal modes (c_1, ..., c_{57}) control the shape $\tilde{X}$ as well as grey-scale $\tilde{G}$, and, therefore, were used as appearance features for each training subject and as input for feature selection later (Section 2.4.2).

The models of shape and grey-scale variation generated from the training process were used to interpret appearance of examples by varying the parameters c_a. We used this model to fit test images. Accordingly, the 57 updated values of the principal modes (c_1, ..., c_{12}) derived from the model as fitted for each testing subject were also used as appearance features for each testing subject (Section 2.4.2).

3.3 Implementing Texture Analysis

3.3.1 Regions of Interest

As introduced in Section 1.2.1, many studies focus on a specific ROI. The ROI chosen in this study were the whole hip, the whole femoral neck, neck slice, Ward's triangle, narrow neck, inter-trochanter and femoral shaft for bone texture analysis using Gabor filters and textons (Figure 3.7).

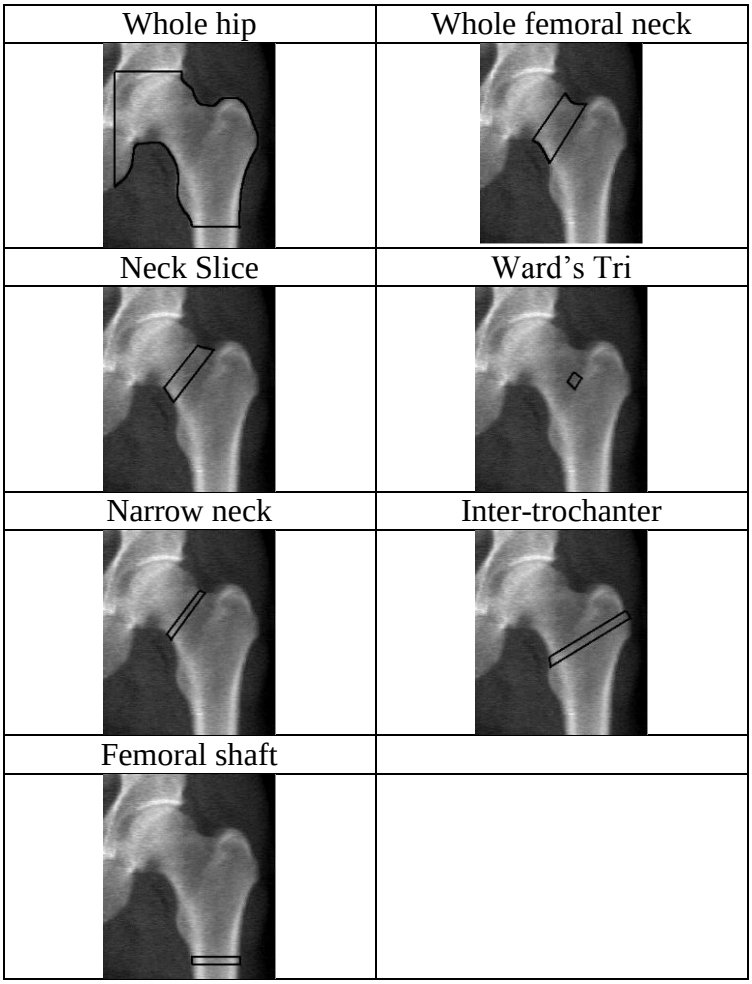


Figure 3.7 : Regions of Interest (ROI) considered for computing texture features using Gabor filters and textons.

3.3.2 Gabor Filters

To implement Gabor filters for texture analysis, the outputs of a symmetric ($\varphi = 0$) and an anti-symmetric filter ($\varphi = \pi/2$) at each image pixel were combined to form the Gabor energy (Section 2.5.1). The value of the remaining parameters for Gabor filters constructed from the Gabor kernel function in Equation 2.18 were as follows. (1) Four spatial frequencies, $1/\lambda$ = 1/16, 1/18, 1/26 and 1/32, were used in the cosine factor of the Gabor kernel function. (2) Six orientations, $\theta = n\pi/6$, n = 0, 1, 2, 3, 4, 5, were used. (3) The spatial aspect ratio was set as γ = 0.5. (4) The value of σ was specified by setting the spatial frequency bandwidth b = 1. This resulted in a bank of 24 Gabor filters applied to extract the texture information. Thus, 24 Gabor-filtered images were produced from each original image (Figure 3.8). For each Gabor-filtered image, the

total Gabor energy was computed using Equation 2.21. Accordingly, Gabor features, (g_1, ..., g_{24}), in the form of Gabor energy, were extracted for each ROI for each subject (Section 2.5.1).

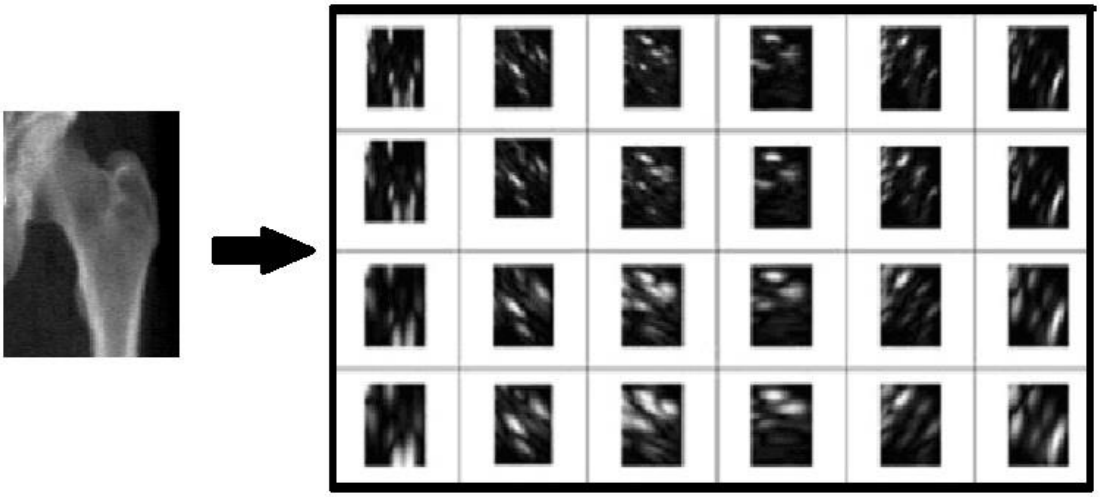


Figure 3.8: Examples of Gabor filter outputs at various frequencies and orientations. (Left) The original image; (Right) 24 Gabor-filtered images with four different spatial frequencies (1/16, 1/18, 1/26 and 1/32 from top to bottom) and six different orientations (left to right).

3.3.3 Textons

Two classes of texton features were extracted from images in this study: Gabor filters based textons and local 3x3 neighbourhood-based textons. To compute Gabor filters textons, the 24 Gabor filters described in the previous section were applied to all the ROI in all the training images. This resulted in 24 filter responses for each pixel in each ROI. For every pixel p, the 24 responses at p were used to form a vector v_p of length 24. For all the single ROI types (for example for the whole femoral neck), the vectors v_p from all the pixels in this ROI for all images in the training set were viewed as points in a 24-dimensional representation space. To form the textons dictionary, K-means clustering was applied on the filter responses of ROI from the training images (Section 2.5.2).

An experiment was conducted to determine the number of textons needed to separate the two classes of ROI, since this was not known ahead of time. K-means clustering was repeatedly applied for four values of K (K = 10, 20, 30, 40) on the filter responses

of ROI on all training images and the AUC scores were compared. The results indicated that 20 textons (K = 20) were suitable in this study (Table 4.3 b).

We also conducted an experiment to compare K-means clustering methods. K-means clustering was applied in two ways: 1) the filter responses for a specific ROI for all the training images (fracture and non-fracture groups) were aggregated and 20 cluster centres (k = 20) were found to identify 20 textons representing the ROI type; and 2) for a specific ROI, 10 cluster centres (k = 10) were found for the fracture group and a set of 10 clusters (k = 10) were found for the non-fracture group. The two sets of 10 clusters were then aggregated to form a single dictionary of 20 textons. For both methods, these 20 clusters were Gabor textons for each ROI type. Next, each pixel in the ROI was associated with the textons closest to v_p in the representation space for that ROI type. Each ROI was then represented by the normalised histogram of the textons' occurrences in the ROI. Thus, each ROI in each training image was represented by a vector of length 20.

The 3x3 textons representation was computed similarly. In this case, for each pixel p, the image intensities of the 8 neighbouring pixels were used to form a vector v_p of length 8. The remaining steps followed those for constructing the Gabor textons representation. Thus, the vectors v_p for a single ROI type were collected into an 8-dimensional representation space and the same two K-means clustering methods mentioned above were used to find 20 textons. Pixels were associated with one of these textons based on the closest textons in the representation space to v_p. Each ROI in each image was represented by a vector of length 20 called the 3x3 neighbourhood textons.

In the testing stage, the same procedure was repeated as in the training stage except that the clustering step was omitted. Thus, the vectors v_p were computed for each pixel, but the pixel p was associated with the textons from the training step closest to v_p in the representation space. The Gabor textons and the 3x3 neighbourhood textons were computed analogously to these quantities for the training images. Accordingly, Gabor textons features (gt_1, ..., gt_{20}) and 3x3 neighbourhood textons features (nt_1, ..., nt_{20}) were extracted for a specified ROI for each subject and they were used as input for feature selection.

3.3.4 Selecting ROI with Better Classification Performance

We also compared the classification performance between six ROI: the whole hip, the whole femoral neck, neck slice, Ward's triangle, the narrow neck, the inter-trochanter and the femoral shaft. The whole hip and the whole femoral neck regions outperformed the other regions in the proximal femur (Table 4.3 b in Results Section), and were therefore selected for combinations of two methods described in the following section.

3.4 Feature Selection and Linear Discriminant Analysis

With 12 ASM features (b_1, ..., b_{12}), 57 AAM features (c_1, ..., c_{57}), 24 Gabor filters features (g_1, ..., g_{24}), 20 Gabor texton features (gt_1, ..., gt_{20}) and 20 3x3 neighbourhood textons features (nt_1, ..., nt_{20}), each subject was represented by a feature vector of length 517. Classification based on this many features with only 60 training and 59 testing examples is unreliable and so feature selection was used to reduce the dimensionality of the feature space to avoid overfitting (Section 2.6).

As a rule of thumb, the number of features should be approximately $\log_{10}(N)$ where N is the number of samples available. Accordingly, three features were seen as the most reasonable number of features to select. In addition, all individual features and all combinations of 6 features were compared. Exhaustive search was used three times for each ROI type to reduce the number of features for classification to 1, 3 and 6 features respectively. The method of exhaustive search is not often discussed in the literature because the method scales factorially and so is not practical in many situations. However, here the number of combinations is reasonable and this method is guaranteed to find the best solution. For each ROI type, the parameters for three Fisher linear classifiers were determined using the 1, 3 and 6-dimensional feature spaces. Only training data was used for the feature selection step and fitting the Fisher classifier. The classifiers trained in this way were then applied to the respective ROI types in the testing images.

The procedure described thus far provided estimates of how well the shape and appearance models and texture measures were able to predict fractures. Also of interest is the performance of combining these methods with aBMD and/or T-score. To assess

the combined risk factors, the steps above were repeated but with standard risk factors included in the classification step. To form a comprehensive picture of the relative contributions from these methods, features from these methods were combined in two ways: (1) by including clinical standards (total aBMD and total T-score) in the pool of features prior to feature selection, and (2) by augmenting clinical standards with the set of shape, appearance and texture features after feature selection.

4 RESULTS

In this study, the area under the ROC curve, AUC, was used as the measure of classification performance. Since one, three and six features were considered for comparison, classification performance was measured for optimal feature subsets of size k=1, 3, 6. The proposed model was validated using 2-fold cross validation techniques.

4.1 Baseline Characteristics of Subjects

There were no significant differences between the fracture group and control group with regard to age, neck aBMD, neck T-score, HAL and NSA. As expected, the fracture group had lower total aBMD and total T-score compared to the control group ($p < 0.05$, two tailed t-test) (Table 4.1).

