UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
ELECTRONICS AND COMPUTER SCIENCE

Quantitative Analysis of Bone Microarchitecture in HR-pQCT Images for Fracture Discrimination

by

Shengyu Lu

Thesis for the degree of Doctor of Philosophy

August 2024

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: [Publications P1-P4] in Section 1.6.

Signed:..... Date:.....

University of Southampton

Abstract

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

Quantitative Analysis of Bone Microarchitecture in HR-pQCT Images for Fracture Discrimination

by Shengyu Lu

Osteoporosis is the pathological disorder of bones, characterised by decreased bone mineral density (BMD) and microarchitectural deterioration of bone tissue, leading to increased fracture risk. Early observation and treatment of fracture risk from potential patients can reduce the incidence and medical expenses. While much of clinical fracture risk assessment is carried out through dual-energy X-ray absorptiometry (DXA) imaging, high-resolution peripheral quantitative computed tomography (HR-pQCT) is becoming increasingly available, providing more detailed analysis of bone microarchitecture at the peripheral skeleton. Traditional analysis of HR-pQCT images requires manual operation and results in a multitude of cortical and trabecular parameters which would be potentially cumbersome to interpret for clinicians. Automated quantitative analysis of HR-pQCT scans to ascertain fracture risk, would be far simpler and more efficient.

Our research work primarily focuses on the dataset from the Hertfordshire Cohort Study (HCS), which comprises 2997 men and women born in Hertfordshire from 1931-1939 and who still lived there in 1998-2004. 376 participants of the HCS attended research visits at which clinical covariates were measured; fracture history was determined via self-report and vertebral fracture assessment. Bone microarchitecture was assessed via HR-pQCT scans of the non-dominant distal tibia, and BMD measurement and lateral vertebral assessment were performed using DXA. In addition, our study utilises the dataset from the Global Longitudinal Study of Osteoporosis in Women (GLOW) which involves 723 physicians and 60,393 women aged 55 years and older in 10 countries. In this cohort, 501 participants completed self-administered questionnaires and underwent HR-pQCT scans of the non-dominant distal radius and tibia, as well as DXA scans of whole body, proximal femur and lumbar spine.

Building upon the correlation between bone microarchitecture and fracture risk, we develop an automatic approach to discriminate previous fractures by using HR-pQCT

measures of bone microarchitecture. We propose a method based on local binary pattern (LBP) to characterise the texture patterns of HR-pQCT images and to quantify bone microarchitecture via statistical distributions. Further, we decouple the relative contributions of cortical and trabecular compartments in fracture discrimination. Our method includes a deep neural network-based segmentation algorithm for separating the cortical and trabecular regions to enable texture features to be extracted separately and their statistical distributions quantified.

In addition to volumetric texture analysis, we present a novel discriminative system to automatically identify individuals with previous fractures from HR-pQCT images using a combination of multi-view convolutional neural networks (CNNs) and the random forest algorithm. Unlike conventional deep learning architectures that require a massive amount of training data, our method based on transfer learning extracts image features from representative views of HR-pQCT scans to characterise bone microarchitecture and then integrate the features for fracture discrimination.

Last but not least, we propose an adaptive threshold strategy to further enhance the accuracy and robustness of our discriminative system for previous fracture. Our method generates adaptive thresholds based on DXA-measured T-scores of the participants within the population to filter out healthy subjects with traumatic fractures and osteoporotic non-fractured subjects. Then we adopt multi-view CNNs to characterise bone microarchitecture in HR-pQCT images to distinguish between non-fractured healthy subjects and subjects with osteoporotic fractures. Furthermore, we evaluate the performance of our discriminative system on an independent cohort.

Contents

List of Figures

List of Tables

Abbreviations

BMD	bone mineral density
DXA	dual-energy X-ray absorptiometry
HR-pQCT	high-resolution peripheral quantitative computed tomography
HCS	Hertfordshire Cohort Study
LBP	local binary pattern
AUC	area under the curve
IoU	Intersection of Union
CT	computed tomography
2D	two dimensional
3D	three dimensional
WHO	World Health Organisation
NIH	National Institutes of Health
BMI	body mass index
BMC	bone mineral content
BA	bone area
aBMD	areal bone mineral density
SD	standard deviation
QCT	quantitative computed tomography
SVM	support vector machine
MRI	magnetic resonance imaging
GMRF	Gaussian Markov random fields
CNN	convolutional neural network
CLBP	completed local binary pattern
ELBP	extended local binary pattern
RLBP	robust local binary pattern
DTLBP	directional thresholded local binary pattern
MRELBP	median robust extended local binary pattern
VLBP	volume local binary pattern
ROC	receiver operating characteristic
TBS	trabecular bone score
ReLU	rectified linear unit
up-conv	up-convolution

SNR signal-to-noise ratio
TP true positive
FP false positive
TN true negative
FN false negative
TPR true positive rate
FPR false positive rate
TNR true negative rate
CI confidence interval
vBMD volumetric bone mineral density
XGBoost extreme gradient boosting
ELM extreme learning machine
BN batch normalization
ReLU rectified linear unit
RCLBP robust completed local binary pattern

Acknowledgements

First and foremost, I would like to express my deep gratitude to my supervisor Dr. Sasan Mahmoodi, who gave me the opportunity to pursue a Ph.D. degree and provided continuous support and guidance during my research. His professional knowledge, down-to-earth attitude and friendly character have benefited me a lot. I am also very thankful to my supervisor Prof. Mahesan Niranjan for his kind support and numerous valuable advice. I have learned not only from his wide knowledge but also from the confident and enthusiastic attitude of his life.

I would like to deeply appreciate Dr. Nicholas R Fuggle, Dr. Leo Westbury and Prof. Cyrus Cooper for the wonderful collaboration experience. Special thanks for their selflessness support and generosity and for teaching me clinical knowledge. I acknowledge the MRC Lifecourse Epidemiology Centre for supplying the clinical data from the Hertfordshire Cohort Study and the Global Longitudinal Study of Osteoporosis in Women. Thanks to Prof. Nicholas Harvey, Prof. Kate A Ward and Prof. Elaine M Dennison for their insightful suggestions and honest help. Without their trust and support, my research work would not have been possible.

I acknowledge the Doctoral Scholarship covering the full tuition fee from the School of Electronics & Computer Science, University of Southampton, and appreciate the China Scholarship Council providing living cost stipends.

I am proud of being a member of the Vision, Learning and Control group at the University of Southampton, which has made my life interesting as well as challenging. I would like to acknowledge the generous help from academic staff and colleagues in the group. I especially thank Dr. Srinandan Dasmahapatra and Dr. Hansung Kim for their valuable and constructive suggestions on reviewing my progression reports.

Last but not least, I would like to gratitude my parents and my girlfriend for their incredible love, continuous encouragement and unconditional support throughout my life.

To all doctors and healthcare workers around the world . . .

Chapter 1

Introduction

1.1 Background

Osteoporosis is characterised by a reduction in bone mineral density (BMD) and microarchitectural deterioration of bone tissue, leading to a predisposition to fracture (Christodoulou and Cooper (2003)). This skeletal disease is associated with substantial morbidity and mortality (Katsoulis et al. (2017), Haentjens et al. (2010)). It has been reported that fractures often occur in the elderly and about 50% of women suffer from osteoporosis in their lifetime (Löffler et al. (2021)). The global prevalence of individuals at high risk of fragility fracture is greater than 158 million and is set to double by the year 2040 (Oden et al. (2015)). This will lead to a substantial increase in economic costs associated with osteoporotic fractures. However, identifying those at high risk of fracture means that they can be treated with effective medications to reduce their fracture risk and improve outcomes (Hoff et al. (2021)).

Traditionally fracture risk prediction to target preventative measures has rested upon clinical risk factors and BMD (Kanis et al. (2008), Nguyen et al. (2008), Hippisley-Cox and Coupland (2009)). Currently, the gold standard for radiologists to assess fracture risk is the measurement and quantitative assessment of areal bone mineral density (aBMD) through dual-energy X-ray absorptiometry (DXA). However, as a two dimensional (2D) imaging modality, DXA cannot provide detailed information about cortical and trabecular bone microarchitecture (Mikolajewicz et al. (2020)). More recently high-resolution peripheral quantitative computed tomography (HR-pQCT) has been proposed, previous studies have demonstrated bone microarchitecture phenotypes associated with a high risk of fracture (Edwards et al. (2016), Westbury et al. (2019)), which suggests that this imaging modality might help predict fracture occurrence. However, HR-pQCT requires manual operation and results in a large number of clinical variables. There is no available way in which information from HR-pQCT images can be adequately integrated into a convenient fracture risk assessment tool. Novel artificial

intelligence techniques have the potential to assist.

Machine learning has become increasingly popular because it can automatically learn the features from current instances and provide predictions for new cases (Kaissis et al. (2020)). Machine learning methods have demonstrated success in many medical tasks such as lesion detection and risk assessment (Tjoa and Guan (2020), De Bruijne (2016)), some researchers have proposed the use of computer-assisted diagnostic algorithms to diagnose osteoporosis or predict fracture risk (Wani and Arora (2020), Cruz et al. (2018), Kruse et al. (2017), Kilic and Hosgormez (2016)). However, in the field of imaging these studies have been limited to specific imaging features (for example finite element analysis) (Nishiyama et al. (2014)) or conventional computed tomography (CT) (Valentinitsch et al. (2019), Muehlematter et al. (2019)), and have not utilised the detailed, textural information on bone microarchitecture which can be gleaned from HR-pQCT.

1.2 Motivations

Identifying subjects at substantial risk of fracture is crucial for preserving bone health. One of the most substantial determinants of sustaining a fracture is indeed a history of having had a fracture in the past (Johansson et al. (2017)). Individuals who have experienced fractures in the past are more likely to have decreased bone strength and compromised bone health, making them more susceptible to future fractures. By discriminating the previous fracture history of individuals, healthcare professionals can identify those at substantial risk of fracture and design tailored prevention plans to reduce healthcare costs.

Over the decades, a number of prediction models have been developed to automatically measure fracture risk. However, these approaches primarily rely on clinical risk factors and DXA-measured BMD, and their diagnostic performance is not perfect (Bolland et al. (2011), Cummins et al. (2011)). HR-pQCT is an advanced and noninvasive imaging instrument that captures detailed bone microarchitectural information and produces volumetric images to visualize and quantify bone structure. However, there is currently no automated tool available for quantitative analysis of HR-pQCT images to assess fracture risk. Traditional analysis of HR-pQCT requires manual operation and results in a multitude of cortical and trabecular parameters which would be potentially cumbersome to interpret for clinicians. Existing HR-pQCT-based fracture risk prediction models have four limitations, and there is still room for improvement.

The first weakness is that existing approaches cannot capture comprehensive bone microarchitectural information from HR-pQCT images for automated fracture risk assessment. They function like black boxes and rely on cortical and trabecular parameters

that provide limited bone microarchitectural information.

The second weakness is that there is a lack of fair comparison and analysis between HR-pQCT and traditional measurements such as DXA and clinical risk factors for fracture risk assessment.

The third weakness is that they do not investigate the individual contributions of cortical and trabecular compartments in HR-pQCT images to fracture risk assessment.

Finally, existing approaches are developed for specific populations and are not tested on independent cohorts (Sornay-Rendu et al. (2017), Langsetmo et al. (2018)). Evaluation of prediction models on a single population is insufficient to ensure their robustness.

By using computer vision techniques to characterise bone microarchitecture, decoupling the relative contributions of cortical and trabecular compartments, and enhancing model robustness, we can better understand fracture risk prediction models and improve discriminative accuracy. In relation to fracture risk assessment, there are three questions of interest:

1. Can fracture risk be identified through automated quantitative analysis of HR-pQCT images?

2. Does image information obtained from HR-pQCT outperform DXA-measured BMD and clinical risk factors in fracture risk assessment?

3. What are the relative contributions of cortical and trabecular compartments in predicting fracture risk?

In HR-pQCT images, each voxel represents a nominal resolution of 82 μm. The variation in grayscale values between adjacent voxels reflects changes in bone microarchitecture. The presence of such variations can be captured by image processing for quantitatively assessing bone health. Since texture features capture spatial patterns and statistical properties of pixel intensities within a local neighborhood, they have the potential to quantify bone microarchitecture and provide richer information beyond traditional bone density measurements to improve fracture risk assessment. Although traditional deep learning models have demonstrated remarkable capabilities in image feature extraction, their success heavily relies on the availability of large and diverse training datasets (Aljabri et al. (2022)). In clinical practice, it is challenging to acquire a large number of bone HR-pQCT images and the corresponding fracture status of those

participants. Therefore, instead of the traditional deep learning framework, we develop two approaches to encode image features to characterise bone microarchitecture in HR-pQCT images for fracture classification, as described in Chapter 4 and Chapter 6. Furthermore, we decouple the relative contributions of cortical and trabecular compartments for fracture discrimination in Chapter 5, and enhance the robustness of our discriminative system in Chapter 7.

Our research is based on the hypothesis that there are differences in bone microarchitecture in HR-pQCT images between fracture and non-fracture groups. Therefore, we use computer vision approaches to extract image features to characterise bone microarchitecture and employ machine learning techniques to distinguish between subjects with and without previous fractures.