Table 4.1: Baseline characteristics of all subjects (29 fracture cases and 90 control cases). Values shown as mean±SD, and p values are from two tailed t-test.

Methods	Fracture group (n=29)	Control group (n=90)	*p* Values
Age (years)	65.90±2.91	65.64±3.09	0.700
Neck T-score	-1.17±0.81	-0.78±1.33	0.140
Total T-score	-0.70±0.87	-0.08±1.10	0.0065[b]
Total aBMD (g/cm2)	0.88±0.12	0.96±0.15	0.0138[a]
Neck aBMD (g/cm2)	0.74±0.10	0.78±0.21	0.320
HAL (mm)	109.10±9.72	110.76±11.01	0.470
NSA (^{0})	128.93±5.18	128.58±5.86	0.770

[a] $p<0.05$, [b] $P<0.01$; HAL= hip axis length; NSA= neck shaft angle.

4.2 Classification Results for Individual Methods

The word ‘method’ will be used to refer to any one of the strategies for determining risk of fracture. Thus, the standard risk factors, total aBMD, neck aBMD, total T-score and neck T-score, are each viewed as a single method for estimating risk. The statistical methods ASM and AAM as well as the texton-based strategies are also viewed as individual methods. Each method results in one or more output values. In keeping with the language of machine learning, the output values will be referred to as features since they are used as inputs to classification schemes. Thus, total aBMD, neck aBMD, total T-score and neck T-score each yield a single feature each while ASM and AAM and the textons methods yield several features each.

As a single feature, total aBMD and total T-score outperformed the other single features including the best single features selected from ASM and AAM (Table 4.2). Increasing the number of features for ASM and AAM improved the training scores but the testing scores were poor in each case (Table 4.2). The results indicate the existence of similar trends for both folds.

Table 4.2 a: Performance of methods using a single feature and optimal sets of 3 and 6 features in estimating fracture risk for training (fold A of 60 subjects) and testing sets (fold B of 59 subjects). The results are reported as AUC.

Method		Best 1 feature	Best 3 features[b]	Best 6 features[b]
Total aBMD	Training	0.651	-	-
	Testing	0.699[a]	-	-
Neck aBMD	Training	0.629	-	-
	Testing	0.651	-	-
Total Tscore	Training	0.688	-	-
	Testing	0.692[a]	-	-
Neck T-score	Training	0.627	-	-
	Testing	0.654	-	-
ASM	Training	0.610	0.715	0.778
	Testing	0.402	0.475	0.563
AAM	Training	0.660	0.830	0.900
	Testing	0.479	0.417	0.487

[a] The best two testing AUC score; [b] Where available – for ASM and AAM only.

Table 4.2 b: Performance of methods using a single feature and optimal sets of 3 and 6 features in estimating fracture risk for training (fold B of 59 subjects) and testing sets (fold A of 60 subjects). The results are reported as AUC.

Method		Best 1 feature	Best 3 features[b]	Best 6 features[b]
Total aBMD	Training	0.699	-	-
	Testing	0.651[a]	-	-
Neck aBMD	Training	0.651	-	-
	Testing	0.629	-	-
Total Tscore	Training	0.692	-	-
	Testing	0.688 [a]	-	-
Neck T-score	Training	0.654	-	-
	Testing	0.627	-	-
ASM	Training	0.683	0.737	0.779
	Testing	0.379	0.384	0.278
AAM	Training	0.706	0.833	0.919
	Testing	0.487	0.579	0.504

[a] The best two testing AUC score; [b] Where available – for ASM and AAM only.

For 3x3 neighbourhood textons and Gabor textons on individual ROI with 4 different values of *K* (*K*=10, 20, 30, and 40) and *K*-means clustering method 1 (common textons over all classes), the best single feature in terms of AUC for the testing data was a Gabor textons feature measured on the whole femoral neck region, AUC=0.640 for *K*=10 (Table 4.3 a), AUC=0.674 for *K*=20 (Table 4.3 b),AUC=0.670 for *K*=30 (Table 4.3 c), AUC=0.665 for *K*=40 (Table 4.3 d) respectively. Overall, the methods (3x3 neighbourhood textons and Gabor textons) *K*=20 (20 textons) provided the best discriminating performance compared to the other three values of *K* (Table 4.3 b).

For 3x3 neighbourhood textons and Gabor textons on individual ROI with *K*-means clustering method 2 and *K*=10 (20 textons), the best single feature in terms of

AUC for the testing data was also a Gabor textons feature measured on the whole femoral neck region, AUC=0.583 (Table 4.3 e). The results also revealed that *K*-means clustering method 1 (common textons over all classes) outperformed *K*-means clustering method 2 (different textons per class) (Table 4.3 b, Table 4.3 e).

Table 4.3 a: Performance of 3x3 neighbourhood textons method using a single feature and optimal sets of 3 and 6 features on the various regions of interest for training (fold A) and testing sets (fold B) with *K*-means clustering method 1 (common textons over all classes) and *K* = 10. The columns labelled 1, 3 and 6 refer to the number of features. The results are reported as AUC.

		3x3 neighbourhood textons			Gabor textons		
ROI		1	3	6	1	3	6
Whole hip	Training	0.657	0.744	0.790	0.645	0.667	0.678
	Testing	0.383	0.336	0.364	0.367	0.454	0.392
Whole femoral Neck	Training	0.579	0.664	0.703	0.707	0.764	0.779
	Testing	0.476	0.349	0.396	0.640[a]	0.475	0.426
Neck Slice	Training	0.555	0.651	0.727	0.582	0.656	0.703
	Testing	0.415	0.350	0.244	0.393	0.457	0.337
Ward's Triangle	Training	0.585	0.666	0.737	0.572	0.659	0.687
	Testing	0.562	0.397	0.422	0.415	0.496	0.475
Narrow Neck	Training	0.559	0.687	0.777	0.588	0.673	0.687
	Testing	0.458	0.202	0.316	0.397	0.336	0.333
Inter-trochanter	Training	0.657	0.680	0.704	0.578	0.639	0.633
	Testing	0.436	0.394	0.395	0.473	0.453	0.523
Femoral shaft	Training	0.592	0.667	0.686	0.611	0.667	0.738
	Testing	0.526	0.380	0.444	0.353	0.458	0.524

[a] The best testing AUC score.

Table 4.3 b: Performance of 3x3 neighbourhood textons method using a single feature and optimal sets of 3 and 6 features on the various regions of interest for training (fold A) and testing sets (fold B) with *K*-means clustering method 1 (common textons over all classes) and $K = 20$. The columns labelled 1, 3 and 6 refer to the number of features. The results are reported as AUC.

ROI		3x3 neighbourhood textons			Gabor textons		
		1	3	6	1	3	6
Whole hip	Training	0.698	0.794	0.873	0.626	0.760	0.821
	Testing	0.429	0.433	0.398	0.501	0.394	0.455
Whole femoral Neck	Training	0.624	0.743	0.859	0.719	0.777	0.808
	Testing	0.504	0.413	0.388	0.674[a]	0.478	0.515
Neck Slice	Training	0.607	0.714	0.872	0.601	0.727	0.785
	Testing	0.402	0.498	0.324	0.387	0.437	0.530
Ward's Triangle	Training	0.607	0.711	0.804	0.616	0.719	0.810
	Testing	0.563	0.481	0.420	0.406	0.462	0.468
Narrow Neck	Training	0.562	0.708	0.801	0.599	0.699	0.756
	Testing	0.455	0.280	0.297	0.391	0.409	0.386
Inter-trochanter	Training	0.678	0.780	0.838	0.587	0.675	0.731
	Testing	0.419	0.513	0.417	0.466	0.433	0.591
Femoral shaft	Training	0.618	0.741	0.807	0.591	0.659	0.707
	Testing	0.558	0.424	0.422	0.480	0.629	0.537

[a] The best testing AUC score.

Table 4.3 c: Performance of 3x3 neighbourhood textons method using a single feature and optimal sets of 3 and 6 features on the various regions of interest for training (fold A) and testing sets (fold B) with *K*-means clustering method 1 (common textons over all classes) and $K = 30$. The columns labelled 1, 3 and 6 refer to the number of features. The results are reported as AUC.

ROI		3x3 neighbourhood textons			Gabor textons		
		1	3	6	1	3	6
Whole hip	Training	0.684	0.787	0.879	0.681	0.773	0.854
	Testing	0.315	0.425	0.394	0.362	0.367	0.376
Whole femoral Neck	Training	0.667	0.833	0.901	0.690	0.764	0.828
	Testing	0.598	0.463	0.514	0.670[a]	0.590	0.455
Neck Slice	Training	0.616	0.730	0.841	0.628	0.713	0.781
	Testing	0.385	0.361	0.407	0.448	0.441	0.481
Ward's Triangle	Training	0.607	0.796	0.868	0.593	0.721	0.810
	Testing	0.54	0.388	0.452	0.425	0.518	0.292
Narrow Neck	Training	0.630	0.757	0.849	0.613	0.704	0.787
	Testing	0.548	0.472	0.356	0.417	0.390	0.449
Inter-trochanter	Training	0.684	0.797	0.846	0.626	0.742	0.790
	Testing	0.453	0.490	0.506	0.435	0.339	0.566
Femoral shaft	Training	0.641	0.781	0.896	0.607	0.735	0.784
	Testing	0.586	0.412	0.263	0.482	0.624	0.611

[a] The best testing AUC score.

Table 4.3 d: Performance of 3x3 neighbourhood textons method using a single feature and optimal sets of 3 and 6 features on the various regions of interest for training (fold A) and testing sets (fold B) with *K*-means clustering method 1 (common textons over all classes) and $K = 40$. The columns labelled 1, 3 and 6 refer to the number of features. The results are reported as AUC.

ROI		3x3 neighbourhood textons			Gabor textons		
		1	3	6	1	3	6
Whole hip	Training	0.709	0.830	0.893	0.663	0.789	0.877
	Testing	0.413	0.448	0.380	0.371	0.382	0.382
Whole femoral Neck	Training	0.664	0.850	0.937	0.716	0.823	0.884
	Testing	0.599	0.560	0.503	0.665[a]	0.545	0.644
Neck Slice	Training	0.631	0.757	0.868	0.629	0.744	0.829
	Testing	0.412	0.335	0.264	0.428	0.382	0.417
Ward's Triangle	Training	0.621	0.790	0.888	0.583	0.717	0.836
	Testing	0.489	0.522	0.433	0.389	0.352	0.325
Narrow Neck	Training	0.631	0.774	0.859	0.581	0.717	0.790
	Testing	0.506	0.428	0.362	0.460	0.375	0.625
Inter-trochanter	Training	0.688	0.785	0.852	0.644	0.787	0.891
	Testing	0.440	0.499	0.450	0.433	0.506	0.483
Femoral shaft	Training	0.661	0.786	0.890	0.601	0.713	0.779
	Testing	0.573	0.390	0.426	0.522	0.502	0.489

[a] The best testing AUC score.

Table 4.3 e: Performance of 3x3 neighbourhood textons method using a single feature and optimal sets of 3 and 6 features on the various regions of interest for training (fold A) and testing sets (fold B) with *K*-means clustering method 2 (different textons per class) and *K* = 10. The columns labelled 1, 3 and 6 refer to the number of features. The results are reported as AUC.

ROI		3x3 neighbourhood textons			Gabor textons		
		1	3	6	1	3	6
Whole hip	Training	0.719	0.801	0.856	0.636	0.753	0.824
	Testing	0.355	0.382	0.333	0.394	0.374	0.392
Whole femoral Neck	Training	0.581	0.735	0.801	0.716	0.808	0.887
	Testing	0.434	0.444	0.448	0.583[a]	0.541	0.499
Neck Slice	Training	0.621	0.734	0.823	0.638	0.740	0.820
	Testing	0.395	0.303	0.359	0.462	0.49	0.59[a]
Ward's Triangle	Training	0.612	0.740	0.842	0.575	0.801	0.870
	Testing	0.559	0.416	0.407	0.424	0.401	0.352
Narrow Neck	Training	0.595	0.741	0.845	0.597	0.689	0.763
	Testing	0.375	0.255	0.229	0.386	0.41	0.539
Inter-trochanter	Training	0.716	0.774	0.825	0.641	0.747	0.853
	Testing	0.439	0.475	0.468	0.423	0.445	0.475
Femoral shaft	Training	0.653	0.762	0.844	0.648	0.756	0.826
	Testing	0.469	0.351	0.387	0.404	0.546	0.504

[a] The best testing AUC score.

For Gabor filters, 3x3 neighbourhood textons and Gabor textons on individual ROI, the best single feature in terms of AUC for the testing data was a Gabor filter feature measured on the whole femoral neck region (AUC=0.700 in Table 4.4 a). For the best

combination of three features, the best performance on the testing data was also provided by the method of Gabor filter applied to the whole femoral neck region (AUC =0.674 in Table 4.4 b). Similarly, for the best six features the highest AUC score for testing data was 0.646 from Gabor filters on the whole femoral neck (Table 4.4 b). The results indicate that similar trends exist for both folds.

Table 4.4 a: Performance of Gabor filters, 3x3 neighbourhood textons and Gabor textons using a single feature and optimal sets of 3 and 6 features on the various regions of interest for training (fold A) and testing sets (fold B). The results are reported as AUC. For each method (for example the whole hip) the top row reports the training scores and the bottom row reports the testing scores. The results are reported as AUC.

	Gabor filters			3x3 neighbourhood textons[b]			Gabor textons[b]		
ROI	1	3	6	1	3	6	1	3	6
Whole hip	0.646	0765	0.841	0.698	0.794	0.873	0.626	0.760	0.821
	0.394	0.426	0.429	0.429	0.433	0.398	0.501	0.394	0.455
Whole femoral Neck	0.736	0.770	0.871	0.624	0.743	0.859	0.719	0.777	0.808
	0.700[a]	0.588	0.603	0.504	0.413	0.388	0.674	0.478	0.515
Neck Slice	0.648	0.721	0.820	0.607	0.714	0.872	0.601	0.727	0.785
	0.358	0.401	0.436	0.402	0.498	0.324	0.387	0.437	0.530
Ward's Triangle	0.621	0.726	0.807	0.607	0.711	0.804	0.616	0.719	0.810
	0.352	0.381	0.439	0.563	0.481	0.420	0.406	0.462	0.468
Narrow Neck	0.647	0.719	0.792	0.562	0.708	0.801	0.599	0.699	0.756
	0.368	0.463	0.391	0.455	0.280	0.297	0.391	0.409	0.386
Inter-trochanter	0.619	0.819	0.858	0.678	0.780	0.838	0.587	0.675	0.731
	0.595	0.545	0.521	0.419	0.513	0.417	0.466	0.433	0.591
Femoral shaft	0.596	0.690	0.745	0.618	0.741	0.807	0.591	0.659	0.707
	0.437	0.464	0.504	0.558	0.424	0.422	0.480	0.629	0.537

[a] The best testing AUC score. [b] *K*-means clustering method 1 (common textons over all classes) with *K*=20 applied.

Table 4.4 b: Exactly the same as Table 4.4 a except that fold B was used for training sets and fold A was used for testing sets. For each method (for example the whole hip) the top row reports the training scores and the bottom row reports the testing scores.

	Gabor filters			3x3 neighbourhood textons[b]			Gabor textons[b]		
ROI	1	3	6	1	3	6	1	3	6
Whole hip	0.648	0.8030	0.860	0.677	0.807	0.857	0.649	0.750	0.813
	0.530	0.502	0.590	0.521	0.454	0.436	0.461	0.547	0.504
Whole femoral Neck	0.723	0.807	0.810	0.677	0.814	0.858	0.733	0.837	0.889
	0.596	0.674[a]	0.646	0.499	0.487	0.478	0.499	0.585	0.550
Neck Slice	0.685	0.782	0.841	0.720	0.871	0.921	0.699	0.795	0.839
	0.494	0.503	0.542	0.389	0.418	0.350	0.484	0.521	0.418
Ward's Triangle	0.669	0.792	0.841	0.611	0.706	0.802	0.657	0.765	0.822
	0.481	0.445	0.388	0.424	0.486	0.501	0.530	0.433	0.516
Narrow Neck	0.723	0.782	0.813	0.740	0.859	0.904	0.683	0.779	0.854
	0.489	0.502	0.469	0.404	0.401	0.313	0.501	0.453	0.455
Inter-trochanter	0.689	0.785	0.839	0.660	0.844	0.913	0.633	0.800	0.850
	0.496	0.508	0.599	0.381	0.531	0.553	0.448	0.449	0.470
Femoral shaft	0.724	0.771	0.836	0.651	0.784	0.860	0.720	0.848	0.882
	0.516	0.472	0.416	0.510	0.559	0.494	0.504	0.404	0.443

[a] The best testing AUC score. [b] *K*-means clustering method 1 (common textons over all classes) with *K*=20 applied.

4.3 Discriminant Analysis Using Combinations of Two Methods

The combination of total T-score and AAM demonstrated higher discriminative ability for fracture than other combinations at the training stage (AUC=0.907 in 6 features in Table 4.5 a, Table 4.5 b; AUC=0.940 in 6 features in Table 4.5 c, Table 4.5 d). The best predictive capacity at the testing stage was obtained when total T-score was combined with Gabor filters on the whole femoral neck (AUC=0.700 in 1 feature in Table 4.5 a; AUC=0.787 in 6 features in Table 4.5 b; AUC=0.711 in 6 features in Table 4.5 c, Table 4.5 d).

Table 4.5 a: Performance of the combinations of the methods proposed (ASM, AAM, Gabor filters or textons) and the standard methods (total aBMD or T-score) using a single feature and optimal sets of 3 and 6 features in estimating fracture risk for a training set (fold A) and testing set (fold B). The features were gathered from different methods, then optimal combinations were selected. The results are reported as AUC.

Methods		Total aBMD			Total T-score		
		1	3	6	1	3	6
ASM	Training	**0.651**	0.732	0.787	**0.688**	0.770	0.809
	Testing	**0.699**	0.591	0.592	**0.692**	0.633	0.598
AAM	Training	0.660	0.830	0.902	**0.688**	0.841	0.907
	Testing	0.480	0.417	0.524	**0.692**	0.558	0.492
Gabor filters on whole hip	Training	**0.651**	0.765	0.847	**0.688**	0.765	0.844
	Testing	**0.699**	0.426	0.524	**0.692**	0.426	0.517
Gabor filters on whole femoral neck	Training	0.736	0.770	0.871	0.736	0.770	0.871
	Testing	0.700	0.588	0.603	0.700[a]	0.588	0.603
3x3 neighbourhood textons on whole hip	Training	0.698	0.801	0.880	0.698	0.808	0.884
	Testing	0.429	0.600	0.506	0.429	0.570	0.571
3x3 neighbourhood textons on whole femoral neck	Training	**0.651**	0.759	0.864	**0.688**	0.797	0.870
	Testing	**0.699**	0.674	0.606	**0.692**	0.646	0.628
Gabor textons on whole hip	Training	**0.651**	0.760	0.838	**0.688**	0.760	0.847
	Testing	**0.699**	0.394	0.537	**0.692**	0.394	0.527
Gabor textons on whole femoral neck	Training	0.719	0.783	0.816	0.719	0.796	0.827
	Testing	0.674	0.514	0.583	0.674	0.513	0.541

[a] The best testing AUC score; Bold indicates that the best feature was either total aBMD or the total T-score (applies to single features only).

Table 4.5 b: Performance of the combinations of the methods proposed (ASM, AAM, Gabor filters or textons) and the standard methods (total aBMD or T-score) using a single feature and optimal sets of 3 and 6 features in estimating fracture risk for a training set (fold A) and testing set (fold B). The optimal combinations were selected within each method, then these were combined. The results are reported as AUC.

Methods		Total aBMD			Total T-score		
		1	3	6	1	3	6
ASM	Training	**0.651**	0.736	0.787	**0.688**	0.770	0.809
	Testing	**0.699**	0.591	0.592	**0.692**	0.633	0.598
AAM	Training	0.660	0.819	0.901	**0.688**	0.841	0.907
	Testing	0.480	0.518	0.524	**0.692**	0.558	0.492
Gabor filters on whole hip	Training	**0.651**	0.745	0.847	**0.688**	0.752	0.844
	Testing	**0.699**	0.483	0.524	**0.692**	0.518	0.517
Gabor filters on whole femoral neck	Training	0.736	0.729	0.862	0.736	0.751	0.870
	Testing	0.700	0.507	0.687	0.700	0.575	0.787[a]
3x3 neighbourhood textons on whole hip	Training	0.698	0.801	0.880	0.698	0.808	0.884
	Testing	0.429	0.593	0.506	0.429	0.570	0.571
3x3 neighbourhood textons on whole femoral neck	Training	**0.651**	0.759	0.864	**0.688**	0.797	0.870
	Testing	**0.699**	0.675	0.606	**0.692**	0.646	0.628
Gabor textons on whole hip	Training	**0.651**	0.740	0.838	**0.688**	0.756	0.847
	Testing	**0.699**	0.522	0.537	**0.692**	0.717	0.527
Gabor textons on whole femoral neck	Training	0.719	0.783	0.816	0.719	0.796	0.827
	Testing	0.674	0.514	0.583	0.674	0.513	0.541

[a] The best testing AUC score; **Bold** indicates that the best feature was either total aBMD or the total T-score (applies to single features only).

Table 4.5 c: Performance of the combinations of the methods proposed (ASM, AAM, Gabor or textons) and the standard methods (total aBMD or T-score) using a single feature and optimal sets of 3 and 6 features in estimating fracture risk for a training set (fold B) and testing set (fold A). The features were gathered from different methods, then optimal combinations were selected. The results are reported as AUC.

Methods		Total aBMD			Total T-score		
		1	3	6	1	3	6
ASM	Training	**0.699**	0.779	0.838	**0.692**	0.750	0.829
	Testing	**0.651**	0.568	0.600	**0.688**	0.535	0.639
AAM	Training	0.706	0.839	0.928	0.706	0.844	0.940
	Testing	0.487	0.578	0.525	0.487	0.538	0.533
Gabor filters on whole hip	Training	**0.699**	0.803	0.860	**0.692**	0.803	0.860
	Testing	**0.651**	0.502	0.590	**0.688**	0.502	0.590
Gabor filters on whole femoral neck	Training	0.723	0.807	0.825	0.723	0.807	0.840
	Testing	0.596	0.674	0.674	0.596	0.674	0.711[a]
3x3 neighbourhood textons on whole hip	Training	**0.699**	0.813	0.892	**0.692**	0.870	0.892
	Testing	**0.651**	0.542	0.528	**0.688**	0.600	0.528
3x3 neighbourhood textons on whole femoral neck	Training	**0.699**	0.817	0.872	**0.692**	0.817	0.870
	Testing	**0.651**	0.510	0.559	**0.688**	0.510	0.554
Gabor textons on whole hip	Training	**0.699**	0.828	0.897	**0.692**	0.828	0.889
	Testing	**0.651**	0.567	0.589	**0.688**	0.567	0.612
Gabor textons on whole femoral neck	Training	0.733	0.842	0.881	0.733	0.841	0.876
	Testing	0.499	0.671	0.629	0.499	0.684	0.655

[a] The best testing AUC score; **Bold** indicates that the best feature was either total aBMD or the total T-score (applies to single features only).

Table 4.5 d: Performance of the combinations of the methods proposed (ASM, AAM, Gabor or textons) and the standard methods (total aBMD or T-score) using a single feature and optimal sets of 3 and 6 features in estimating fracture risk for a training set (fold B) and testing set (fold A). The optimal combinations were selected within each method, then these were combined. The results are reported as AUC.