1.3 Goals

In order to address the issues shown above, we conduct a series of studies. Specifically,

Aim 1: to assess the association between information obtained from HR-pQCT and fracture risk. To achieve this objective, we first develop an automated approach based on three dimensional (3D) texture representations to quantify bone microarchitecture measured by HR-pQCT to identify previous fractures. In addition, we propose a method based on deep learning techniques to automatically encode feature representations of bone HR-pQCT images to discriminate between subjects with and without previous fractures.

Aim 2: to compare the performance of HR-pQCT with traditional methods of DXA and clinical risk factors for fracture risk assessment. We use the same data partition method for these three approaches and conduct a comparative analysis for fracture discrimination.

Aim 3: to assess the relative contributions of cortical and trabecular compartments in fracture discrimination. We develop a deep neural network-based segmentation algorithm to automatically separate various regions in HR-pQCT images. This enables us to quantify texture patterns from cortical and trabecular regions separately for fracture classification.

Aim 4: to improve the accuracy and robustness of our approach for fracture classification. We propose an efficient strategy to filter out incorrectly labeled data from original

cohorts. Subsequently, we employ multi-view CNNs to characterise bone microarchitecture in HR-pQCT images for discriminating between non-fractured healthy subjects and subjects with osteoporotic fractures.

1.4 Contributions

The primary contribution of our research is to develop discriminative systems to automatically quantify bone microarchitecture in HR-pQCT images using texture representations and deep learning techniques for fracture classification. Therefore, we highlight the following contributions:

Firstly, considering the prior knowledge that fracture occurrences are associated with microarchitectural deterioration of bone tissue, we propose a 3D local binary pattern (LBP) model to characterise the texture patterns of bone HR-pQCT images. Then histograms are constructed to quantify bone microarchitecture through statistical distributions. Our discriminative system can automatically identify individuals with previous fractures from HR-pQCT images. Furthermore, our approach applied to HR-pQCT images improves fracture discrimination compared to DXA-measured BMD and clinical risk factors.

Secondly, we decouple the relative contributions of cortical and trabecular compartments in fracture discrimination. Our method includes a deep neural network for automatic segmentation of cortical and trabecular regions in HR-pQCT images, and then extracts texture features separately and quantifies their statistical distributions. We find that the cortical compartment outperforms the trabecular compartment in terms of fracture discrimination.

Thirdly, we propose an automatic approach based on deep learning techniques to characterise bone microarchitecture and combine it with the random forest classifier for fracture discrimination. Unlike traditional neural networks that require a massive amount of training data, our approach can automatically discriminate between people with and without previous fractures based on a few of HR-pQCT images. Furthermore, our method outperforms DXA measurement and clinical risk assessment methods in fracture classification.

Last but not least, we introduce a learning system that exploits DXA BMD and HR-pQCT images to further improve fracture discrimination. Our approach generates adaptive DXA-measured T-score thresholds to separate non-fractured healthy subjects from osteoporotic fracture subjects in the fracture group and to take out osteoporotic non-fractured patients from the healthy individual group within the original cohort.

By effectively filtering out incorrectly labeled data, we improve the discriminative capacity of our approach to accurately identify osteoporotic fractures. Furthermore, we evaluate the performance of our prediction model on an independent cohort, and it maintains high accuracy in fracture classification.

1.5 Thesis Outline

The structure of this thesis is organized as follows: Chapter 2 reviews the related literature about fracture risk assessment, volumetric texture analysis and deep learning methods; The study design, data collection and data processing for the HCS and GLOW cohorts, are then presented in Chapter 3. In Chapter 4, we develop an automatic discriminative system to identify previous fractures from HR-pQCT images using a combination of volumetric texture analysis and machine learning techniques; Chapter 5 introduces a novel approach that automatically segments cortical and trabecular regions in HR-pQCT images and separately extracts texture features from these two regions for fracture discrimination; An automatic method based on deep learning techniques is proposed in Chapter 6 to characterise bone microarchitecture in HR-pQCT images for fracture classification; Chapter 7 presents an enhanced fracture discriminative system and evaluates it on an independent cohort. Finally, in Chapter 8, we conclude this study and describe future work.

1.6 Publications

Publications based on this research include:

P1. Nicholas R Fuggle, **Shengyu Lu**, Michael O Breasail, Leo D Westbury, Kate A Ward, Elaine Dennison, Sasan Mahmoodi, Mahesan Niranjan and Cyrus Cooper. OA22 machine learning and computer vision of bone microarchitecture can improve the fracture risk prediction provided by DXA and clinical risk factors. Rheumatology, 61 (Supplement 1):keac132–022, 2022b.

P2. Shengyu Lu, Nicholas R Fuggle, Leo D Westbury, Michael O Breasail, Gregorio Bevilacqua, Kate A Ward, Elaine M Dennison, Sasan Mahmoodi, Mahesan Niranjan and Cyrus Cooper. Machine learning applied to HR-pQCT images improves fracture discrimination provided by DXA and clinical risk factors. Bone, page 116653, 2022a.

P3. Shengyu Lu, Sasan Mahmoodi and Mahesan Niranjan. Robust 3D rotation invariant local binary pattern for volumetric texture classification. In 2022 26th International Conference on Pattern Recognition (ICPR), pages 578–584. IEEE, 2022b.

P4. Nicholas R Fuggle, **Shengyu Lu**, Michael O Breasail, Leo D Westbury, Kate A Ward, Elaine Dennison, Sasan Mahmoodi, Mahesan Niranjan and Cyrus Cooper. Machine learning and computer vision of bone microarchitecture can improve the fracture risk prediction provided by DXA and clinical risk factors. In Aging Clinical and Experimental Research, volume 34, pages S43–S43, 2022a.

P5. Nicholas R Fuggle, **Shengyu Lu**, Michael O Breasail, Leo D Westbury, Kate A Ward, Elaine Dennison, Sasan Mahmoodi, Mahesan Niranjan and Cyrus Cooper. A deep learning, computer vision approach to segmentation of bone microarchitecture highlights the role of the cortical compartment in fracture discrimination, In preparation.

P6. **Shengyu Lu**, Sasan Mahmoodi, Mahesan Niranjan, Nicholas R Fuggle, Leo D Westbury, Kate A Ward, Elaine M Dennison and Cyrus Cooper. Enhanced fracture discrimination from HR-pQCT images using deep learning with adaptive thresholds, In preparation.

Chapter 2

Literature Review

2.1 Osteoporosis and Fracture

Osteoporosis was initially defined as a skeletal disorder characterised by a reduction in BMD, leading to an increased risk of fracture by the World Health Organisation (WHO) (Kanis et al. (1994)). Then, the National Institutes of Health (NIH) updated the definition of osteoporosis as a skeletal disorder characterised by compromised bone strength, which predisposes individuals to fractures (Klibanski et al. (2001)). The latter definition considers not only bone mass but also the reflection of bone structure in the loss of trabecular connectivity and thickness. Bone consists of a hard outer shell called cortex and a spongy tissue called trabecular. These two components combine to make bone strong but relatively flexible. Figure 2.1 illustrates examples of normal bone and bone with osteoporosis.

BMD, which characterises the amount of bone mass, is commonly used to assess bone quality and evaluate fracture risk. The decrease in BMD is common in people over 50 years of age and menopausal women who experience a decline in reproductive hormones (Finkelstein et al. (2008)). As a result, these populations often suffer from osteoporosis and fractures. Apart from BMD, osteoporosis is also affected by other risk factors such as physical activity, body mass index (BMI), height, weight and dietary calcium. Among these factors, BMI, advancing ages and BMD, play major roles in bone health assessment. The Dubbo Osteoporosis Epidemiology Study reported that bone loss increased with low BMD or low body weight, physical inactivity and advancing age (Ho (2018)).

Fracture, defined as a break at any skeletal site, is the ultimate consequence of osteoporosis. The distal forearm and spine are common sites affected by osteoporosis, and they have a higher risk of fracture. Fracture is affected by multiple risk factors (Litwic (2020)), as shown in Table 2.1. Modifiable risk factors, which can be changed

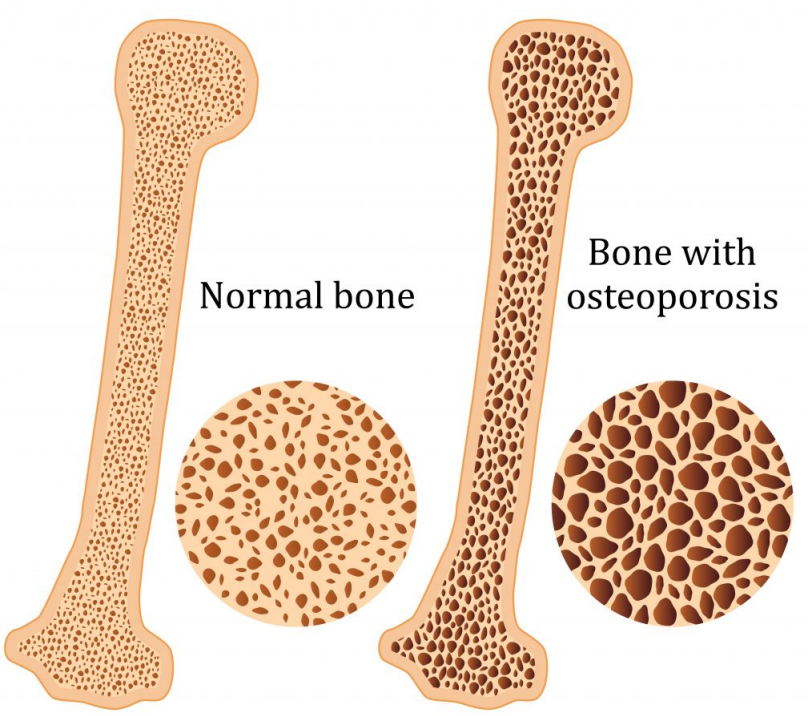

FIGURE 2.1: A diagram comparing normal bone and bone with osteoporosis. Low bone mass and microarchitectural deterioration of bone tissue predispose to fractures. Source: https://myfamilyphysio.com.au/osteoporosis/

TABLE 2.1: Risk factors for fracture.

Modifiable	Non-modifiable
Low BMD	A history of fracture
Low BMI	Ageing
Insufficient calcium intake	Being a woman
Physical inactivity	A high-risk genetic profile
High alcohol consumption	
Smoking	

through intervention, mainly include low BMD, low BMI, insufficient calcium intake, physical inactivity, high alcohol consumption and smoking. Non-modifiable factors are unchangeable makers such as a history of fracture and ageing. Once someone suffers a fracture, it will be challenging for the patient to recover completely. Therefore, there is increasing focus on identifying subjects at high risk of fracture to prevent such events.

The tibia and fibula are integral components of the human skeletal system, located in the lower leg. The tibia is the larger, weight-bearing bone on the inside, while the fibula is the smaller bone on the outside (Bardeen (1905)). Both bones are susceptible to fractures, commonly resulting from traumatic injuries. Tibia fractures are more prevalent than fibula fractures due to the tibia's weight-bearing role. By utilizing advanced clinical imaging techniques, healthcare providers can accurately assess the condition of

TABLE 2.2: T-score values for different categories of osteoporosis.

No.	Categories	T-score
1	Normal	>-1
2	Osteopenia	Between -2.5 and -1
3	Osteoporosis	<-2.5

the tibia, facilitating early diagnosis and effective treatment.

The gold standard for radiologists to diagnose osteoporosis is the quantitative assessment of the amount of X-ray absorption by minerals inside the bone through DXA (Watts (2004)). The results are then interpreted according to the T-score, with diagnostic criteria shown in Table 2.2. Furthermore, DXA-measured BMD combined with clinical factors such as age, gender and BMI is used for fracture risk assessment. However, the application of DXA has some limitations. DXA as a 2D measurement, areal BMD is a composite of cortical and trabecular bone that may be affected by cortex and hyperosteogeny (degenerative age-related change such as osteoarthritis) during BMD measurement and underestimate the loss of bone mass (Sukumar et al. (2011), Tsujii et al. (2017), Zhang et al. (2020)); meaning many individuals with fracture do not have osteoporosis by definition (Kanis et al. (1994)). In addition, as a two dimensional (2D) imaging technology, it may not fully capture spatial information regarding bone geometry and microarchitecture.

The use of 3D imaging techniques has the potential to overcome the limitations of DXA measurements and improve the accuracy of fracture risk assessment. In a related study, Löffler et al. (2021) compared routine CT with DXA in discriminating 192 patients who had suffered vertebral fractures. They used a CNN model to automatically segment vertebrae in CT scans and extracted various volumetric measures from vertebral bodies. The results demonstrated that CT-based measures performed significantly better than DXA in identifying patients with vertebral fractures. However, this study was limited to vertebral fracture assessment.

HR-pQCT is a non-invasive 3D imaging modality that provides detailed cortical and trabecular bone microarchitecture. In clinical practice, it is very difficult to capture bone microarchitectural information from HR-pQCT images via visual inspection. Although a multitude of cortical and trabecular parameters provide an opportunity to quantify bone microarchitecture (Mikolajewicz et al. (2020)), these clinical parameters would be potentially cumbersome to interpret for clinicians. Therefore, there is a need to develop convenient tools that capture bone microarchitectural information from HR-pQCT images and enable fully automated fracture risk assessment.

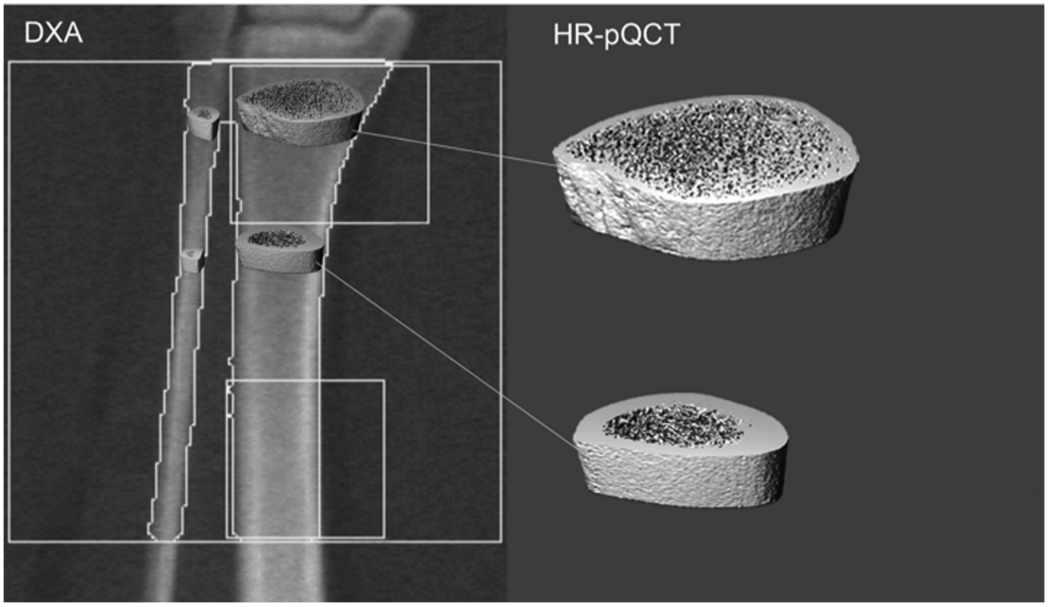

FIGURE 2.2: Comparing clinical imaging of the distal tibia with DXA and HR-pQCT. HR-pQCT provides more detailed information about the bone microarchitecture of cortical and trabecular compartments. Modified from source (Popp et al. (2014)): https://doi.org/10.1371/journal.pone.0088946.g002

2.2 Instruments to Measure Bone Health

Bone health assessment is critical for the diagnosis, management and prevention of fragility fracture. Various instruments and imaging techniques have been developed to measure different aspects of bone health such as bone density, microarchitecture and strength. Among the commonly utilised instruments, DXA provides precise measurements of bone density (see Figure 2.2), while HR-pQCT offers detailed 3D images of bone microarchitecture (Popp et al. (2014)). In our research, the femoral neck of the hip is measured using DXA, and BMD values are derived to quantify bone density information of subjects. In traditional DXA measurements, femoral neck BMD values are used as input data for fracture risk assessment. The distal tibia in the leg is measured using HR-pQCT to provide 3D reconstructions of bone microarchitecture. Our methods are applied to tibial HR-pQCT images to extract features and quantify bone microarchitecture for fracture classification.

2.2.1 Dual-Energy X-ray Absorptiometry

DXA is the gold standard method to measure BMD. The good accuracy of DXA makes it a reliable tool for evaluating bone health, assisting clinicians in identifying individuals at high risk of fracture and guiding medical decisions. DXA employs a high-energy X-ray beam and a low-energy X-ray beam directed from the radiation source to the radiation detector to assess BMD (Litwic (2020)). During the DXA scan, these two X-ray

TABLE 2.3: Common HR-pQCT parameters for fracture risk assessment.

Parameters	Descriptions
Cortical area	The cross-sectional area of the cortical bone
Cortical thickness	The thickness of the cortical bone
Cortical density	The mineral density within the cortical bone
Cortical porosity	The presence of pores or openings within the cortical bone
Trabecular area	The cross-sectional area of the trabecular bone
Trabecular density	The mineral density within the trabecular bone
Trabecular number	The number of trabeculae within a given volume
Trabecular thickness	The thickness of the trabecular bone
Trabecular separation	The distance between individual trabeculae

beams with different energies pass through the body of the participants. Some photons are absorbed while others that pass through the body are detected by the X-ray detectors. The absorption levels of each X-ray beam by bone mineral and soft tissues are calculated. The scanner then calculates crucial clinical parameters, including bone mineral content (BMC), bone area (BA) and aBMD. aBMD is defined as the bone mineral mass per unit image area (Litwic (2020)). The T-score is the standardised BMD measurement calculated for each individual by subtracting the population mean (Mean) and dividing by the standard deviation (SD) of the reference population (Ward et al. (2023)):

$$\text{T-score} = \frac{\text{aBMD} - \text{Mean}}{\text{SD}} \tag{2.1}$$

2.2.2 High-Resolution Peripheral Quantitative Computed Tomography

HR-pQCT is an advanced imaging technique designed to capture detailed 3D images of the skeletal structure (Nishiyama and Shane (2013)). This instrument employs a 360-degree rotating X-ray tube that generates two X-ray beams with different energies to pass through the body of the participant and assess bone microarchitecture. After passing through the soft tissues, the X-ray beams are measured by detectors, forming a profile of transmitted radiation. The spatial distribution of radiation absorption is then calculated based on the profile and reconstructed into an image as a 2D slice. These plane slices are combined to form a 3D image that visualizes microscopic details of the participant's skeletal structure (Litwic (2020)). In addition, HR-pQCT provides a multitude of clinical parameters through manual operation and complex processing to quantify bone microarchitecture. Common HR-pQCT parameters utilised for fracture risk assessment are cortical area, cortical thickness, cortical density, cortical porosity, trabecular area, trabecular density, trabecular number, trabecular thickness and trabecular separation, as shown in Table 2.3.

2.3 Artificial Intelligence for Fracture Risk Assessment

With the development of artificial intelligence, clinical experts are increasingly interested in utilising computer-aided diagnosis techniques to analyse and interpret patients' medical data for fracture risk assessment. Various types of medical data such as medical images, clinical variables and gene sequences are collected, and relevant features are extracted to train models for fracture risk prediction (Cruz et al. (2018)). They aim to leverage advanced techniques to establish a relationship between medical data and fracture risk to enhance early identification and prevention of fractures.

Over the decades, machine learning has become increasingly popular because it can automatically learn features from current instances and make predictions for new cases. Referring to the success of machine learning techniques in many medical tasks such as disease diagnosis (Criminisi (2016)), numerous computer-aided diagnosis methods based on machine learning have been developed to identify individuals at high risk of fracture.

Compared with clinical variables and gene sequences, medical imaging with visual bone structure and clinical interpretation has demonstrated better outcomes and has attracted more attention (Wani and Arora (2020)). For instance, in a study by Nishiyama et al. (2014), a method was proposed that utilised finite element analysis based on quantitative computed tomography (QCT) images to distinguish between 50 women with and without previous hip fractures. Both bone material and structural properties were captured through QCT images. Bone stiffness and bone failure load under various loading conditions were estimated using finite element analysis. The support vector machine (SVM) was used to classify subjects with and without pooled fractures, achieving an area under the curve (AUC) of 0.83. Similarly, in another study by Muehlematter et al. (2019), bone texture analysis was combined with machine learning classifiers in standard CT images to identify vertebral insufficiency fractures. The trabecular bone of vertebrae on CT scans was manually segmented, and image features were extracted using the MaZda software to build prediction models for classifying vertebral insufficiency fractures. Bone texture analysis involved six different types of feature descriptors calculated by image histogram, absolute gradient, gray level co-occurrence matrix, gray level run length matrix, autoregressive model and wavelet transformation. Classification of fractured and remained intact vertebrae from 58 patients using SVM shows an AUC of 0.64. Nonetheless, the methods proposed in both studies required manual operation and did not enable fully automated assessment of fracture risk.

In addition to CT measurements, artificial intelligence techniques have also been applied to other imaging modalities for fracture risk assessment. Ferizi et al. (2019) conducted a comparison of various machine learning algorithms to identify patients who

had suffered from fragility fractures based on magnetic resonance imaging (MRI) data. The study systematically explored the application of artificial intelligence in osteoporosis diagnosis and emphasized image features extracted from MRI images that contributed to predicting fragility fractures. However, it lacked comparison with clinical methods such as DXA measurement and clinical risk assessment. Meanwhile, Kilic and Hosgormez (2016) developed a novel method that leveraged ensemble learning techniques to detect osteoporosis and assist in estimating osteoporotic fractures. Six bone densitometry parameters were extracted from DXA and fed to ensemble classifiers to accurately distinguish between osteoporosis, osteopenia and healthy subjects. This study demonstrated the efficacy of combining multiple learning models to improve accuracy in osteoporosis diagnosis. However, the proposed method did not have the capability to directly identify osteoporotic fracture subjects.

Although deep learning techniques have been applied to fracture risk assessment, their performances largely depend on the availability of sufficient training data (Lee et al. (2020)). Therefore, traditional deep neural networks may not be suitable for some cases, particularly when dealing with small or imbalanced datasets. Texture analysis and few-shot learning, which are suitable for small datasets, have the potential to outperform traditional deep-learning models in our tasks. Therefore, in our study, we adopt these methods to encode image features to characterise bone microarchitecture in HR-pQCT images for fracture discrimination.

2.3.1 Texture Feature Extraction

Texture is an important property of medical images, and texture analysis has been widely used in various medical image tasks (Riaz et al. (2015)), especially when regional textures are the diagnostic criterion for experts. Over the past two decades, texture analysis techniques have been applied to bone imaging, providing a valuable reference for clinical experts.

Texture features extracted from X-ray images have been applied to quantify bone density information. For example, Le Corroller et al. (2012) extracted three bone texture parameters including the fractal parameter Hmean, co-occurrence and run-length matrices from 21 digital X-ray images. These parameters were then combined with BMD measurements to predict fracture load in human femurs. The results demonstrated that the combination of bone texture parameters and BMD measurements significantly outperformed BMD alone in fracture load prediction. Zheng and Makrogiannis (2016) proposed a novel method that computed various texture descriptors such as wavelet and LBP using digital radiography from 116 patients to diagnose osteoporosis and prevent fracture. Feature selection was then employed to find the most discriminant subset, and classification techniques were adopted to separate healthy subjects from osteoporotic

patients. The study showed that texture features provided an opportunity to capture the deterioration of trabecular bone. Nonetheless, these methods that used 2D texture analysis on X-ray images could not capture spatial bone structure information, and there was still room for the improvement of fracture risk assessment.

Texture analysis has therefore been extended for quantitative analysis of 3D clinical imaging of bone. In a study, Valentinitsch et al. (2019) proposed an automatic method that combined global and local texture features extracted from multi-detector CT images to identify osteoporotic vertebral fractures from 154 patients. Histogram of gradients, gray level co-occurrence matrix Haralick, LBP and wavelet transformation were used to extract texture features from CT scans. This study demonstrated that texture information obtained from CT images outperformed vBMD for fracture discrimination. However, it was limited to identifying osteoporotic vertebral fractures and did not consider other types of fractures. In another study, Valentinitsch et al. (2013) developed an algorithm to quantify trabecular microarchitecture characteristics via texture features in HR-pQCT scans. They utilised 3D gray level co-occurrence matrices and partial derivatives to extract texture features from HR-pQCT images of 36 postmenopausal women. This study demonstrated that clustering of trabecular bone by 3D-texture analysis for HR-pQCT images was feasible. Nonetheless, it did not investigate the discriminative performance of texture information obtained from HR-pQCT scans for fracture classification.

2.3.2 Medical Image Segmentation

Image segmentation plays a crucial role in the field of medical imaging, facilitating the localisation of the regions of interest and reducing error rates (Minaee et al. (2021)). Since the introduction of deep convolutional neural networks (CNNs) with powerful feature extraction capabilities, medical image segmentation has witnessed significant progress (Tajbakhsh et al. (2020)).

The U-Net model, initially proposed by Ronneberger et al. (2015), is a popular framework for semantic segmentation. Its main novelty lies in the upsampling operation, which enables the network to propagate contextual information to higher layers, thus reducing information loss. With its advantages of high accuracy and fast training on small datasets, U-Net is widely applied in medical image tasks such as brain tumor and lung lesion segmentation. In recent years, various U-Net variants have been proposed to enhance segmentation performance. Typical representations include attention U-Net, residual U-Net, recurrent U-Net, dense U-Net and U-Net++ (Siddique et al. (2021)).

Several studies have proposed the use of U-Net to segment skeletal regions of interest in clinical imaging to assist expert decision-makers. For instance, Fang et al. (2021) introduced an automatic approach using deep neural networks to analyse multi-detector CT images to identify osteoporosis, osteopenia and normal subjects. The U-Net model was employed to segment the lumbar vertebral body, and the DenseNet-121 was then utilised for calculating BMD via regression prediction. The results demonstrated that the proposed model could provide high accuracy in vertebral body segmentation and generate BMD values highly correlated with those derived from QCT. Similarly, Noguchi et al. (2020) developed a deep neural network based on the U-Net architecture to segment bone regions in whole-body CT images. The incorporation of data augmentation techniques enhanced the accuracy and robustness of the model. The proposed method showed high segmentation accuracy on whole-body CT and exhibited generalisability under different scan conditions.

The relative contributions of cortical and trabecular compartments measured by HR-pQCT to fracture discrimination have not yet been investigated. In Chapter 5, we employ the 2D U-Net model to automatically segment cortical and trabecular regions in HR-pQCT slices. Subsequently, we separately quantify cortical and trabecular bone microarchitecture through statistical distributions.