Methods		Total aBMD			Total T-score		
		1	3	6	1	3	6
ASM	Training	**0.699**	0.779	0.838	**0.692**	0.752	0.829
	Testing	**0.651**	0.568	0.600	**0.688**	0.535	0.639
AAM	Training	0.706	0.839	0.928	0.706	0.844	0.940
	Testing	0.487	0.578	0.525	0.487	0.538	0.533
Gabor filters on whole hip	Training	**0.699**	0.772	0.846	**0.692**	0.795	0.855
	Testing	**0.651**	0.527	0.562	**0.688**	0.592	0.617
Gabor filters on whole femoral neck	Training	0.723	0.794	0.825	0.723	0.802	0.840
	Testing	0.596	0.624	0.674	0.596	0.657	0.711[a]
3x3 neighbourhood textons on whole hip	Training	**0.699**	0.813	0.887	**0.692**	0.840	0.877
	Testing	**0.651**	0.542	0.641	**0.688**	0.614	0.636
3x3 neighbourhood textons on whole femoral neck	Training	**0.699**	0.850	0.872	**0.692**	0.799	0.870
	Testing	**0.651**	0.512	0.559	**0.688**	0.621	0.554
Gabor textons on whole hip	Training	**0.699**	0.828	0.897	**0.692**	0.828	0.889
	Testing	**0.651**	0.567	0.589	**0.688**	0.567	0.612
Gabor textons on whole femoral neck	Training	0.733	0.842	0.881	0.733	0.841	0.876
	Testing	0.499	0.671	0.629	0.499	0.684	0.655

[a] The best testing AUC score; **Bold** indicates that the best feature was either total aBMD or the total T-score (applies to single features only).

4.4 Classification Performance Using Combinations of Several Methods

Feature combinations from all methods were considered next. With a total of 517 individual features from all the methods, there are 22,897,930 possible combinations of three features. This number was too large to conduct an exhaustive search for the best combination, and so the total number of features was reduced by considering only the top six features from each method according to individual performance. In addition, poor performing methods such as 3x3 neighbourhood textons were excluded. This resulted in a pool of 20 features: 6 features from ASM, 6 features from AAM, 6 features from Gabor filters on the whole hip plus the standard measures of aBMD and total T-score. Exhaustive searching was used on the training set, and we found that a perfect training performance (AUC=1.000 in Table 4.6 a) and near perfect performance (AUC=0.960 in Table 4.6 b) was achieved with a combination of 11 features. In addition, all individual features were considered and all combinations of 6 features were considered for comparison. The resulting features and classifiers were used to estimate performance using the testing set (Table 4.6). The results indicate that similar trends exist for both folds.

Table 4.6 a: Performance of the combinations of ASM (6 features), AAM (6 features), Gabor filters on whole femoral neck (6 features), total aBMD and total T-Score (2 features) using a single feature and optimal sets of 6 and 11 features in estimating fracture risk for training (fold A) and testing sets (fold B). The results are reported as AUC.

Number of features	1	3	6	11
Training	0.732	0.850	0.916	1.000
Testing	0.705[a]	0.571	0.603	0.537

[a] The best testing AUC score.

Table 4.6 b: Exactly the same as Table 4.6 a except that fold B of 59 subjects was used for the training set and fold A of 60 subjects was used for the testing set. The results are reported as AUC.

Number of features	1	3	6	11
Training	0.710	0.821	0.927	0.960
Testing	0.658[a]	0.550	0.553	0.539

[a] The best testing AUC score.

4.5 Comparison between ASM and AAM and with Previous Studies

Additional comparisons were made between ASM and AAM and with previous studies [72, 103, 111]. Comparisons with other studies were made in terms of study parameters (Table 4.7 a), performance on training data (Table 4.7 b) and performance on testing data (Table 4.7 c).

Table 4.7 a: Comparison of study parameters.

	Gregory [72]	Goodyear [111]	B.-L. [103]	This study
Database	26 fracture 24 non-fracture (all females)	182 fracture 364 non-fracture (all females)	168 fracture 231 non-fracture (all females)	29 fracture 90 non-fracture (both genders)
Fracture types	Hip fractures	Hip fractures	Hip fractures	All fractures
Point-point error analysis	Yes	No	No	No
landmarks	29 points head included	72 points head & pelvis	60 points head included	44 points head excluded

Table 4.7 b: Comparison of training AUC.

Method	Gregory [72]	Goodyear [111]	B.-L. [103]	This study
ASM	0.81 (4 features)	0.57 (1 feature)	0.81 (10 features)	0.78 (6 features)
AAM	n/a	0.57 (1 feature)	n/a	0.90 (6 features)
Ward aBMD	0.95 (1 feature)	n/a	n/a	n/a
Total aBMD	0.63(1 feature)	0.62 (1 feature)	n/a	0.65 (1 feature)
Neck aBMD	0.79 (1 feature)	n/a	0.68	0.63 (1 feature)
ASM & Ward aBMD	0.96 (5 features)	n/a	n/a	n/a
ASM & Neck aBMD	0.89 (5 features)	n/a	0.84(11 features)	n/a
ASM & Total aBMD	n/a	0.79 (6 features)	n/a	0.79 (6 features)
Multi-combination	n/a	0.65 (3 features) ASM & AAM & Total aBMD	n/a	1.00 (11 features) Gabor & ASM & AAM & Total aBMD

Table 4.7 c: Results for testing dataset in terms of AUC. The other studies [72, 103, 111] did not include testing results, and the methods used are as Table 4.7 b.

	Gregory [72]	Goodyear [111]	B.-L. [103]	This study
	n/a	n/a	n/a	0.79[a]

[a] AUC score using 6 features selected from Gabor filters on whole femoral neck & total T-score

4.6 Association of Principal Modes with Bone Fracture

The 60 shape and appearance principle modes explaining the shape and appearance variation of proximal femur images were compared (Figure 4.1). The first six principal modes of shape and appearance explained 95% of shape variation and 27% of appearance variation, respectively (Figure 4.1). In addition, the best combination of 6 shape principal modes were modes 1, 3, 6, 7, 9, and 12 with a training AUC score of 0.778, while the best combination of 6 appearance principal modes were modes 7, 37, 39, 43, 47, and 48 with a training AUC score of 0.900 (Table 4.8 a). Principal mode 10, which explained 0.49% of the variance in the shape of proximal femur, had the best training AUC score (AUC=0.61) (Table 4.8 b). Additional comparisons were made for each of the 12 principal modes in terms of visible variance in the shape of the proximal femur (Figure 4.2).

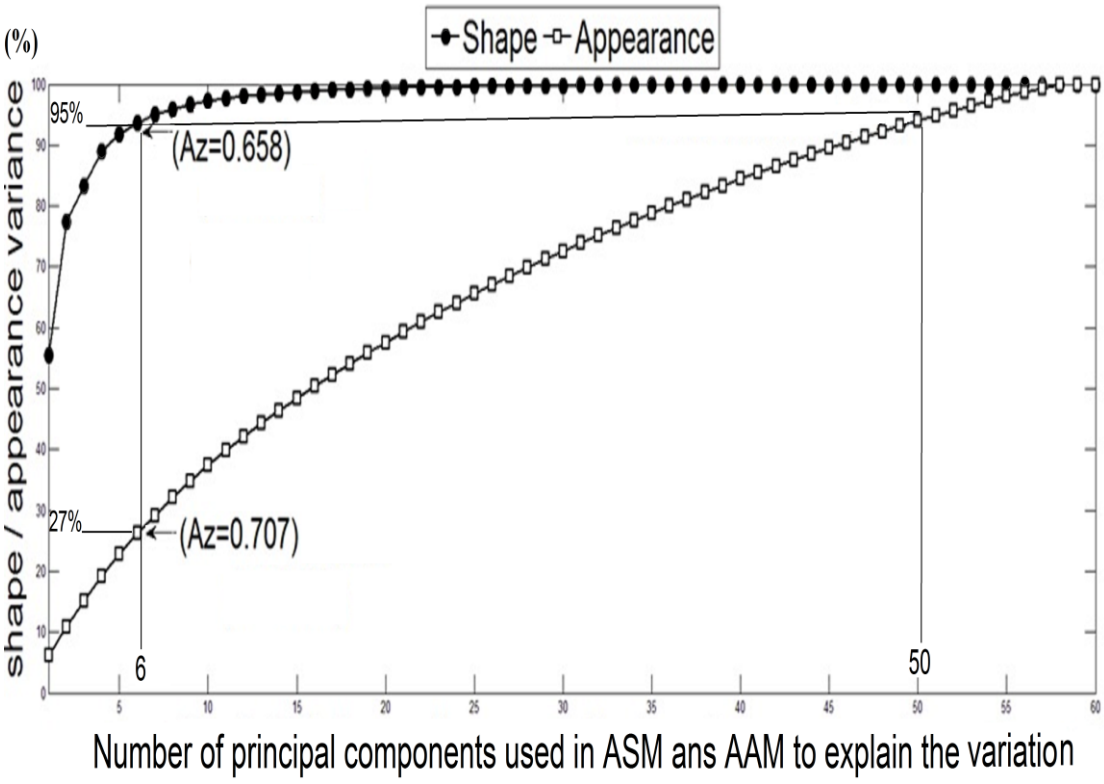


Figure 4.1: Cumulative percentage of proximal femur shape and appearance variation explained by principal modes in ASM and AAM respectively.

Table 4.8 a: The best combination of 6 principal modes selected by exhaustive search for the ASM and AAM respectively.

ASM	AAM
Principal mode 1	Principal mode 7
Principal mode 3	Principal mode 37
Principal mode 6	Principal mode 39
Principal mode 7	Principal mode 43
Principal mode 9	Principal mode 47
Principal mode 12	Principal mode 48

Table 4.8 b: Percentage of variance in the shape of the proximal femur explained by each principal mode of variation, and the corresponding training AUC score. The principal mode marked with * was the one selected with the best training AUC score.

Principal mode	% of variance explained	Training AUC score
Principal mode 1	56.57	0.55
Principal mode 2	22.21	0.48
Principal mode 3	5.84	0.54
Principal mode 4	5.72	0.57
Principal mode 5	2.85	0.55
Principal mode 6	1.79	0.59
Principal mode 7	1.22	0.60
Principal mode 8	0.83	0.58
Principal mode 9	0.65	0.54
Principal mode 10*	0.49	0.61
Principal mode 11	0.44	0.60
Principal mode 12	0.39	0.59
Total	99.00	

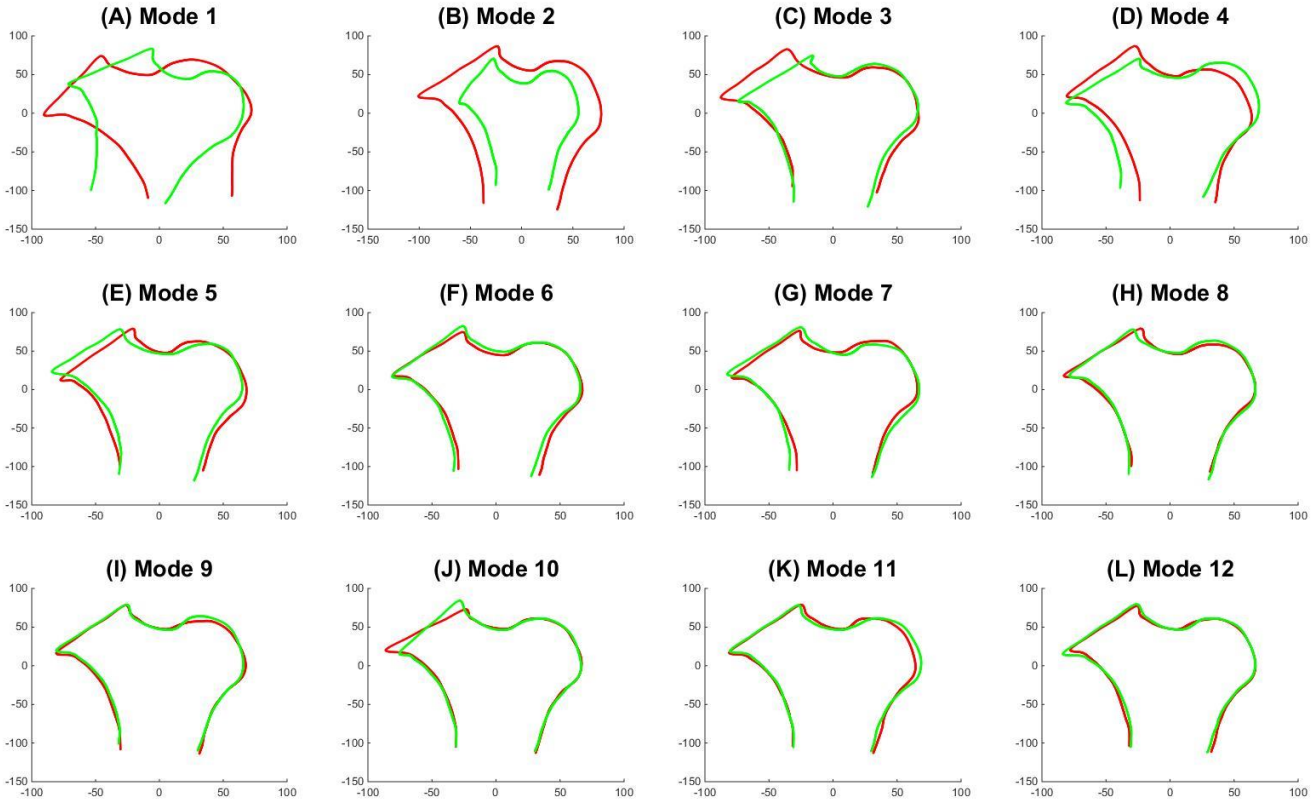


Figure 4.2: Visible variance in the shape of the proximal femur in each principal mode. Each figure shows the +2 SD (red) shapes and the -2 SD (green) shapes for each of the 12 principal modes. Principal mode 10 was the one selected with the best AUC score.

5 DISCUSSION AND CONCLUSIONS

5.1 Discussion

Texture information derived from Gabor filters in combination with the total T-score provided a better estimate of fracture status than the standard measures of aBMD (AUC=0.651 and 0.699) or total T-score (AUC=0.688 and 0.692) alone. In particular, 6 features selected from Gabor filters computed over the whole femoral neck and total T-score gave an AUC more than 10% higher than aBMD alone, total T-score alone or combinations of textons features alone. The combination of Gabor filters and total T-score also outperformed risk estimates based on ASM and AAM, which, on their own, performed very poorly (Table 4.2, Table 4.5). In addition, risk estimates based on the femoral neck were more reliable than estimates based on other regions of interest tested (Table 4.3, Table 4.4). These conclusions are based on estimates of risk obtained by testing the features and classifier parameters found during the training stage on an independent set of images. Thus, these AUC scores provide indications of how well the methods described here are able to estimate risk in new cases.

In this study, two strategies for combinations of two methods were considered (Table 4.5). The first strategy was to gather features from different methods, and then select optimal combinations of one, three and six features (Table 4.5 a, Table 4.5 c). The second strategy was to select optimal combinations, then gather features from different methods (Table 4.5 b, Table 4.5 d). The experimental results showed that there is not much difference in classification performance between the two strategies for combinations of two methods. This could be explained by the fact that the features from the current clinical standard measures (total aBMD and total T-score) still play such an important role in discriminating the risk of bone fractures that the classification performance was not affected no matter what the strategies are for combinations of two methods. However, the results also indicate that texture-based analysis applied on the whole femoral neck has the potential to complement the current clinical standard measures.

While reporting performance results on an independent testing set is standard in image analysis generally, this has not been the practice in the field of fracture risk assessment. Thus, while the scores reported for testing data provide the best indication

of how the methods presented here will perform in practice, the training scores must be used to compare results with other studies. We compared our findings with three other studies [72, 103, 111] (Table 4.7). Differences must be interpreted with care due to disparity in the type of fractures considered, the landmarks used and the regions over which aBMD was computed. In this study, measurements taken from DXA images of the femur were used to estimate the risk of all fractures (femur, vertebrae and wrist), not just of the femur, whereas other studies, some of those mentioned in Section 2.3 (Table 2.1), specifically focussed on the risk of fracture of the femur and provided only training results. Nevertheless, several observations may be made.

First, the results for ASM alone are consistent over all the studies except Goodyear et al [111], which differs in that the pelvis was included in the shape model. Including the pelvis increased the total information available from an information theory perspective but whether this additional information is relevant to fracture risk is not clear. Including the pelvis also introduces more opportunity for error, especially since the outline of the pelvis is not always clear in DXA images. Since better results were obtained in this study where the pelvis was not included, it seems that, in balance, including the pelvis is detrimental for estimating fracture risk. This may also explain the superior training results for the AAM in this study (AUC=0.90) compared to previous use of the AAM in Goodyear et al [111] (AUC=0.57), but the discrepancy may also be due to the final number of features used in classification. In Goodyear et al [111], only one AAM feature was used whereas six features were used in this study (Table 4.7 b).

Second, total aBMD was consistent over the studies for which this information was available and this supports the tacit assumption that the images used in all these studies are comparable in quality. In the study performed by Gregory et al, better prediction of hip fractures was achieved using neck and Ward's triangle aBMD [72]. Improved performance for predicting all fractures using neck aBMD was not supported in this study and Ward's triangle aBMD was not available (Table 4.7 b).

Third, in all studies, combining a form of aBMD with ASM resulted in better performance. Performance using total aBMD discovered in this study was about the same as reported in the study by Goodyear et al [111], but, as noted above, the performance of ASM was much lower in the study by Goodyear et al [111] than in this

study. However, the combination of ASM and total aBMD was the same in these studies. These results indicate that, if ASM is used effectively (pelvis not included, for example), then total aBMD does not contribute extra information, but if ASM is used less efficiently, then aBMD may compensate (Table 4.7 b).

Fourth, these studies do not necessarily support the notion that increasing the number of features improves performance. Notably, combining AAM, ASM and total aBMD in the study by Goodyear et al led to poorer results than combining just ASM and total aBMD [111] (Table 4.7 b).

The comparisons made with other studies were based on training results instead of testing results (Table 4.7 b). This was necessary because only training scores were presented in the cited papers. In the study performed by Baker-Lepain et al [103], a bootstrap procedure was used to determine that the effects were different from what could be expected by chance alone, but that does not address the robustness of the reported performance scores [103]. The parameters determined during the development stage were not tested on an independent data set. In the study performed by Gregory et al, models were trained and tested on independent data but final feature selection and classification was only performed for the resulting best model [72]. Thus, the selection of the best model was validated but feature selection and classification were not tested on an independent dataset and so results must be viewed as training results. Classification results based on training only may not be reliable, especially if feature selection is used in the training process. This is well documented in the literature [109] and there were many examples in this study. By selecting six of the 3x3 neighbourhood texton features on the whole hip ROI, an AUC score of 0.873 was found during the training stage but when the resulting classifier was applied to the set of test images, the AUC score reduced to 0.398 (Table 4.3 b). Perfect classification was achieved by combining 11 ASM, AAM and Gabor filters features during training but performance dropped to AUC=0.537 in testing (Table 4.6 a). For this reason, only properly validated results should be taken as indicators of actual performance. Thus, only the testing results reported in this study should be taken as indicative of true performance.

The ASM was designed to consider information along the boundary of target objects to represent the variation of shape only, while the AAM takes advantage of

shape as well as the intensity information of the region within a target object to represent variation in appearance. The principal modes for shape and appearance were computed using PCA on output from the ASM and AAM models respectively. Results showed that 95% of shape variations could be explained by the first six principal modes of the shape (Figure 4.1). In contrast, the first six principal modes of appearance described a mere 27% of the variation in appearance. The first 50 principal modes of appearance were needed to explain 95% of the variation in appearance, in other words, the AAM is much slower in describing the shape and bone density. The fact that 50 principal modes were needed in the AAM model when the size of the training set was approximately 60 shows that this isn't great. This has been seen before. For example, Bryan et al, who used a 3D AAM model, found similar performance for an AAM [112, 113]. We also compared the classification performance between the ASM and the AAM using the first six principal modes of variation, and found that the AAM produced better discrimination (AUC=0.707) than the ASM (AUC=0.658) (Figure 4.1). In addition, we compared the best combinations of 6 principal modes selected by exhaustive search for the ASM and the AAM based on AUC score, and found that the features suitable for representation of target subjects may not always be the features selected by exhaustive search for classification between the two groups (Table 4.8 a).

We also compared the percentage of variance in the proximal femur shape explained by each principal mode of variation and the corresponding training AUC score and found that the features explaining more variance in the proximal femur shape may not always achieve higher classification performance (Table 4.8 b). For instance, principal mode 1 and principal mode 2 explained a much higher percentage of shape variance (56.57% and 22.21% respectively) than others, but achieved lower classification performance (AUC=0.550 and 0.484 respectively). On the contrary, the principal modes with less shape variance, such as principal mode 10 and principal mode 11, had higher classification performance (AUC= 0.61 and 0.60 respectively).

These illustrate the fact that, while PCA finds features (called principal modes) that are optimal for representing information, these features are not necessarily optimal for classification [109]. We used PCA as part of AAM and ASM because these methods were historically designed to use PCA. Hence, we also used PCA in order to extend and compare to previous literature on AAM and ASM. The problem with PCA

is not that it considers linear combinations, but that it chooses linear combinations to maximally represent the full data rather than linear combinations that separate two subclasses of the data. While PCA is not necessarily optimal for classification, the method is reasonable in this context since the amount of redundancy in the original 44 shape parameters (points on the boundary of the femur) is large. The principal modes found by PCA are decorrelated and this property (rather than optimal representation) facilitates feature selection. Thus, PCA was used as described in the literature for ASM and AAM to reduce the number of shape features (and appearance features) before they were included in the pool of features. However, once these PCA features were obtained, and combined with features from other sources (texture, etc.), further feature selection was based on exhaustive search rather than PCA.

We examined the shape features that provided good classification by generating examples of femurs according to the variation of individual principal modes. We used the ASM method to examine the variation in shape of the proximal femur and the association of hip shape with fracture risk. This led to the identification of distinct hip shapes or principal modes that were significantly associated with higher fracture risk (Table 4.8 a and Figure 4.2). With the ASM model, the top 12 principal modes (out of 60) explained 99% of the variance in the shape of the proximal femur (Table 4.8 b).

Principal mode 10, which explained 0.49% of the variance in the shape of the proximal femur, was associated with the change in angle of the femoral head-neck rotation movement in relation to the trochanters and shaft (Figure 4.2J). Principal mode 11 explained 0.44% of the variance in the proximal femur shape, reflecting the relative sizes in the femoral greater trochanter compared to the femoral neck and shaft (Figure 4.2K). Principal mode 7 explained 1.22% of the variance in the proximal femur shape and was related not only to the relative sizes of the femoral neck, trochanter and shaft but also the relative curvatures of these parts (Figure 4.2G). Accordingly, ASM appears to be a powerful tool for comparing the difference in the shape of these ROI from DXA images between fracture and control subjects for identifying those subjects at a significantly higher risk of suffering a bone fracture. Of particular interest, the ASM can also be used as a way to measure and quantify the shape changes within subjects over a certain period from baseline to followup and, therefore, offers the opportunity of early identification of those subjects who may be at higher risk of

developing osteoporosis and fracture in the near future.

Although the ASM and the AAM may be useful as methods for extracting features of shape and grey-scale, several pitfalls need to be considered. One pitfall is that subjects with different ethnic backgrounds may have different bone morphologies and, therefore, the ASM and the AAM created based on one ethnic group may not optimally represent the individuals of other ethnic backgrounds. Mahfouz et al. conducted statistical shape analysis and identified differences in three-dimensional knee morphology among Caucasian, African American, and East Asian populations [177]. For this reason, creating separate ASM and AAM for different ethnic groups may be necessary to address this problem. In addition, the analysis with the model is inherently limited in its ability to represent 3D objects in 2D images, and this might cause the apparent variation in hip rotation seen in the DXA images and therefore bias the results.

Due to the coarse spatial resolution of the images, only gross texture features could be measured. The advantage of automatic texture analysis is that subtle intensity patterns, regardless of spatial size, may be discovered. Two texture analysis techniques, Gabor filters and textons, were chosen in this study to extract a set of texture features on several ROI. Our work indicates that there are texture patterns in DXA images that provide information regarding fracture risk. The discriminating accuracy using the features derived from the whole femoral neck region was found to be higher than using other regions. Furthermore, we found that the features from Gabor filters on the whole femoral neck region gave the best prediction of factures among all texture methods. Gabor filters are a powerful tool for extracting spatial orientation so that patterns of oriented structure may be recognised even if they appear indistinct to the human eye. Thus, this result indicates that orientation information is more important in the whole femoral neck than in other ROI. Compared with Gabor filters, the features from textons did not provide much information to further improve the performance.