2.4 Volumetric Texture Analysis

Texture analysis has been an active topic in a wide variety of applications such as face recognition, remote sensing and medical image analysis (Pan et al. (2021)). High discriminative power, strong model robustness and low computational complexity are three crucial properties of efficient texture representations. Over the decades, numerous local and global descriptors have been proposed to characterise the texture information of natural images, in which wavelets and LBP are commonly used.

However, it's important to note that most current texture analysis methods are designed for 2D images, and there are only a few approaches proposed for volumetric texture analysis. Volumetric textures, composed of a series of 2D slices, provide richer object information. However, treating volumetric texture as 2D slices using conventional 2D approaches results in the loss of spatial information and reduced classification accuracy. Therefore, developing volumetric texture analysis methods to characterise 3D images is pretty valuable, but such an extension faces many challenges such as significant increases in computational cost. Among various texture descriptors, LBP has received extensive attention due to its efficiency and low computational complexity (Pan et al. (2021)).

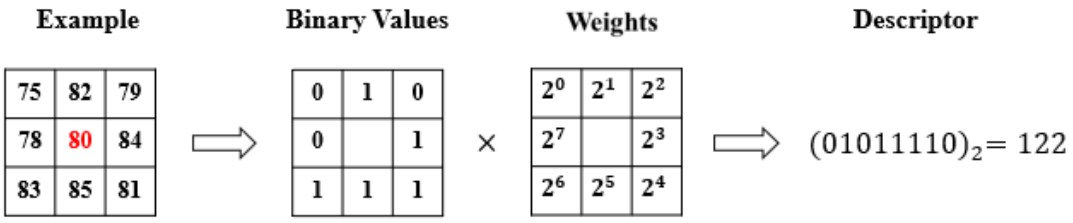

FIGURE 2.3: An example of the original local binary pattern. The centre pixel is compared to its neighboring pixels to generate binary values.

2.4.1 Local Binary Pattern

LBP, invented by Ojala et al. (2002), compares the difference between the centre pixel and its neighbors in a local region to encode texture patterns. Figure 2.3 shows the original LBP operator which works in a 3×3 neighborhood. The centre grayscale value is regarded as the threshold to encode binary codes (0 or 1). Then this LBP code and corresponding weights are converted into a decimal number to represent the local structural information. The formulas are shown as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) \times 2^p \tag{2.2}$$

and

$$s(x) = \begin{cases} 1 & \mathrm{x} \geq 0 \\ 0 & \mathrm{x} < 0 \end{cases} \tag{2.3}$$

where, R represents the radius of the circle neighborhood, and P is the number of sampled points. g_c and g_p correspond to the grayscale values of the centre pixel and the neighbor pixels respectively.

Then the rotation invariant property was considered in the LBP framework. Pietikäinen et al. (2000) first proposed to apply a circular bitwise right shift and search for minimum LBP to achieve rotation invariance:

$$LBP^{ri}_{P,R} = \min\{ROR(LBP_{P,R}, i) \mid i = 0, 1, ..., P-1\} \tag{2.4}$$

where, $ROR(LBP_{P,R}, i)$ is the circular bitwise right shift operation. Superscript ^{ri} represents the use of rotation invariant patterns.

However, this method does not provide an excellent discriminative performance. According to the statistical results of texture patterns, Ojala et al. (2002) found that uniform LBP patterns occurred more frequently than others. To improve rotation invariance, they proposed the $LBP^{riu2}_{P,R}$, which is defined as:

$$LBP^{riu2}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U \leq V \\ P & \text{else } U > V \end{cases} \tag{2.5}$$

and

$$U = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |f(g_p - g_c) - s(g_{p-1} - g_c)| \tag{2.6}$$

where, U represents the measurement that counts the number of bit transitions from 1 to 0 or vice versa. V is the threshold that defines the LBP as "uniform" or "non-uniform" patterns. Superscript ^{riu2} represents the use of rotation invariant "uniform" patterns with V set to 2.

2.4.2 Local Binary Pattern Variants

The original LBP texture descriptor only considers the joint distribution of the neighborhood, and other information is discarded. To address this issue, researchers have suggested using a combination of multiple detectors to enhance the discriminative capability of LBP. Over the last two decades, various LBP variants have been proposed (Li et al. (2011)).

Guo et al. (2010) proposed a completed local binary pattern (CLBP) model that consisted of the centre pixel and a local difference sign-magnitude transform. The centre pixel, namely CLBP_C, represents the image grayscale level, which is converted to a binary code through global thresholding. The local differences are decomposed into the sign operator CLBP_S and the magnitude operator CLBP_M (See Figure 2.4). These three operators are defined as:

$$CLBP\_C = s(g_c - m) \tag{2.7}$$

$$CLBP\_S = \sum_{p=0}^{P-1} s(g_p - g_c) \times 2^p \tag{2.8}$$

$$CLBP\_M = \sum_{p=0}^{P-1} s(m_p - n) \times 2^p \tag{2.9}$$

with

$$m_p = |g_p - g_c| \tag{2.10}$$

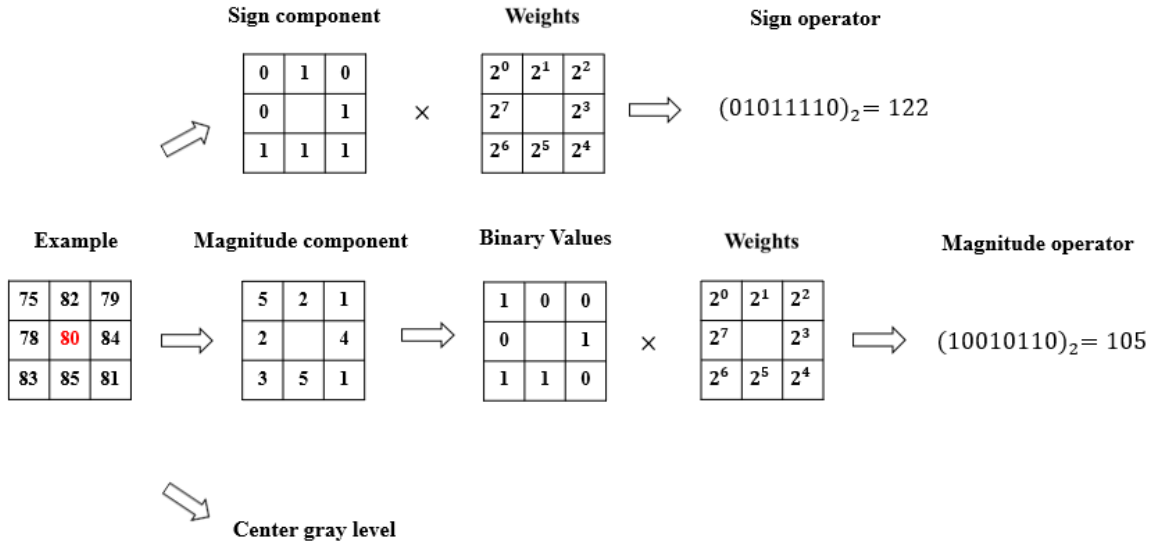

FIGURE 2.4: An example to illustrate the completed local binary pattern. The sign operator equals the original LBP, while the magnitude and centre pixel operators are incorporated to enrich the texture representation.

where, m denotes the threshold, set as the average grayscale value of the entire image. n represents the mean value of m_p from the entire image. g_c, g_p and P are defined as in Equation 2.2.

The CLBP_C, CLBP_S and CLBP_M operators are combined to form the feature map, and a histogram is then constructed to represent the texture image. Unlike the original LBP descriptor, which only contains the sign component, the inclusion of the magnitude component and image intensity provides efficient contrast information and enhances the discriminative power.

Similarly, Liu et al. (2012) developed an extended local binary pattern (ELBP) descriptor that consisted of two intensity-based operators and two difference-based operators to generalise the conventional LBP. The intensity-based features represent the intensity of the centre pixel (CI-LBP) and the neighbor pixels (NI-LBP), while the difference-based features consider the radial-difference (RD-LBP) and the angular-difference (AD-LBP). The CI-LBP, NI-LBP, RD-LBP and AD-LBP descriptors are defined as follows:

$$CI - LBP = s(g_c - m) \tag{2.11}$$

$$NI - LBP = \sum_{p=0}^{P-1} s(g_p - \frac{\sum_{p=0}^{P-1} g_p}{P}) \times 2^p \tag{2.12}$$

$$RD - LBP = \sum_{p=0}^{P-1} s(g_p^R - g_p^{R-1}) \times 2^p \tag{2.13}$$

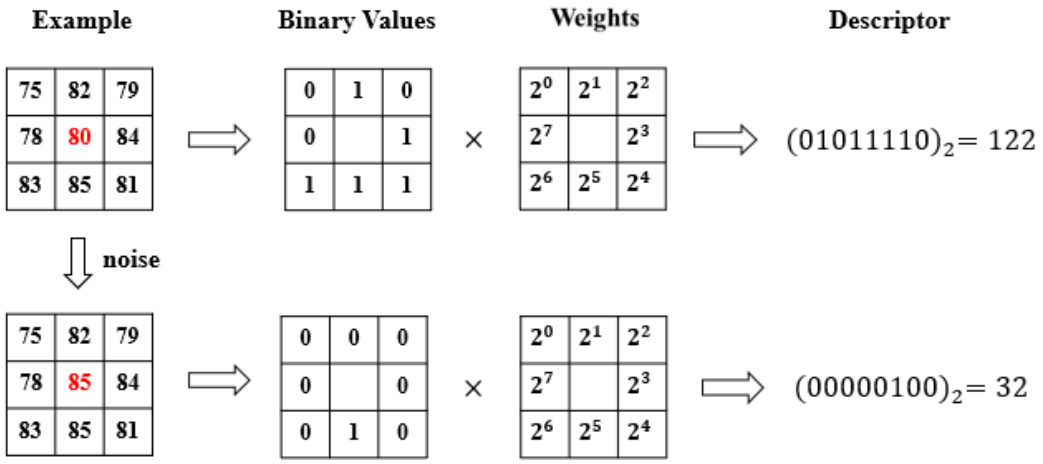

FIGURE 2.5: An example that local binary pattern is susceptible to noise. The centre pixel value is used as a threshold to generate binary values and encode the descriptor.

$$AD - LBP = \sum_{p=0}^{P-1} s(g_p - g_{mod(p+d,P-1)}) \times 2^p \tag{2.14}$$

where, g_c, g_p, m, r and P are defined as in Equation 2.2 and Equation 2.7. d is an integer, $d \in [1, P/2]$. The function $mod(x, y)$ represents the modulus x of y.

The histograms of these four LBP operators are calculated separately and then concatenated into a joint histogram to represent the image. Combining these four types of features can extract complementary texture information of local regions and enhance the classification performance of LBP.

2.4.3 Robust Local Binary Pattern

The original LBP is highly susceptible to noise in the image due to its thresholding mechanism. As illustrated in Figure 2.5, even a minor change in the centre pixel value (from 80 to 85) significantly affects the LBP value. To enhance its noise tolerance, several robust LBP frameworks have been proposed (Guo et al. (2015)). For example, Zhao et al. (2013) proposed the robust local binary pattern (RLBP) descriptor that replaced the grayscale value of the centre pixel with its average local grayscale value to reduce noise. The $RLBP_{P,R}$ is presented as:

$$RLBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - \frac{\sum_{p=0}^{P-1} g_p + g_c}{P+1}) \times 2^p \tag{2.15}$$

where, g_c, g_p, P and R are defined as in Equation 2.2.

Moreover, some other approaches have been developed to enhance model robustness

against noise (Qiang et al. (2021)). Tabatabaei and Chalechale (2020) presented a directional thresholded local binary pattern (DTLBP) that replaced the centre pixel value with the average values of directional neighboring pixels. The information of neighboring pixels was used to reduce noise and enhance the discriminative capacity of the texture descriptor. The proposed method demonstrated high classification accuracy both in noisy and noise-free images. Analogously, Liu et al. (2016) proposed a median robust extended local binary pattern (MRELBP) that compared regional image medians instead of raw image intensities to enhance noise robustness. A novel sampling scheme was introduced to capture both microstructure and macrostructure texture information. Experimental results on benchmark datasets showed that the proposed method was highly robust to noise and outperformed other state-of-the-art LBP variants in texture classification. However, these methods were limited to the 2D LBP framework.

2.4.4 3D Local Binary Pattern

Zhao and Pietikäinen (2006) first introduced the concept of 3D LBP and proposed the volume local binary pattern (VLBP) to extract texture features in a local neighborhood of the centre volume. Figure 2.6 illustrates the generation process of the VLBP texture descriptor. Its key point is to stack several consecutive frames and select neighboring points to encode binary codes. The $VLBP_{Q,R,L}$ is defined as follows:

$$VLBP_{Q,R,L} = \sum_{q=0}^{3\times Q+1} (v_q - v_c) \times 2^q \tag{2.16}$$

where, v_c and v_q represent the grayscale values of the centre voxel and the neighbor voxels respectively. Q is the number of sampled neighbors on a circle of radius R in consecutive L frames.

LBP faces the challenge of high computational cost in 3D space due to the large number of texture units. Compared to other 3D LBP variants based on the sphere, the VLBP texture descriptor, without interpolation operation, has a lower computational cost while preserving raw image information.