Textural analysis of bone in assessment of fracture risk is a topical area; trabecular bone score is a recently developed tool that performs grey-level texture analysis on lumbar spine DXA images providing information relating to trabecular micro-architecture. Trabecular bone score has been shown to relate to fracture risk independent of clinical risk factors and aBMD, and has predictive value for fracture independent of fracture probabilities using FRAX [178]. The methodology discussed

in this thesis provides information at another site assessed by DXA images, the femoral hip.

There were no significant differences between the case and control groups with regard to neck aBMD and neck T-score (Table 4.1). However, the experimental results from this study show that the whole femoral neck ROI is much more important than the other ROI considered. These results are in line with the literature in that the findings using the most examined geometric parameters to discriminate fracture cases from control cases have often been inconsistent [68, 94]. In addition, the literature review (Section 2.3) reveals that the most examined geometric parameters, such as HAL and NSA, measured on the femoral neck do not give consistent results in predicting fracture risk. This indicates that the femoral neck is important for estimating risk, not because of its geometric properties but because of properties leading to DXA image texture.

Evaluating many features with a few data samples runs the risk of overfitting. To avoid overfitting, we used feature selection in this study to reduce the number of original features from more than 100 to 1, 3 or 6. For the amount of data samples available, 3 features is about the maximum that should reasonably be used in the classification step to avoid overfitting so we considered 1 or 3 features and then considered 6 features for comparison (Section 2.8.3). However, feature selection alone does not mitigate overfitting. Two more essential ingredients are necessary: the choice of classifier and testing the optimal feature subset on independent data to demonstrate that the fitting is not due to chance. Here we chose Fisher's linear classifier since it is the most robust against overfitting compared to popular methods such as neural networks, SVM, *k*-nearest neighbours, etc. Second, final conclusions are based entirely on performance scores form testing the feature combination on data that is independent of the data used to train the feature selection and classification steps. The final performance results reported in this thesis are based on independent testing data, thus, guarding against overfitting.

Fracture cases in the dataset included fractures of wrist, hip, lower limb, spine, etc. The number of these within each group was too small to explore the possibility that different image features, or combinations of features, might be optimal for estimating risk for these different fracture types separately. Accordingly, the results indicate that

the methods presented here apply to predicting the risk of fracture generally. In many situations, predicting fracture generally is more useful than predicting only femur fracture. However, predicting femur fracture from DXA images of the femur is likely to be more accurate than predicting general fracture from the same DXA images. Unfortunately, sufficient data was not available in this study to measure femur fracture risk prediction.

This study was limited by the quality of DXA images available. DXA images are widely available, but a study of the type reported here requires DXA images of fracture and matched non-fracture subjects. Only a few data sets exist satisfying this requirement and, unfortunately, all such image datasets known to the author are stored in proprietary image formats that prevent the application of novel image analysis methods. Accordingly, this study was based on relatively few images of low quality. While this limits the ability to properly estimate the full potential of the methods presented for estimating fracture risk, the application of these methods to higher resolution images can only improve results. Since improvement over existing methods was demonstrated on low quality images, this study suggests that estimates of fracture risk better than those reported here could be expected in practice using full resolution DXA images.

This study included elderly white subjects only and, thus, the results in this study do not automatically extend to younger people or other ethnic groups. However, if image texture characteristics differ in other groups then the specific weighting and parameters reported here may need to be recomputed, but the methods presented here provide the framework for doing so. Finally, the individuals recruited were selected because they had been born in Hertfordshire, and continued to live there at the age of 60–75 years. However, a previous study have demonstrated that the Hertfordshire populations studied have similar smoking characteristics and bone density compared with national figures [175], suggesting that selection bias is minimal.

A larger study would be needed to determine the best model for various ages or ethnic groups but once such models exist, all future processing is fully automatic and could be done in real time or close to real time. Accordingly, the methods introduced in this study hold great potential to address clinical needs in the future.

5.2 Conclusions

The objective of this research was to consider a broad range of parameters related to shape, density and texture of various regions of interest of the femur in order to improve fracture risk assessment. We have identified and assessed methodology that can be applied to derive further information from routinely collected DXA scan data. According to the results and discussion, the conclusion for this study are listed as follows:

1. In a modest sample, texture features derived from Gabor filters in combination with total T-score better separates the fracture and control groups with an AUC more than 10% higher than the standard measures of aBMD or total T-score alone. This indicates that the texture-based fracture risk estimation method presented in this thesis has the potential to improve upon current standard clinical practice.

2. Estimates of risk were more accurate when shape and texture were measured on the whole femoral neck in comparison with other feature combinations and other regions.

3. This research acts as a proof of concept and could be applied to images of other bone types or other imaging modalities.

4. Finally, with images at higher resolution, little, if any, improvement is expected in the contribution from AAM and ASM since these report gross features of the femur. However, greater resolution has the potential to capture much more texture information.

5.3 Future Work

Although the results indicate that including texture information with total T-score improves risk assessment, the full potential impact of image analysis techniques on risk has not been addressed due to the quality of the images and small sample size (for each type of fracture). Accordingly, in future work we hope to acquire a database of full resolution DXA images reflecting a wide range of fracture types. Full resolution DXA images that allow us to understand how well the shape and detailed structure of the femoral neck predicts the likelihood of fracture of the proximal femur and also how well it predicts low energy fracture generally.

If sufficient numbers of images are available, K-fold cross validation will be used to determine an estimate of performance of the techniques on unseen data. All these methods proposed are standard in machine learning.

In addition, the relationship between changes over time and risk of fracture has received very little attention. If possible, we would like to conduct a temporal research to investigate if the risk of fracture deduced from DXA images is constant or changes over time. If constant, this would indicate substantial robustness of the DXA based risk assessment. If variable, then the changes over time may in themselves provide additional information regarding risk of fracture.

Finally, the population used in this study include elderly white subjects only and, thus, further studies using younger people or other ethnic groups can be included to predict fracture risk. Additionally, this study was conducted solely on proximal femur from DXA images, but the technique developed in this study can be applied to other bones, such as vertebra or acetabulum, and other medical imaging applications, such as breast cancer risk assessment using mammographic images.

BIBLIOGRAPHY

1. Cummings, S., et al., *Bone density at various sites for prediction of hip fractures.* The Lancet, 1993. 341(8837): p. 72-75.
2. Stone, K.L., et al., *BMD at multiple sites and risk of fracture of multiple types: long-term results from the Study of Osteoporotic Fractures.* Journal of Bone and Mineral Research, 2003. 18(11): p. 1947-1954.
3. Cooney, L. and R. Marottoli. *Functional decline following hip fracture*. in *Fourth International Symposium on Osteoporosis and Consensus Development Conference Proceedings*. 1993.
4. Cummings, S., et al., *Epidemiology of osteoporosis and osteoporotic fractures.* Epidemiologic reviews, 1985. 7(1): p. 178-208.
5. Chrischilles, E., T. Shireman, and R. Wallace, *Costs and health effects of osteoporotic fractures.* Bone, 1994. 15(4): p. 377-386.
6. Holbrook, T.L. and K.L. Grazier, *The frequency of occurrence, impact, and cost of selected musculoskeletal conditions in the United States*. 1984: Amer Academy of Orthopaedic.
7. Peck, W., E. Barrett-Connor, and J. Buckwalter, *NIH Consensus Development Conference Statement.* Jama, 1984. 252: p. 799-802.
8. LJIII, M., *How many women have osteoporosis now*. J Bone Miner Res1995, 1995. 10: p. 175-7.
9. Cooper, C., *Osteoporosis--an epidemiological perspective: a review*. Journal of the Royal Society of Medicine, 1989. 82(12): p. 753.
10. Hammett-Stabler, C.A., *Osteoporosis: From Pathophysiology to Treatment: Special Topics in Diagnostic Testing*. 2004: Amer. Assoc. for Clinical Chemistry.
11. Bouillon, R., et al., *Consensus development conference: prophylaxis and treatment of osteoporosis.* Am. J. Med., 1991. 90: p. 107-110.
12. Kanis, J.A., *Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: report of a WHO study group.* Osteoporosis International, 1994. 4(6): p. 368-381.
13. Sanders, K.M., et al., *Health burden of hip and other fractures in Australia beyond 2000. Projections based on the Geelong Osteoporosis Study.* The Medical Journal of Australia, 1999. 170(10): p. 467-470.
14. Randell, A., et al., *Direct clinical and welfare costs of osteoporotic fractures in elderly men and women.* Osteoporosis International, 1995. 5(6): p. 427-432.
15. Stevens, J.A., et al., *Surveillance for injuries and violence among older adults.* MMWR CDC Surveill Summ, 1999. 48(8): p. 27-50.
16. Congress, U., *Office of Technology Assessment, Hip Fracture Outcomes in People Age 50 and Over–Background Paper*. US Government Printing Office, Washington, DC, 1994.

17. Brainsky, A., et al., *The economic cost of hip fractures in community-dwelling older adults: a prospective study.* Journal of the American Geriatrics Society, 1997. 45(3): p. 281-287.

18. Ray, N.F., et al., *Medical expenditures for the treatment of osteoporotic fractures in the United States in 1995: report from the National Osteoporosis Foundation.* Journal of Bone and Mineral Research, 1997. 12(1): p. 24-35.

19. Baker, P.N., et al., *Evolution of the hip fracture population: time to consider the future? A retrospective observational analysis.* BMJ open, 2014. 4(4): p. e004405.

20. Fracture, H., *Costing report. Implementing NICE guidance.* National Institute for Health and Care Excellence clinical guideline, 2011. 124.

21. Cooper, C., G. Campion, and L. Melton III, *Hip fractures in the elderly: a world-wide projection.* Osteoporosis international, 1992. 2(6): p. 285-289.

22. Johnell, O., *The socioeconomic burden of fractures: today and in the 21st century.* The American journal of medicine, 1997. 103(2): p. S20-S26.

23. Johnell, O. and J. Kanis, *An estimate of the worldwide prevalence and disability associated with osteoporotic fractures.* Osteoporosis international, 2006. 17(12): p. 1726-1733.

24. Van Staa, T., et al., *Epidemiology of fractures in England and Wales.* Bone, 2001. 29(6): p. 517-522.

25. Compston, J., *Action Plan for the prevention of osteoporotic fractures in the European Community.* Osteoporosis International, 2004. 15(4): p. 259-262.

26. Hans, D., et al., *Skeletal sites for osteoporosis diagnosis: the 2005 ISCD Official Positions.* Journal of Clinical Densitometry, 2006. 9(1): p. 15-21.

27. A, P., R.G. Josse, and S.A.C.o.t.O.S.o. Canada, *2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada:summary.* Canadian Medical Association Journal, 2002. 182(17): p. 1864-1873.

28. Genant, H.K., et al., *Interim report and recommendations of the World Health Organization task-force for osteoporosis.* Osteoporosis International, 1999. 10(4): p. 259-264.

29. Leib, E.S., et al., *Position Development Conference of the International Society for Clinical Densitometry. Vancouver, BC, July 15-17, 2005.* The Journal of rheumatology, 2006. 33(11): p. 2319-2321.

30. Duboeuf, F., et al., *Bone mineral density of the hip measured with dual-energy X-ray absorptiometry in normal elderly women and in patients with hip fracture.* Osteoporosis International, 1991. 1(4): p. 242-249.

31. Di Monaco, M., et al., *Total lymphocyte count and femoral bone mineral density in postmenopausal women.* Journal of bone and mineral metabolism, 2004. 22(1): p. 58-63.

32. Boehm, H.F., et al., *Prediction of the fracture load of whole proximal femur specimens by topological analysis of the mineral distribution in DXA-scan images.* Bone, 2008. 43(5): p. 826-831.