LBP has not yet been applied to analyse bone microarchitecture measured by HR-pQCT. In Chapter 4, we develop a 3D LBP method to characterise bone microarchitecture in HR-pQCT images for automated fracture discrimination. In Chapter 5 and Appendix C, we present a robust framework of 3D LBP to enhance the noise tolerance of our discriminative system.

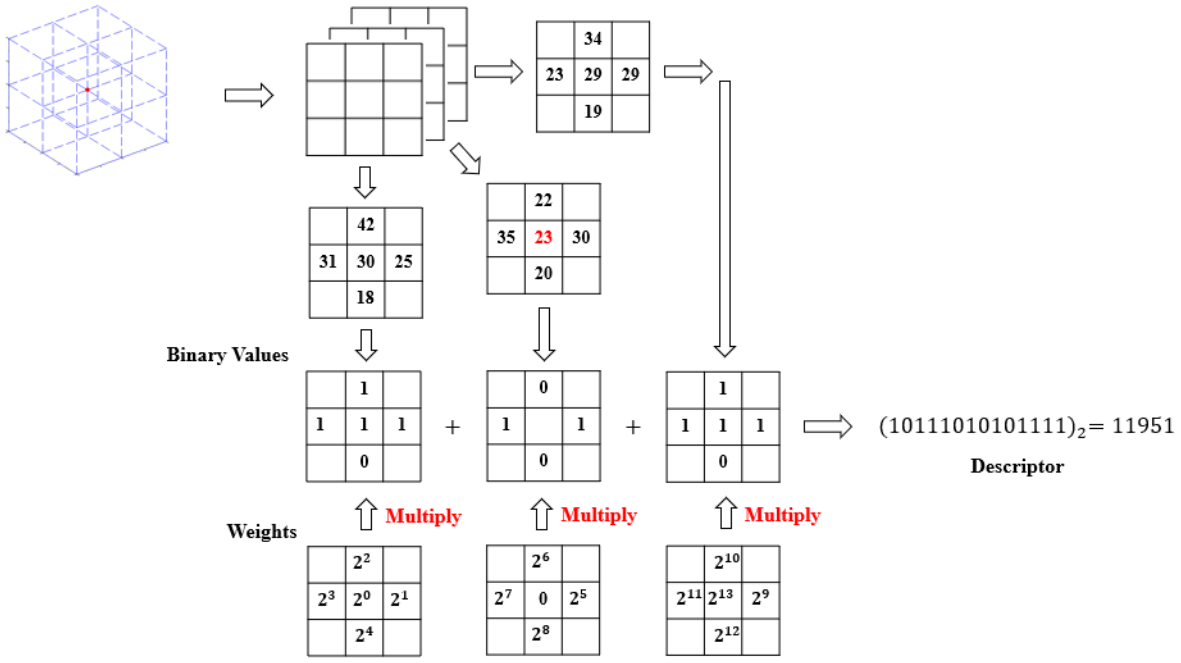

FIGURE 2.6: The generation process of the volume local binary pattern. Neighbors in the volume are sampled to encode the binary values with the centre voxel value as the threshold.

2.5 Convolutional Neural Networks

CNN is one of the most popular deep learning architectures, and has been successfully used in various computer vision tasks such as image classification, segmentation, registration and object detection. The success of CNN can be primarily attributed to its capability to automatically learn hierarchical features from low-level to high-level abstractions (Yan et al. (2015)). Moreover, the use of local receptive fields and weight sharing enables it to effectively capture intricate patterns in data and optimize parameter efficiency.

Initially, LeCun et al. (1998) proposed the LeNet-5, which demonstrated great success in handwritten digit recognition. Then the deep CNN architecture AlexNet was presented by Krizhevsky et al. (2012). This innovative work significantly outperformed traditional methods and established a new benchmark in image recognition. Subsequent architectures such as VGGNet, GoogLeNet and ResNet, further promoted the development of CNNs. These architectures built upon earlier models, introducing deeper network structures and innovative architectural designs to enhance feature representation.

Figure 2.7 illustrates the basic architecture of the CNN, which comprises convolutional layers, pooling layers and fully connected layers (Albelwi and Mahmood (2017)). The

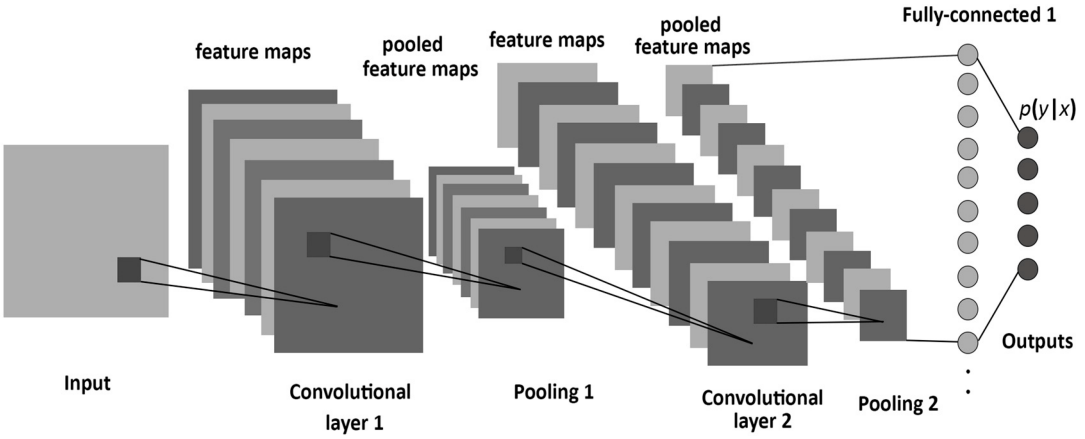

FIGURE 2.7: Basic architecture of the convolutional neural network. Alternating convolutional layers and pooling layers are followed by the fully connected layer to learn hierarchical representations of the input data. Source: https://www.mdpi.com/1099-4300/19/6/242

convolutional layer uses multiple filters to learn various features from the data in parallel. Subsequently, the pooling layer downsamples the feature map to reduce the computational complexity while retaining important information. Finally, the fully connected layer connects each neuron in the layer to every neuron in the previous layer to integrate global information and produce the final output.

Over the past decade, CNNs have demonstrated remarkable success in medical image analysis tasks such as disease diagnosis and lesion localisation (Criminisi (2016)). With powerful feature extraction capabilities, CNNs can automatically learn and identify pathological features in medical images that might be challenging via visual inspection. This assists clinicians in enhancing the accuracy and efficiency of medical decision-making.

Nevertheless, CNNs encounter some challenges (Litjens et al. (2017)). One of the most concerning issues is that CNNs require a large amount of training data to prevent overfitting. In recent years, the complex architecture and massive parameters of CNN models have increased their reliance on rich and diverse data. To address this issue, several strategies have been proposed (Han et al. (2018)). Transfer learning with pre-trained CNN models has become a prevalent strategy. This approach leverages knowledge gained from large datasets to enhance the classification performance of prediction models on limited data. Another strategy is data augmentation, which increases the size and diversity of training data through transformations such as rotation, flipping and scaling. This approach enables CNNs to learn a broader range of patterns from data and enhance model generalisability and robustness.

CNNs have not yet been applied to analyse bone microarchitecture measured by HR-pQCT. In Chapter 6, we propose a deep learning framework to automatically extract

bone microarchitectural features from HR-pQCT images to discriminate previous fractures.

2.6 Summary

This chapter gives an overview of osteoporosis and fracture risk, methods to measure bone health and artificial intelligence techniques. Current fracture risk prediction models primarily focus on DXA measurements, which have some limitations. Although HR-pQCT provides a more detailed bone microarchitecture in three dimensions, there is no automated tool available for bone HR-pQCT images to measure fracture risk. Volumetric texture analysis and deep learning techniques have the potential to automatically identify bone fragility from HR-pQCT images and to improve fracture discrimination provided by DXA and clinical risk factors.

Chapter 3

Datasets

3.1 Study Design

Our research project primarily focuses on the Hertfordshire Cohort Study (HCS). The HCS is an extraordinary and nationally unique research initiative of 3225 participants who were born in the English county of Hertfordshire between 1931 and 1939 and were still resident there at the end of the 20th century (Syddall et al. (2019)). This longitudinal study primarily focuses on investigating the relationship between early life factors, ageing and chronic disease risk in later life, and has provided valuable insights into the understanding of age-related diseases (Clynes et al. (2014)). The HCS was supported by the Medical Research Council, British Heart Foundation, Versus Arthritis UK, International Osteoporosis Foundation, NIHR Southampton Biomedical Research Centre, NIHR Oxford Biomedical Research Centre, and the University of Southampton.

The HCS is a comprehensive research endeavor that encompasses a wide range of chronic diseases related to the musculoskeletal system, including sarcopenia, osteoporosis, osteoarthritis and metabolic syndrome. In particular, the study has been a pioneering investigation into bone health, exploring risk factors influencing skeletal well-being throughout the life course. It revealed that early-life factors such as birth weight and childhood nutrition played important roles in determining BMD and later fracture risk (Dennison et al. (2005)). Lifestyle choices, especially physical activity, were identified as crucial determinants of bone density, while smoking and alcohol consumption were linked to lower bone density (Moinuddin et al. (2008)). In addition, the study reported genetic contributions to individual variations in BMD and fracture susceptibility (Lips et al. (2007)). These findings underscore the significance of early-life interventions, healthy lifestyle choices and personalized approaches for improving bone health and preventing age-related musculoskeletal disorders.

The expanded objective of this study is to investigate the influence of bone microarchitecture on fracture risk. To achieve this goal, our research characterises bone microarchitecture in HR-pQCT images using computer vision techniques and develops automatic approaches for fracture discrimination.

The Global Longitudinal Study of Osteoporosis in Women (GLOW) is a comprehensive and prospective research initiative designed to delve into the intricacies of osteoporosis in women (Hooven et al. (2009)). Globally, GLOW enrolled 60,393 women aged 55 years and older through 723 physicians in 10 countries in Australia, Europe and North America, with an annual follow-up conducted for up to 5 years. This wide-ranging enrollment not only ensures the representation of diverse populations but also facilitates the exploration of regional differences in osteoporosis prevalence, risk factors and fracture outcomes. Our study utilises the GLOW dataset from an independent cohort as testing data to evaluate the model robustness of our discriminative systems.

TABLE 3.1: Participant characteristics of the Hertfordshire Cohort Study.

Characteristics	Mean (std)	Number of non-missing subjects
Age (years)	76.3 (2.6)	376
Height (cm)	167.0 (9.2)	376
Weight (kg)	77.0 (13.6)	376
BMI (kg/m^2)	27.6 (4.2)	376
Dietary calcium (g)	8.3 (2.4)	376
Physical activity time (minutes)	212.2 (116.2)	325
Number of comorbidities	1.6 (1.4)	350
Femoral neck BMD (g/cm^2)	0.9 (0.1)	361
Characteristics	**Number (%)**	**Number of non-missing subjects**
Men	198 (52.7%)	376
Women	178 (47.3%)	376
Smoking history (ever)	171 (48.9%)	350
High alcohol consumption	32 (9.1%)	350
Bisphosphonate (since baseline)	37 (10.6%)	350
Social class (manual)	207 (55.1%)	376
Previous fracture	97 (28.1%)	345

High alcohol consumption (per week): more than 21 units for men and more than 14 units for women.
Social class was categorised into the manual group (classes IIIM, IV and V) and the non-manual group (classes I, II and IIINM).
Previous fracture: the participant had a vertebral fracture or a self-reported fracture.
Std: standard deviations.
Non-missing subject: the subject has no missing value for that characteristic.

3.2 Data Collection

3.2.1 Hertfordshire Cohort Study

The baseline recruitment for the HCS took place between 1998 and 2004 (Syddall et al. (2019)). 2997 participants in Hertfordshire agreed to a home interview, and subsequently attended a clinic for a detailed health survey. In 2011–2012, 376 participants agreed to take part in a further follow-up study, and DXA and HR-pQCT scans were taken at this time point. The data used in our study was part of the HCS, which included HR-pQCT scans of the radius and tibia, clinical covariates, femoral neck BMD values and fracture history of participants, as shown in Table 3.1. The case index of subjects included in our study is provided by Table A.1 in Appendix A. This research was approved by the Hertfordshire and Bedfordshire Local Research Ethics Committee and the East and North Hertfordshire Ethical Committees. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The written informed consent was given and signed by all subjects.

HR-pQCT scans (XtremeCTi Scanco Medical AG, Switzerland) of the non-dominant distal radius and tibia were performed. Dominant limbs were scanned if the non-dominant limb had fractured. 110 parallel CT slices were obtained, representing a volume of bone 9 mm in the z-axis, with a nominal resolution (voxel size) of 82 μm in both the x and y axes. The scan protocol was in accordance with manufacturer's guidelines and as described by Boutroy et al. (2005). Using the method of Pauchard et al. (2012), scans with excessive motion artefact (grade 5) were excluded. Manufacturer standard evaluation and cortical porosity scripts were used for image analysis. Representative 2D slices of the distal radius and tibia are shown in Figure 3.1.

Clinical covariates of participants include age, sex, height, weight, BMI, dietary calcium, smoking history, alcohol consumption, physical activity, bisphosphonate usage, number of comorbidities and occupational social class. Among them, dietary calcium, time since menopause and occupational social class were ascertained through a home interview at baseline recruitment (1998-2004), while all other variables were ascertained in 2011–2012. Height (cm) was measured using a wall-mounted SECA stadiometer along with weight (kg) using calibrated SECA 770 digital floor scales (SECA Ltd, Hamburg). These measurements were then used to derive BMI (kg/m^2) (Nuttall (2015)). Social class was coded from the 1990 OPCS Standard Occupational Classification unit group for occupation (of Population Censuses and Surveys (1995)). More detail about data acquisition has been described in Syddall et al. (2005) and Syddall et al. (2019).

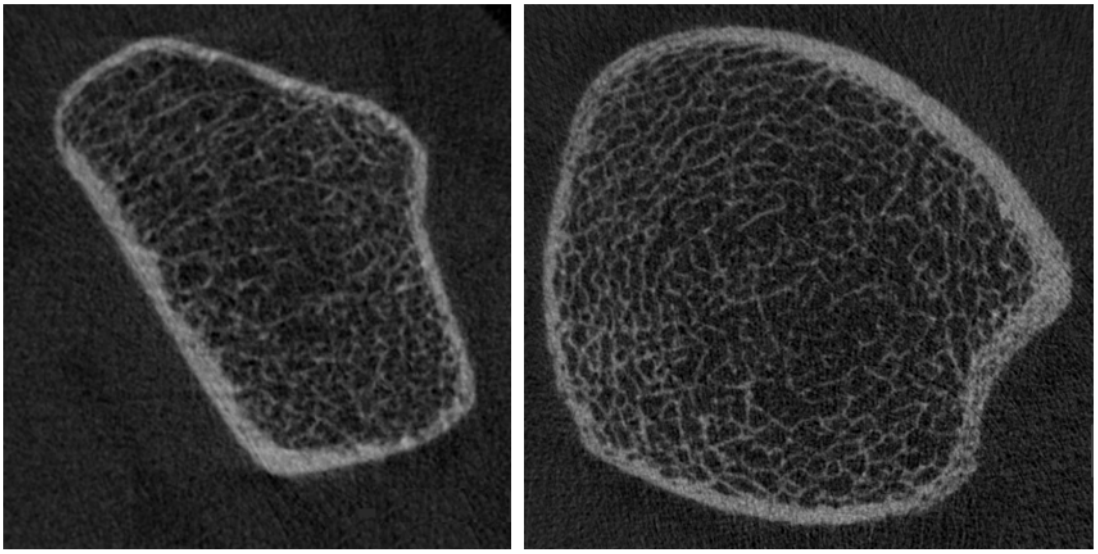

FIGURE 3.1: Representative 2D slices taken from the HR-pQCT scans of the non-dominant distal radius (a) and tibia (b). The radius is positioned in the forearm, while the tibia is situated in the lower leg.

Total hip and femoral neck BMD (g/cm^2) were measured for each participant following conventional procedures. Both the left and right sides of the hip were scanned, and the minimum femoral neck BMD value between the two sides was used for analysis. The T-score was derived using NHANES III data and Equation 2.1 in Chapter 2 (Ward et al. (2023)). The scans were performed using a Lunar Prodigy Advance DXA scanner manufactured by GE Medical Systems (Westbury et al. (2019)). A representative DXA image of the hip is depicted in Figure 3.2.

Fracture history was determined via self-report and vertebral fracture assessment. Participants with a vertebral or self-reported fracture were regarded as having had a previous fracture. Fractures since aged 45 years were ascertained through self-report. Morphometric vertebral fractures were diagnosed from a lateral spine view imaged using a Lunar Prodigy Advance DXA scanner (GE Medical Systems) and graded according to the Genant semi-quantitative method (Genant et al. (1993)).

3.2.2 Global Longitudinal Study of Osteoporosis in Women

The baseline surveys for the GLOW were distributed via mail to more than 140000 potential subjects between October 2006 and February 2008 (Hooven et al. (2009)). After excluding 3,265 individuals who were ineligible or had died, 60,393 subjects agreed to participate in the study, and were followed annually for up to 5 years. After completing 5 years of follow-up, 1367 participants with baseline data and follow-up questionnaires in Southampton were invited for a follow-up study. Between April 2014 and

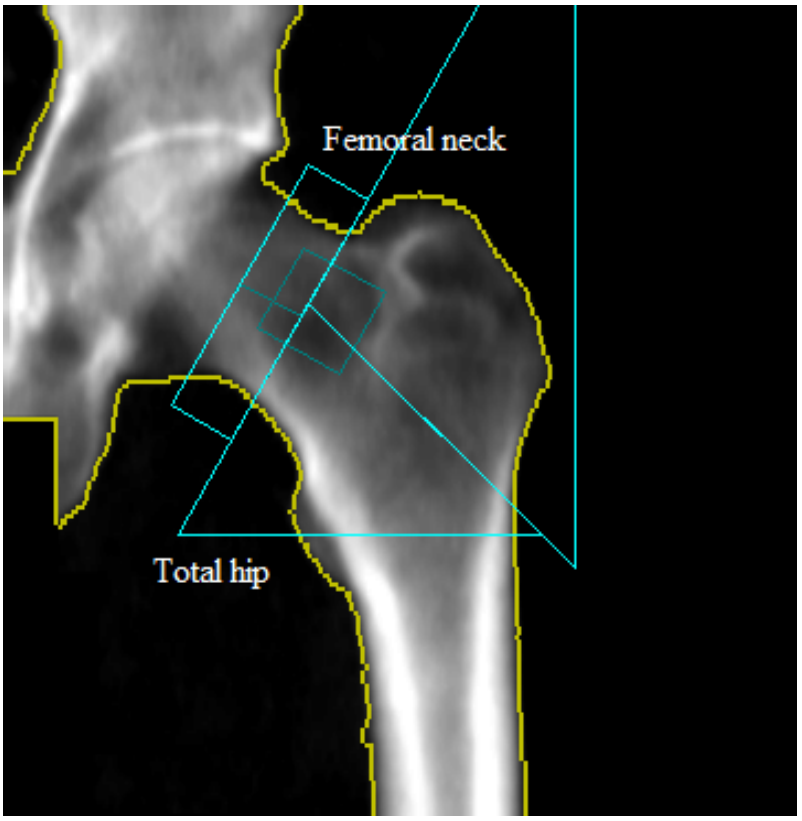

FIGURE 3.2: A representative DXA image of the hip. The total hip refers to the area within the blue lines, while the femoral neck corresponds to the area the diameter across the smallest part of the top rectangle.

December 2017, participants were scanned with HR-pQCT and DXA. The data used in our research work was part of the collected data from Southampton, which included HR-pQCT scans of the radius and tibia, clinical risk factors, femoral neck BMD values and fracture history of participants, as shown in Table 3.2. The case index of subjects included in our study is provided by Table A.2 in Appendix A. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The written informed consent was given and signed by all subjects.

HR-pQCT scans (XtremeCTi Scanco Medical, Basserdorf, Switzerland) of the non-dominant distal radius and tibia were performed (Cappelle et al. (2021)). If there was a fracture on the non-dominant limb, the non-fractured limbs were scanned. 110 parallel HRpQCT slices were acquired with a nominal resolution (voxel size) of 82 μm in both the x and y axes and 9 mm in the z-axis, according to the manufacturer's guideline. The motion artefact of each scan was assessed on a scale of 1 (no artefact) to 5, and scans rated as Grade 5 were excluded. Manufacturer standard evaluation and Image Processing Language (Version 6.1, ScancoMedical) were carried out for image analysis.

Clinical risk factors include age, sex, BMI, current smoker, alcoholic drinks, cortisone or prednisone usage, rheumatoid history, colitis history, diabetes history, coeliac history and premature menopause. Participant information was collected through self-administered questionnaires (Litwic et al. (2021)).

TABLE 3.2: Participant characteristics of the Global Longitudinal Study of Osteoporosis in Women.

Characteristics	Mean (std)	Number of non-missing subjects
Age (years)	70.9 (5.5)	501
BMI (kg/m^2)	26.8 (5.0)	491
Alcoholic drinks per week (levels)	1.2 (1.0)	493
Femoral neck BMD (g/cm^2)	0.7 (0.1)	466
Characteristics	**Number (%)**	**Number of non-missing subjects**
Women	501 (100.0%)	501
Current smoker	30 (6.1%)	493
Cortisone or prednisone usage	10 (2.0%)	493
Rheumatoid history	47 (9.8%)	481
Colitis history	10 (2.0%)	490
Diabetes history	10 (2.1%)	486
Coeliac history	7 (1.4%)	492
Premature menopause	47 (9.7%)	486
Previous fracture	106 (22.8%)	464

Alcoholic drinks per week were categorised into five levels: level 0 (none), level 1 (1-6 units), level 2 (7-13 units), level 3 (14-20 units) and level 4 (more than 20 units).
Previous fracture: the participant had a self-reported fracture.
Std: standard deviations.
Non-missing subject: the subject has no missing value for that characteristic.

Total hip, lumbar spine and femoral neck BMD (g/cm^2) were measured for each participant using DXA Hologic Horizon W (Litwic et al. (2021)). The left side of the hip was scanned; the right side was only scanned if there was a hip replacement or a fracture on the left side. The femoral neck BMD value was used for analysis, and the corresponding T-score was derived using NHANES III data and Equation 2.1 in Chapter 2 (Ward et al. (2023)).

Fracture history was obtained at baseline, and further information on fractures was ascertained after 5-year follow-up. Participants from age 45 to 5-year follow-up with a self-reported fracture were regarded as having had a previous fracture.

3.3 Data Processing

The raw data collected from the HCS and GLOW cohorts, including tibial HR-pQCT scans of participants and their corresponding clinical covariates, DXA-measured femoral neck BMD and fracture history, are used for fracture analysis. We process the raw data from both cohorts to construct the HCS and GLOW datasets.

The raw HR-pQCT scans from the HCS and GLOW contain 110 parallel slices, in which a volume of bone with a nominal resolution of 82 um. Although each HR-pQCT image is in high resolution 1536×1536, the limb in the image is very tiny and most regions actually are soft tissues. To optimize the memory usage and computational cost, the KHKs MicroCT tool is used to crop the radius and tibia in the image and remove the fibula and surrounding soft tissues (Kilic and Hosgormez (2016)). The preprocessed HR-pQCT images are archived in HDF5 format for further analysis. Typical examples of tibial CT slices from participants with and without previous fractures are shown in Figure 3.3.

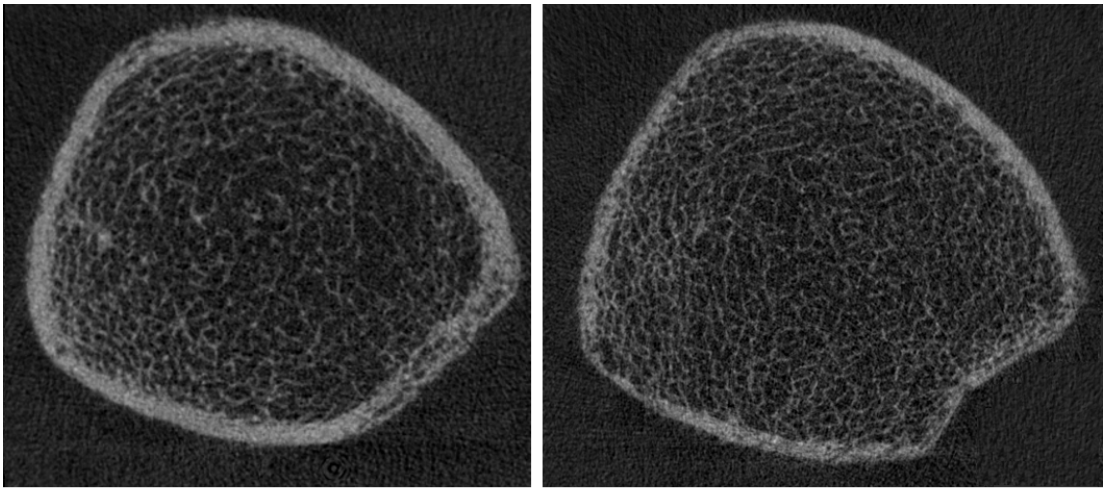

FIGURE 3.3: Typical examples of tibial HR-pQCT slices from fracture (a) and non-fracture (b) cases. There is a difference in bone microarchitecture between (a) and (b).

3.3.1 The HCS and Segmentation Datasets

In the HCS, no subject has more than one tibial HR-pQCT scan, and 193 subjects have tibial HR-pQCT scans. Referring to the method of Pauchard et al. (2012), two tibial CT scans of quality five are excluded, and only the scans of quality four and above are included. 24 subjects who lack DXA-measured BMD or fracture history are also excluded. Finally, the HCS dataset includes 167 subjects with tibial HR-pQCT images, DXA-measured BMD, clinical risk factors and fracture history.

TABLE 3.3: Participant characteristics of the Hertfordshire Cohort Study for fracture analysis.