33. El Maghraoui, A. and C. Roux, *DXA scanning in clinical practice.* Qjm, 2008. 101(8): p. 605-617.

34. Vijay, A., et al. *Evaluation of osteoporosis using CT image of proximal femur compared with dual energy x-ray absorptiometry (DXA) as the standard*. in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*. 2011. IEEE.

35. Long, Y., W.D. Leslie, and Y. Luo, *Study of DXA-derived lateral–medial cortical bone thickness in assessing hip fracture risk.* Bone Reports, 2015. 2: p. 44-51.

36. Mayhew, P., et al., *Discrimination between cases of hip fracture and controls is improved by hip structural analysis compared to areal bone mineral density. An ex vivo study of the femoral neck.* Bone, 2004. 34(2): p. 352-361.

37. Gilsanz, V., *Bone density in children: a review of the available techniques and indications.* European journal of radiology, 1998. 26(2): p. 177-182.

38. Hawkinson, J., et al., *Technical white paper: bone densitometry.* Journal of the American College of Radiology, 2007. 4(5): p. 320-327.

39. Njeh, C., et al., *Radiological assessment of a new bone densitometer—the Lunar Expert.* The British journal of radiology, 1996. 69(820): p. 335-340.

40. Organization, W.H., *Prevention and management of osteoporosis: report of a WHO scientific group*. 2003: Diamond Pocket Books (P) Ltd.

41. Faulkner, K.G., M. Mcclung, and S.R. Cummings, *Automated evaluation of hip axis length for predicting hip fracture.* Journal of bone and mineral research, 1994. 9(7): p. 1065-1070.

42. Faulkner, K.G., et al., *Simple measurement of femoral geometry predicts hip fracture: the study of osteoporotic fractures.* Journal of bone and mineral research, 1993. 8(10): p. 1211-1217.

43. Cummings, S.R., et al., *Racial differences in hip axis lengths might explain racial differences in rates of hip fracture.* Osteoporosis International, 1994. 4(4): p. 226-229.

44. Adams, J.E., *Advances in bone imaging for osteoporosis.* Nature Reviews Endocrinology, 2013. 9(1): p. 28-42.

45. Barbu, C., *DUAL-ENERGY X-RAY ABSORPTIOMETRY APPLICATIONS BEYOND BONE DENSITOMETRY-SOMETHING OLD, SOMETHING NEW, SOMETHING BORROWED.* Acta Endocrinologica (1841-0987), 2014. 10(3).

46. BECK, T.J., et al., *Predicting femoral neck strength from bone mineral data: a structural approach.* Investigative radiology, 1990. 25(1): p. 6-18.

47. Bonnick, S.L., *Hsa: Beyond bmd with dxa.* Bone, 2007. 41(1): p. S9-S12.

48. Bouxsein, M.L. and D. Karasik, *Bone geometry and skeletal fragility.* Current osteoporosis reports, 2006. 4(2): p. 49-56.

49. Beck, T.J., *Extending DXA beyond bone mineral density: understanding hip structure analysis.* Current osteoporosis reports, 2007. 5(2): p. 49-55.

50. Mrgan, M., A. Mohammed, and J. Gram, *Combined vertebral assessment and bone densitometry increases the prevalence and severity of osteoporosis in patients referred to DXA scanning.* Journal of Clinical Densitometry, 2013. 16(4): p. 549-553.

51. Laster, A.J. and E.M. Lewiecki, *Vertebral Fracture Assessment by Dual-Energy X-ray Absorptiometry: Insurance Coverage Issues in the United States A White Paper of the International Society for Clinical Densitometry.* Journal of Clinical Densitometry, 2007. 10(3): p. 227-238.

52. El Maghraoui, A., et al., *Vertebral fracture assessment in healthy men: prevalence and risk factors.* Bone, 2008. 43(3): p. 544-548.

53. Rea, J., et al., *Visual assessment of vertebral deformity by X-ray absorptiometry: a highly predictive method to exclude vertebral deformity.* Osteoporosis international, 2000. 11(8): p. 660-668.

54. Fuerst, T., et al., *Evaluation of vertebral fracture assessment by dual X-ray absorptiometry in a multicenter setting.* Osteoporosis international, 2009. 20(7): p. 1199-1205.

55. Kanis, J., et al., *Case finding for the management of osteoporosis with FRAX®—assessment and intervention thresholds for the UK.* Osteoporosis international, 2008. 19(10): p. 1395-1408.

56. Kanis, J., et al., *FRAX™ and the assessment of fracture probability in men and women from the UK*. Osteoporosis International, 2008. 19(4): p. 385-397.

57. Kanis, J.A., et al., *Interpretation and use of FRAX in clinical practice.* Osteoporosis International, 2011. 22(9): p. 2395-2411.

58. Pothuaud, L., P. Carceller, and D. Hans, *Correlations between grey-level variations in 2D projection images (TBS) and 3D microarchitecture: applications in the study of human trabecular bone microarchitecture.* Bone, 2008. 42(4): p. 775-787.

59. Leslie, W., et al., *Adjustment of FRAX probability according to lumbar spine trabecular bone score (TBS): The Manitoba BMD Cohort.* Journal of Clinical Densitometry, 2013. 3(16): p. 267-268.

60.Guglielmi, G., *Osteoporosis and bone densitometry measurements*. 2013: Springer.

61. Krueger, D., et al., *Spine trabecular bone score subsequent to bone mineral density improves fracture discrimination in women.* Journal of Clinical Densitometry, 2014. 17(1): p. 60-65.

62. Hans, D., et al., *Bone microarchitecture assessed by TBS predicts osteoporotic fractures independent of bone density: the Manitoba study.* Journal of Bone and Mineral Research, 2011. 26(11): p. 2762-2769.

63. Lamy, O., et al., *What is the performance in vertebral fracture discrimination by Bone mineral density (BMD), micro-architecture estimation (TBS), and FRAX in stand-alone, combined or adjusted approaches: The OsteoLaus Study.* Journal of Clinical Densitometry, 2013. 16(3): p. 265.

64. Bousson, V., et al., *Trabecular bone score (TBS): available knowledge, clinical relevance, and future prospects.* Osteoporosis International, 2012. 23(5): p. 1489-1501.

65. Crabtree, N., et al., *Improving risk assessment: hip geometry, bone mineral distribution and bone strength in hip fracture cases and controls. The EPOS study.* Osteoporosis International, 2002. 13(1): p. 48-54.

66. Bergot, C., et al., *Hip fracture risk and proximal femur geometry from DXA scans.* Osteoporosis international, 2002. 13(7): p. 542-550.

67. Pulkkinen, P., et al., *BMD T-score discriminates trochanteric fractures from unfractured controls, whereas geometry discriminates cervical fracture cases from unfractured controls of similar BMD.* Osteoporosis international, 2010. 21(7): p. 1269-1276.

68. Gnudi, S., et al., *Geometry of proximal femur in the prediction of hip fracture in osteoporotic women.* The British journal of radiology, 1999. 72(860): p. 729-733.

69. Heaney, R., et al., *Bone dimensional change with age: interactions of genetic, hormonal, and body size variables.* Osteoporosis International, 1997. 7(5): p. 426-431.

70. Cootes, T.F. and C.J. Taylor. *Statistical models of appearance for medical image analysis and computer vision*. in *Medical Imaging 2001*. 2001. International Society for Optics and Photonics.

71. Gregory, J.S., et al., *Early identification of radiographic osteoarthritis of the hip using an active shape model to quantify changes in bone morphometric features: can hip shape tell us anything about the progression of osteoarthritis?* Arthritis & Rheumatism, 2007. 56(11): p. 3634-3643.

72. Gregory, J., et al., *A method for assessment of the shape of the proximal femur and its relationship to osteoporotic hip fracture.* Osteoporosis international, 2004. 15(1): p. 5-11.

73. Waarsing, J., et al., *A statistical model of shape and density of the proximal femur in relation to radiological and clinical OA of the hip.* Osteoarthritis and Cartilage, 2010. 18(6): p. 787-794.

74. Cooper, C. and L. Melton, *Magnitude and impact of osteoporosis and fractures.* Osteoporosis. Academic Press, San Diego, 1996: p. 419-434.

75. Bergmann, G., et al., *Hip contact forces and gait patterns from routine activities.* Journal of biomechanics, 2001. 34(7): p. 859-871.

76. Crowninshield, R., et al., *A biomechanical investigation of the human hip.* Journal of biomechanics, 1978. 11(1): p. 75-85.

77. Drake, R.L., W. Vogl, and A.W. Mitchell, *Gray's anatomy for students*. 3rd ed. 2005, Philipdelphia; London: Elsevier/Churchill Livingstone.

78. Jenkins, D., *Hollinshead's Functional anatomy of the limbs and back 8th ed.* Philadelphia: WB Saunders Company of Elsevier, 2002: p. p. 269.

79. Koch, J.C., *The laws of bone architecture.* American Journal of Anatomy, 1917. 21(2): p. 177-298.

80. Saladin, K., *The Appdendicular Skeleton.* Human Anatomy. International Edition. ed, 2007: p. 207-226.

81. Bell, K., et al., *Structure of the femoral neck in hip fracture: cortical bone loss in the inferoanterior to superoposterior axis.* Journal of Bone and Mineral Research, 1999. 14(1): p. 111-119.

82. Eventov, I., et al., *Osteopenia, hematopoiesis, and bone remodelling in iliac crest and femoral biopsies: a prospective study of 102 cases of femoral neck fractures.* Bone, 1991. 12(1): p. 1-6.

83. Joan, M.Z., et al., *Methods for the histological study of femoral neck bone remodelling in patients with fractured neck of femur.* Bone, 1993. 14(3): p. 249-255.

84. Brunner, L.C., L. Eshilian-Oates, and T.Y. Kuo, *Hip fractures in adults.* American family physician, 2003. 67(3): p. 537-542.

85. Larsen, D., *Assessment and management of hand and wrist fractures.* Nursing standard, 2002. 16(36): p. 45.

86. Hunter, D., *Triage nurse X-ray protocols for hand and wrist injuries.* Emergency nurse: the journal of the RCN Accident and Emergency Nursing Association, 2010. 17(9): p. 20-24.

87. Cooper, C., et al., *Incidence of clinically diagnosed vertebral fractures: A population-based study in rochester, minnesota, 1985-1989.* Journal of Bone and Mineral Research, 1992. 7(2): p. 221-227.

88.Ziegler, R., C. Scheidt-Nave, and G. Leidig-Bruckner, *What is a vertebral fracture?* Bone, 1996. 18(3): p. S169-S177.

89. Eastell, R., et al., *Classification of vertebral fractures.* Journal of Bone and Mineral Research, 1991. 6(3): p. 207-215.

90. Marshall, D., O. Johnell, and H. Wedel, *Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures.* Bmj, 1996. 312(7041): p. 1254-1259.

91. Dequeker, J. and F. Luyten, *Bone densitometry is not a good predictor of hip fracture.* BMJ-British Medical Journal, 2001. 323(7316): p. 795-797.

92. Robbins, J., et al., *Risk factors for hip fracture in women with high BMD: EPIDOS study.* Osteoporosis international, 2005. 16(2): p. 149-154.

93. Cummings, S.R., et al., *Risk factors for hip fracture in white women.* New England journal of medicine, 1995. 332(12): p. 767-774.

94. Dinçel, V.E., et al., *The association of proximal femur geometry with hip fracture risk.* Clinical Anatomy, 2008. 21(6): p. 575-580.

95. Alonso, C.G., et al., *Femoral bone mineral density, neck-shaft angle and mean femoral neck width as predictors of hip fracture in men and women.* Osteoporosis International, 2000. 11(8): p. 714-720.

96. Pande, I., et al., *Bone mineral density, hip axis length and risk of hip fracture in men: results from the Cornwall Hip Fracture Study.* Osteoporosis international, 2000. 11(10): p. 866-870.

97. Michelotti, J. and J. Clark, *Femoral neck length and hip fracture risk.* Journal of bone and mineral research, 1999. 14(10): p. 1714-1720.

98. Gregory, J.S. and R.M. Aspden, *Femoral geometry as a risk factor for osteoporotic hip fracture in men and women.* Medical engineering & physics, 2008. 30(10): p. 1275-1286.