Characteristics	Mean (standard deviations)	
	Men (n=86)	Women (n=81)
Age (years)	76.0 (2.3)	76.5 (2.7)
Height (cm)	174.4 (6.6)	160.8 (5.4)
Weight (kg)	84.3 (13.8)	72.6 (13.4)
BMI (kg/m^2)	27.7 (4.1)	28.1 (5.0)
Dietary calcium (g)	8.6 (2.1)	7.8 (2.2)
Physical activity time (minutes)	206.9 (110.0)	215.5 (120.8)
Number of comorbidities	1.5 (1.1)	1.7 (1.5)
Femoral neck BMD (g/cm^2)	0.9 (0.1)	0.8 (0.1)

Characteristics	Number (%)	
	Men (n=86)	Women (n=81)
Smoking history (ever)	45 (52.3%)	32 (39.5%)
High alcohol consumption	10 (11.6%)	1 (1.2%)
Bisphosphonate (since baseline)	4 (4.7%)	16 (19.8%)
Social class (manual)	43 (50.0%)	46 (56.8%)
Previous fracture	23 (26.7%)	23 (28.4%)

High alcohol consumption was defined as drinking more than 21 units per week for men and more than 14 units per week for women.
Participants with a vertebral fracture or a self-reported fracture since age 45 years were regarded as having had a previous fracture.

Participant characteristics of the HCS dataset are illustrated in Table 3.3. Overall, 86 males and 81 females with detailed previous fracture status are included; all participants are over 72 years old. Women have lower height, weight, dietary calcium intake, femoral neck BMD values, smoking and alcohol consumption compared to men. However, women tend to have higher age, BMI, physical activity, comorbidities, bisphosphonate usage and social class compared to men. In addition, fractures occur more frequently in older women, especially after menopause, consistent with previous studies (Areeckal et al. (2018)).

A segmentation dataset with pixel-level annotations in CT slices is constructed. The open annotation tool LabelMe is used to manually mark the position of various regions in a series of HR-pQCT slices and generate corresponding pixel-level labels (Russell et al. (2008)). Figure 3.4 illustrates the surrounding soft tissue, cortical and trabecular regions in the CT transverse slice. Within the HCS cohort, 24 subjects lack a fracture history or DXA-measured BMD, and their tibial HR-pQCT scans are used to construct the segmentation dataset. A total of 30 tibial CT scans are manually marked with pixel-level annotations, and each HR-pQCT image contains 110 parallel slices. Therefore, 3300 CT transverse slices with pixel-level annotations are produced.

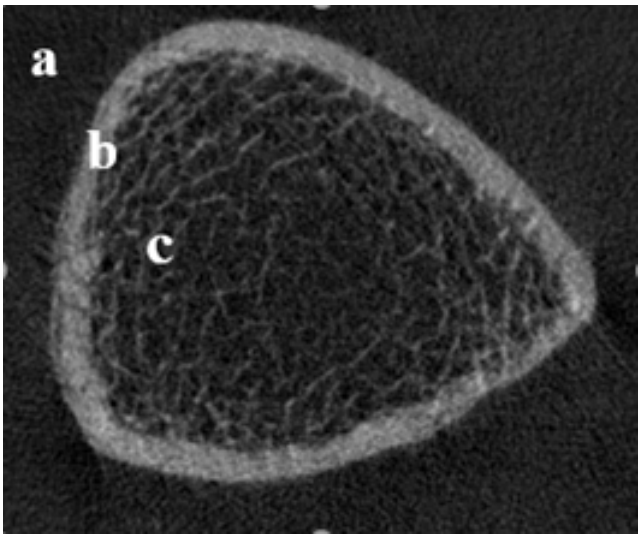

FIGURE 3.4: Cross section of bone taken from a tibial CT image, showing different regions: surrounding soft tissue (a), cortex (b) and trabecular (c).

TABLE 3.4: Participant characteristics of the Global Longitudinal Study of Osteoporosis in Women for fracture analysis.

Characteristics	Mean (standard deviations) Women (n=381)
Age (years)	70.9 (5.5)
BMI (kg/m^2)	26.7 (5.0)
Alcoholic drinks per week (levels)	1.2 (1.0)
Femoral neck BMD (g/cm^2)	0.7 (0.1)
Characteristics	**Number (%) Women (n=381)**
Current smoker	20 (5.2%)
Cortisone or prednisone usage	6 (1.6%)
Rheumatoid history	36 (9.4%)
Colitis history	7 (1.8%)
Diabetes history	9 (2.4%)
Coeliac history	6 (1.6%)
Premature menopause	35 (9.2%)
Previous fracture	84 (22.0%)

Alcoholic drinks per week were categorised into five levels: level 0 (none), level 1 (1-6 units), level 2 (7-13 units), level 3 (14-20 units) and level 4 (more than 20 units).
Previous fracture: the participant had a self-reported fracture.

3.3.2 The GLOW Dataset

In the GLOW, no subject has more than one tibial HR-pQCT scan, and our study includes tibial HR-pQCT scans from all subjects. 15 tibial CT scans are excluded either due to poor scan quality or the absence of scan quality (Pauchard et al. (2012)). In addition, 105 subjects who lack DXA-measured BMD or fracture history are excluded.

Finally, the GLOW dataset comprises 381 subjects with tibial HR-pQCT images, clinical risk factors, DXA-measured BMD and fracture history.

Participant characteristics of the GLOW dataset are presented in Table 3.4. Overall, the GLOW dataset includes 381 females with detailed previous fracture status, and all participants are over 62 years old.

3.4 Summary

In this chapter, we provide an overall review of the HCS and extend its scope to investigate the relationship between bone microarchitecture and fracture risk. In addition, we present details of data collection and data processing for the HCS and GLOW cohorts, including HR-pQCT scans of the distal tibia, clinical risk factors, DXA-measured BMD values and fracture history of participants. Our study utilises the HCS and GLOW datasets to perform fracture analysis and derive meaningful findings.

Chapter 4

Volumetric Texture Analysis for Fracture Discrimination

4.1 Inspirations and Introduction

HR-pQCT is a 3D imaging modality capable of assessing vBMD and bone microarchitecture (Mikolajewicz et al. (2020)). Compared to DXA and clinical risk assessment methods, HR-pQCT has the potential to provide higher accuracy in predicting fracture risk, especially in patients with deterioration in bone microarchitecture. However, traditional analysis of HR-pQCT imaging requires manual operation, and HR-pQCT interpretation remains a challenge. A computer vision approach to ascertain fracture risk from CT scans would be far simpler.

In this chapter, we propose a method that automatically extracts texture features from HR-pQCT images and exploits the random forest classifier to identify previous fractures. Our research objective is to deploy a computer vision approach to HR-pQCT images in order to predict those at risk of fracture and to compare the discriminative performance of this approach against the traditional methods of clinical risk factors and femoral neck BMD. The study of this chapter is nested in the HCS, a group of community-dwelling older adults. The results of our work demonstrate that using a computer vision method to HR-pQCT scanning improves fracture discrimination compared to clinical risk factors and DXA-measured BMD. This approach has the potential to make the information offered by HR-pQCT more accessible (and therefore) applicable to healthcare professionals in the clinic if the technology becomes more widely available.

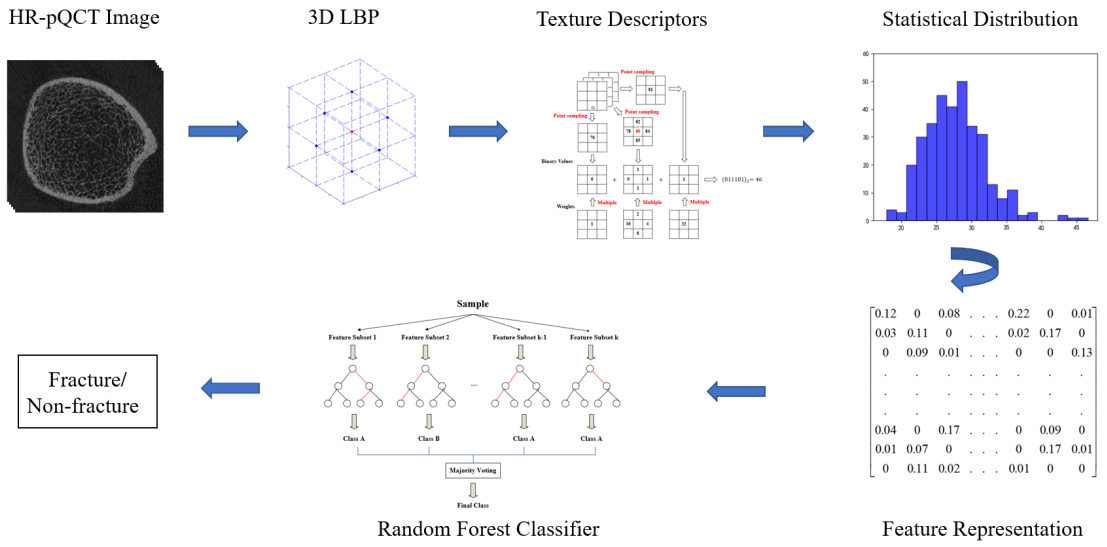

FIGURE 4.1: A combination of volumetric texture analysis and machine learning for fracture classification. Texture information obtained from HR-pQCT images is used to characterise bone microarchitecture.

4.2 Method

4.2.1 Design and Overview

The present study here is designed to test the hypothesis that fracture risk is determined in part by texture representations of HR-pQCT images. Therefore, if texture features are derived from tibial HR-pQCT images of individuals, those with previous fractures can be identified.

The framework of our method is illustrated in Figure 4.2. We propose a 3D LBP model to characterise texture patterns of HR-pQCT images and to quantify bone microarchitecture through statistical distributions. These texture features extracted from HR-pQCT images are then fed into the random forest classifier to distinguish between subjects with and without previous fractures. The HCS dataset is used to evaluate the discriminative performance of our method. Data collection and data processing are detailed in Chapter 3.

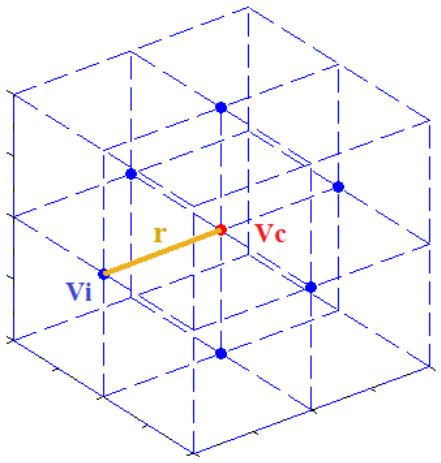

FIGURE 4.2: A centre volume and its surrounding voxels. The texture descriptor is encoded by comparing the difference between the centre voxel v_c and sampled points v_i in the neighborhood.

4.2.2 3D Texture Feature Extraction

As depicted in Figure 3.3, there are differences in bone microarchitecture in HR-pQCT images between typical fracture and non-fracture subjects. Since texture features capture spatial patterns and statistical properties of pixel intensities within a local neighborhood, they have the potential to quantify bone microarchitecture in HR-pQCT images to automatically ascertain fracture risk. Inspired by the successful application of LBP in many tasks (Tang et al. (2013)), we extend the LBP method to extract volumetric texture features from HR-pQCT images. Several 3D rotation invariant LBP descriptors have been developed (Citraro et al. (2017), Fehr and Burkhardt (2008)), which construct a sphere through linear interpolation and automatically adjust the angle to encode texture patterns. However, this process destroys the raw bone microarchitecture to a certain extent and also leads to high computational costs. For these defects, we develop a 3D LBP texture descriptor to characterise bone microarchitecture for fracture discrimination.

$\Omega_v = \{v_{i,j,k} | 1 \leq i \leq W,\ 1 \leq j \leq H,\ 1 \leq k \leq Z\}$ represents a set of points in a $W \times H \times Z$ image. Based on each voxel $v_{i,j,k}$ in the image, we construct a cube with the size of $E \times E \times E$ to encode the local patterns. We let r denote the Euclidean distance between the centre location v_c and sampled points v_i on the cube surface. The configuration of voxels in the neighborhood is shown in Figure 4.2. As observed in this figure, the sampled voxels are positioned at the centre on each side of the neighborhood cube and are shown in bold points. For each voxel v_c in the image, v_i represents the sampled neighboring point of v_c. Therefore, the 3D LBP descriptor can be encoded as in the following

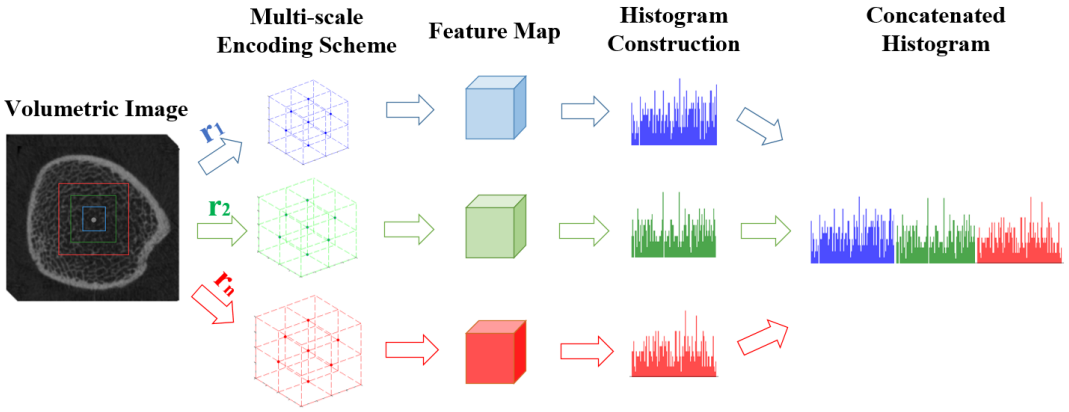

FIGURE 4.3: Multi-scale 3D local binary pattern descriptors. Texture information from various scales is integrated to enrich feature representation.

equation:

$$LBP_{p,r} = \sum_{i=0}^{p-1} F(v_i - v_c) \times w_i \tag{4.1}$$

$$F(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{4.2}$$

where, p is the number of sampled points around the centre voxel, and w_i denotes the weight parameter given to the corresponding neighborhood voxel v_i. We give a threshold to the $E \times E \times E$ neighbourhood voxels in accordance with the value of v_c to encode binary codes and calculate the LBP operator (Ojala et al. (2002)).