99. Beck, T.J., et al., *Structural trends in the aging femoral neck and proximal shaft: analysis of the Third National Health and Nutrition Examination Survey dual-energy X-ray absorptiometry data.* Journal of Bone and Mineral Research, 2000. 15(12): p. 2297-2304.

100. Cootes, T.F., et al., *Active shape models-their training and application.* Computer vision and image understanding, 1995. 61(1): p. 38-59.

101. Cootes, T.F., G.J. Edwards, and C.J. Taylor, *Active appearance models.* IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001(6): p. 681-685.

102. Smyth, P.P., C.J. Taylor, and J.E. Adams, *Automatic measurement of vertebral shape using active shape models.* Image and Vision Computing 1997: p. 575-581.

103. Baker-LePain, J.C., et al., *Active shape modeling of the hip in the prediction of incident hip fracture.* Journal of Bone and Mineral Research, 2011. 26(3): p. 468-474.

104. Sarkalkan, N., H. Weinans, and A.A. Zadpoor, *Statistical shape and appearance models of bones.* Bone, 2014. 60: p. 129-140.

105. Lynch, J., et al., *The association of proximal femoral shape and incident radiographic hip OA in elderly women.* Osteoarthritis and Cartilage, 2009. 17(10): p. 1313-1318.

106. Gregory, J.S., et al., *Bone shape, structure, and density as determinants of osteoporotic hip fracture: a pilot study investigating the combination of risk factors.* Investigative radiology, 2005. 40(9): p. 591-597.

107. Van Ginneken, B., et al., *Active shape model segmentation with optimal features.* medical Imaging, IEEE Transactions on, 2002. 21(8): p. 924-933.

108. Cootes, T.F., et al., *The use of active shape models for locating structures in medical images.* Image and Vision Computing 1994. 12: p. 355-365.

109. Trevor, H., T. Robert, and F. Jerome, *The elements of statistical learning: data mining, inference and prediction.* New York: Springer-Verlag, 2001.

110. Bredbenner, T.L., et al., *Fracture risk predictions based on statistical shape and density modeling of the proximal femur.* Journal of Bone and Mineral Research, 2014. 29(9): p. 2090-2100.

111. Goodyear, S., et al., *Can we improve the prediction of hip fracture by assessing bone structure using shape and appearance modelling?* Bone, 2013. 53(1): p. 188-193.

112. Bryan, R., et al., *Statistical modelling of the whole human femur incorporating geometric and material properties.* Medical engineering & physics, 2010. 32(1): p. 57-65.

113. Bryan, R., P.B. Nair, and M. Taylor, *Use of a statistical model of the whole femur in a large scale, multi-model study of femoral neck fracture risk.* Journal of biomechanics, 2009. 42(13): p. 2171-2176.

114. Gower, J.C., *Generalized procrustes analysis.* Psychometrika, 1975. 40(1): p. 33-51.

115. Dryden, I.L. and K.V. Mardia, *Statistical shape analysis*. Vol. 4. 1998: Wiley Chichester.

116. Johnson, R. and D. Wichern, *Multivariate statistics, a practical approach*. 1988, Chapman & Hall Boca Raton.

117. Jolliffe, I., *Principal component analysis*. 2002: Wiley Online Library.

118. Lee, D.-T. and B.J. Schachter, *Two algorithms for constructing a Delaunay triangulation.* International Journal of Computer & Information Sciences, 1980. 9(3): p. 219-242.

119. Cootes, T.F., G.J. Edwards, and C.J. Taylor. *Comparing Active Shape Models with Active Appearance Models*. in *BMVC*. 1999.

120. Shapiro, L. and G.C. Stockman, *Computer vision. 2001*. ed: Prentice Hall, 2001.

121. Clausi, D.A. and M.E. Jernigan, *Designing Gabor filters for optimal texture separability.* Pattern Recognition, 2000. 33(11): p. 1835-1849.

122. Daugman, J.G., *Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters.* JOSA A, 1985. 2(7): p. 1160-1169.

123. Dunn, D., W.E. Higgins, and J. Wakeley, *Texture segmentation using 2-D Gabor elementary functions.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1994. 16(2): p. 130-149.

124. Dunn, D. and W.E. Higgins, *Optimal Gabor filters for texture segmentation.* Image Processing, IEEE Transactions on, 1995. 4(7): p. 947-964.

125. Daugman, J.G., *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1988. 36(7): p. 1169-1179.

126. Manjunath, B., C. Shekhar, and R. Chellappa, *A new approach to image feature detection with applications.* Pattern Recognition, 1996. 29(4): p. 627-640.

127. Heeger, D.J., *Model for the extraction of image flow.* JOSA A, 1987. 4(8): p. 1455-1471.

128. Grigorescu, S.E., N. Petkov, and P. Kruizinga, *Comparison of texture features based on Gabor filters.* Image Processing, IEEE Transactions on, 2002. 11(10): p. 1160-1167.

129. Yap, D.W.-H., et al. *Detecting femur fractures by texture analysis of trabeculae*. in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. 2004. IEEE.

130. Lim, S.E., et al. *Detection of femur and radius fractures in x-ray images*. in *Proc. 2nd Int. Conf. on Advances in Medical Signal and Info. Proc*. 2004.

131. Pramudito, J., et al., *Trabecular pattern analysis of proximal femur radiographs for osteoporosis detection.* Journal of biomedical and pharmaceutical engineering, 2007. 1(1): p. 45-51.

132. Petkov, N., *Biologically motivated computationally intensive approaches to image pattern recognition.* Future Generation Computer Systems, 1995. 11(4): p. 451-465.

133. Fogel, I. and D. Sagi, *Gabor filters as texture discriminator.* Biological cybernetics, 1989. 61(2): p. 103-113.

134. Turner, M.R., *Texture discrimination by Gabor functions.* Biological cybernetics, 1986. 55(2-3): p. 71-82.

135. Julesz, B., *Textons, the elements of texture perception, and their interactions.* Nature, 1981. 290(5802): p. 91-97.

136. Varma, M. and A. Zisserman, *A statistical approach to texture classification from single images.* International Journal of Computer Vision, 2005. 62(1-2): p. 61-81.

137. Malik, J., et al., *Contour and texture analysis for image segmentation.* International journal of computer vision, 2001. 43(1): p. 7-27.

138. Shotton, J., M. Johnson, and R. Cipolla. *Semantic texton forests for image categorization and segmentation*. in *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.

139. Petroudi, S., T. Kadir, and M. Brady. *Automatic classification of mammographic parenchymal patterns: A statistical approach*. in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*. 2003. IEEE.

140. Gangeh, M.J., et al., *A texton-based approach for the classification of lung parenchyma in CT images*, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. 2010, Springer. p. 595-602.

141. Varma, M. and A. Zisserman. *Texture classification: Are filter banks necessary?* in *Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on*. 2003. IEEE.

142. Gong, Y.C., M. Brady, and S. Petroudi, *Texture based mammogram classification and segmentation*, in *Digital Mammography*. 2006, Springer. p. 616-625.

143. Petroudi, S. and M. Brady. *Breast density characterization using texton distributions*. in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. 2011. IEEE.

144. Li, X.-Z., S. Williams, and M.J. Bottema, *Texture and region dependent breast cancer risk assessment from screening mammograms.* Pattern Recognition Letters, 2014. 36: p. 117-124.

145. Dana, K.J., et al., *Reflectance and texture of real world surfaces.* ACM Transactions on Graphics, 1999. 18(1): p. 1-34.

146. Fisher, R.A., *The use of multiple measurements in taxonomic problems.* Annals of eugenics, 1936. 7(2): p. 179-188.

147. Tahmasebi, P., A. Hezarkhani, and M. Mortazavi, *Application of discriminant analysis for alteration separation; sungun copper deposit, East Azerbaijan, Iran.* Australian Journal of Basic and Applied Sciences, 2010. 6(4): p. 564-576.

148. David, D.E., et al., *Evaluation of virulence factor profiling in the characterization of veterinary Escherichia coli isolates.* Applied and environmental microbiology, 2010. 76(22): p. 7509-7513.

149. Feinberg, F.M., *Discriminant Analysis for Marketing Research Applications.* Wiley International Encyclopedia of Marketing, 2010.

150. Holmström, L., et al., *Neural and statistical classifiers-taxonomy and two case studies.* Neural Networks, IEEE Transactions on, 1997. 8(1): p. 5-17.

151. McLachlan, G., *Discriminant analysis and statistical pattern recognition*. 1992, New York: John Wiley & Sons.

152. Metz, C.E. *Basic principles of ROC analysis*. in *Seminars in nuclear medicine*. 1978. Elsevier.

153. Metz, C.E., *ROC methodology in radiologic imaging.* Investigative radiology, 1986. 21(9): p. 720-733.

154. Rajeswari, S. and K. Theiva Jayaselvi, *Support vector machine classification for MRI images.* International Journal of Electronics and Computer Science Engineering, 2012. 1(3): p. 1534-1539.

155. Zou, K.H., A.J. O'Malley, and L. Mauri, *Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models.* Circulation, 2007. 115(5): p. 654-657.

156. Zweig, M.H. and G. Campbell, *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.* Clinical chemistry, 1993. 39(4): p. 561-577.

157. Jilaihawi, H., et al., *Cross-sectional computed tomographic assessment improves accuracy of aortic annular sizing for transcatheter aortic valve replacement and reduces the incidence of paravalvular aortic regurgitation.* Journal of the American College of Cardiology, 2012. 59(14): p. 1275-1286.

158. Bellman, R.E., *Adaptive control processes: a guided tour*. 1961: Princeton university press Princeton.

159. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* The Journal of Machine Learning Research, 2003. 3: p. 1157-1182.

160. Bermingham, M., et al., *Application of high-dimensional feature selection: evaluation for genomic prediction in man.* Scientific reports, 2015. 5.

161. Cohen, P.R. and D. Jensen. *Overfitting explained*. in *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*. 1997.

162. Bottema, M.J., G. Lee, and S. Lu, *Automatic image feature extraction for diagnosis and prognosis of breast cancer.* SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE, 2000. 39: p. 17-54.

163. Keinosuke, F., *Introduction to statistical pattern recognition.* Academic Press, Boston, 1990.

164. Doak, J., *An evaluation of feature selection methods and their application to computer security*. 1992: University of California, Computer Science.

165. Theodoridis S., K.K., *Pattern Recognition* 2006: Academic press.

166. Langley, P. *Selection of relevant features in machine learning*. in *AAAI Fall Symposium on Relevance*. 1994.

167. Smith, L.A. *Feature subset selection: a correlation based filter approach*. in *Neural Information Processing and Intelligent Information Systems*. 1997.

168. Hall, M.A., *Correlation-based feature selection for machine learning*. 1999, The University of Waikato.

169. Kittler, J., *Feature set search algorithms.* Pattern recognition and signal processing, 1978: p. 41-60.

170. Sambur, M.R., *Selection of acoustic features for speaker identification.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1975. 23(2): p. 176-182.

171. Siedlecki, W. and J. Sklansky, *On automatic feature selection.* International Journal of Pattern Recognition and Artificial Intelligence, 1988. 2(02): p. 197-220.

172. Pudil, P., J. Novovičová, and J. Kittler, *Floating search methods in feature selection.* Pattern recognition letters, 1994. 15(11): p. 1119-1125.

173. Aha, D.W. and R.L. Bankert, *A comparative evaluation of sequential feature selection algorithms*, in *Learning from Data*. 1996, Springer. p. 199-206.

174. Devijver, P.A. and J. Kittler, *Pattern recognition: A statistical approach*. Vol. 761. 1982: Prentice-Hall London.

175. Syddall, H., et al., *Cohort profile: the Hertfordshire cohort study.* International journal of epidemiology, 2005. 34(6): p. 1234-1242.

176. Kim, D., *Automated Face Analysis: Emerging Technologies and Research: Emerging Technologies and Research*. 2009: IGI Global.

177. Mahfouz, M., et al., *Three-dimensional morphology of the knee reveals ethnic differences.* Clinical Orthopaedics and Related Research®, 2012. 470(1): p. 172-185.

178. Harvey, N., et al., *Trabecular bone score (TBS) as a new complementary approach for osteoporosis evaluation in clinical practice.* Bone, 2015.