Multi-scale feature extraction, which contains the descriptors from various scales, can capture richer information about local patterns and increase classification performance (Ojala et al. (2002)). In this study, we extend the multi-scale strategy to 3D space and propose the multi-scale 3D LBP texture descriptor. Through extensive experiments, we employ LBP descriptors from three different scales, as detailed in Section 4.3.2. The multi-scale 3D LBP framework is illustrated in Figure 4.3. We construct multiple cubes for each voxel in the solid image, and select its neighbors from various scales. Then the LBP values of the entire image are calculated, and multiple 3D feature maps are generated. Finally, we combine the texture features from different scales for image classification. The histogram is adopted to characterise the distribution of local patterns in the image, and the texture features are represented as a vector H_r. We calculate the histograms of various feature maps separately and concatenate them together to form the fused feature vector H. The equation is defined as follows:

$$H = H_1 \oplus H_2 \oplus ... \oplus H_r \tag{4.3}$$

where $\oplus$ represents the concatenation operation.

4.2.3 Classification

Among all classification algorithms, we adopt the random forest to group participants according to whether or not they have experienced a previous fracture (Valentinitsch et al. (2019)). The random forest classifier is an ensemble learning algorithm that consists of a number of decision trees (Paul et al. (2018)). Decision trees are constructed using random bootstrap samples from the training dataset. Each decision tree votes for the possible options, and the random forest classifier selects the option that has the highest number of votes. Compared to the single decision tree and other machine learning classifiers such as k-nearest neighbour and multilayer perceptron, random forest is less prone to overfitting because the integrated results from multiple weak classifiers result in fewer errors. Therefore, it is used to deal with small and unbalanced datasets, and demonstrates superior performance to a single classifier (Sagi and Rokach (2018)).

In our study, there are fewer participants with previous fractures compared to those without. Therefore, an under-sampling strategy is used for individuals without previous fractures to balance the data (Lin et al. (2017)). Then 80% of tibial HR-pQCT images from the balanced data are selected randomly as the training set, while the remaining data is equally divided into the validation and testing sets. The experiments are repeated ten times to evaluate performance. All analyses to assess the discriminative performance of the random forest classifier are based on the testing set.

4.2.4 Statistical Analysis

Participant characteristics between fracture and non-fracture groups are compared. The receiver operator characteristic (ROC) curve, sensitivity, specificity, accuracy and AUC are used to assess the discriminative capability of our approach for previous fracture. In addition, a 95% confidence interval (CI) for the AUC is calculated, representing a range of values where the true parameter value is expected to lie with a 95% probability. It uses bootstrap resampling to estimate the parameter and then constructs a distribution of results to determine the interval (Robin et al. (2011)). Sensitivity, specificity and accuracy are defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{4.5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.6}$$

where, TP, TN, FP and FN represent the true positive, true negative, false positive and false negative respectively.

A participant is considered to have a previous fracture if the predicted fracture probability exceeds the threshold. The optimal threshold is determined by the Youden's Index (Youden (1950)) defined as in Equation 4.7, which summarises the overall performance of a diagnostic test or classifier across all threshold values.

$$\text{Youden's index} \;=\; \text{Sensitivity} + \text{Specificity} - 1 \tag{4.7}$$

The Youden's index is calculated for each value of the threshold on the validation set. The threshold that corresponds to the highest Youden's index is selected as the optimal threshold.

The discriminative capacity of different fracture risk assessment techniques is compared. Of particular interest is whether the AUC for our approach is substantially greater than the AUC for DXA measurement and clinical risk assessment. Therefore, the statistical significance of the difference in these two AUCs is examined using DeLong's test (DeLong et al. (1988)). DeLong's test is chosen due to its effectiveness in handling the correlation between the AUCs of two ROC curves. This non-parametric method is widely applied in medical statistics for evaluating the performance differences between diagnostic tests. The p-value represents the probability of observing a difference in AUC values as extreme as the one observed in the data, assuming that the null hypothesis of no difference between the two methods is true. A significance level of 0.05 is used; a p-value <0.05 is considered strong evidence against the null hypothesis, indicating a statistically significant difference in the AUCs obtained by the two methods.

Sensitivity analyses involve stratifying the analyses by sex, including distal tibial HR-pQCT scans, clinical risk factors and DXA-measured BMD. Clinical risk factors include age, sex, height, weight, BMI, dietary calcium, smoking history, alcohol consumption, physical activity, bisphosphonate usage, number of comorbidities and occupational social class. The ascertainment of these clinical factors in HCS has been described previously (Fuggle et al. (2021)). Bone microarchitecture variables have been previously demonstrated to relate to fracture risk independently of DXA-measured BMD (Samelson et al. (2019)).

The analysis sample comprises 167 participants with data on previous fractures. Python 3.7 is used to extract image features from participants and train the random forest classifier. All statistical analysis for predicted results is implemented in R, version 4.0. All analyses are performed on Intel (R) Core (TM) i5-6600 CPU 3.30GHz with HD Graphics 530.

4.3 Results

4.3.1 Parameter Settings

When encoding the local patterns in tibial HR-pQCT images, we construct three different cubes with $3 \times 3 \times 3$, $4 \times 4 \times 4$ and $5 \times 5 \times 5$, and set the corresponding Euclidean distance r to 1, 2 and 3 separately to sample neighboring points. Considering the amount of calculation and feature dimensions in 3D space, we assign the number of sampling points N in a cube to 6. The histogram bins of each texture operator are 64.

4.3.2 Performance of Fracture Risk Assessment

Three different measurements are used to assess fracture risk and make a fair comparison. The standard DXA measurement uses femoral neck BMD values as input data. In the clinical risk assessment approach, 12 clinical covariates are normalized and used as input features for the random forest classifier to discriminate previous fractures. Our method extracts texture features from tibial HR-pQCT image data and then feeds them into the random forest classifier for fracture discrimination.

The sensitivity, specificity and AUC (95% CI) results from various measurements are presented in Table 4.1, according to the participant input information used; the corresponding ROC curves are shown in Figure 4.4. Specificity, sensitivity and accuracy are calculated using the predicted probability of fracture at the optimal threshold as described in Section 4.2.4, summarised in Table 4.1 and Figure 4.5.

All measurements capture valuable information regarding fracture risk. Compared to DXA-measured BMD (AUC: 0.63, 95% CI: 0.54-0.69) and clinical data (AUC: 0.60, 95% CI: 0.52-0.67), HR-pQCT image data demonstrate a higher classification accuracy (AUC: 0.73, 95% CI: 0.65-0.79). Furthermore, there is a statistically significant difference between the AUCs obtained from tibial HR-pQCT image data and both DXA-measured BMD (p-value$<$0.05) and clinical data (p-value$<$0.03). When the FPR is allowed to be 20%, the standard DXA detects only 25% of individuals with previous fractures. However, HR-pQCT measurement significantly improves the TPR to 50%. In addition to the random forest classifier and the optimal threshold, classification results for various machine learning algorithms and thresholds are given in Table B.1 and Table B.2 of Appendix B.

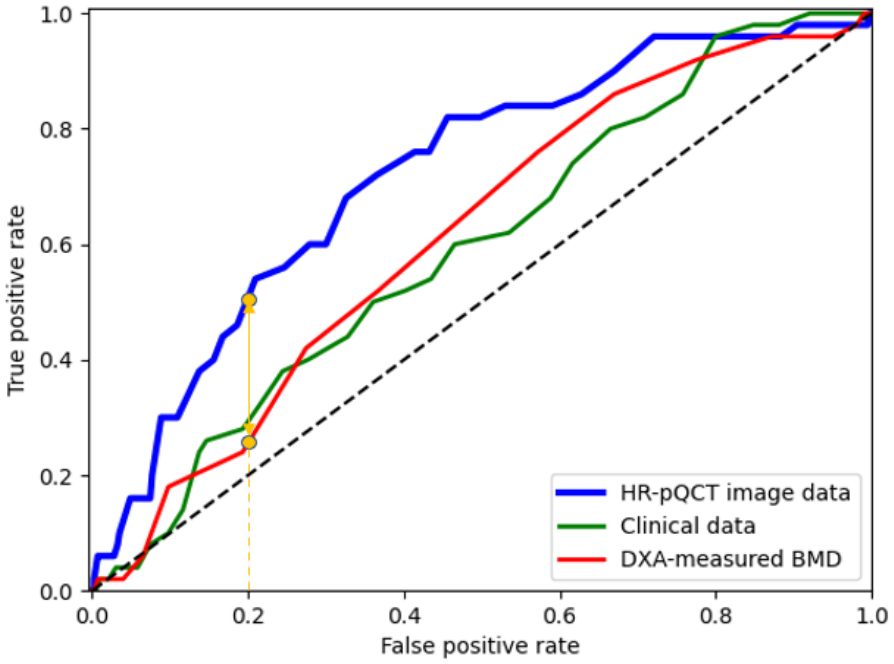

FIGURE 4.4: Receiver operating characteristic (ROC) curves for previous fracture from DXA-measured BMD, clinical data and HR-pQCT image data. HR-pQCT image data from tibial scans are used. Notably, at a false positive rate of 20%, the true positive rate of the gold standard DXA is poor. HR-pQCT shows substantial improvement compared to DXA (depicted by the yellow line).

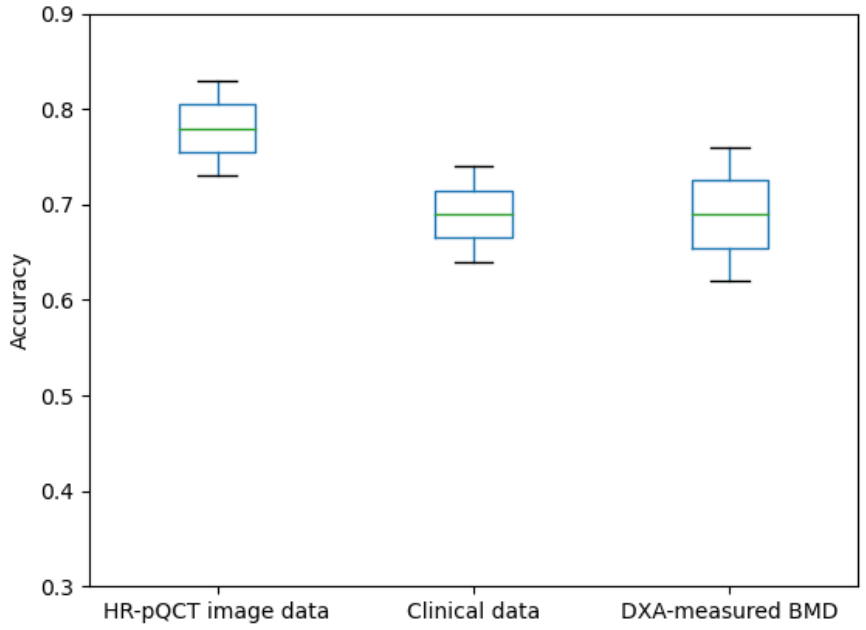

FIGURE 4.5: Classification accuracy of DXA-measured BMD, clinical data and HR-pQCT image data for previous fracture. HR-pQCT image data from tibial scans are used.

TABLE 4.1: Discriminative performance of DXA-measured BMD, clinical data and HR-pQCT image data for previous fracture.

Input data	AUC (95% CI)	Sensitivity	Specificity
DXA-measured BMD	0.63 (0.54-0.69)	0.42	0.73
Clinical data	0.60 (0.52-0.67)	0.40	0.66
HR-pQCT image data	**0.73 (0.65-0.79)**	**0.46**	**0.81**

BMD: femoral neck BMD.
Clinical data includes 12 clinical covariates such as age and gender.
HR-pQCT image data from tibial scans are used.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

TABLE 4.2: Discriminative performance of DXA-measured BMD, clinical data and HR-pQCT image data for previous fracture (sex-specific analyses).

Input data	Males		
	AUC (95% CI)	Sensitivity	Specificity
DXA-measured BMD	0.64 (0.53-0.71)	0.13	0.78
Clinical data	0.58 (0.50-0.67)	0.23	0.79
HR-pQCT image data	**0.67 (0.57-0.79)**	**0.37**	**0.86**
Input data	Females		
	AUC (95% CI)	Sensitivity	Specificity
DXA-measured BMD	0.67 (0.55-0.75)	0.30	**0.83**
Clinical data	0.61 (0.50-0.71)	0.30	0.80
HR-pQCT image data	**0.70 (0.58-0.79)**	**0.43**	0.78

BMD: femoral neck BMD.
Clinical data includes 12 clinical covariates such as age and gender.
HR-pQCT image data from tibial scans are used.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

4.3.3 Sensitivity Analyses

Similar results are observed in the stratification of analyses by sex (Table 4.2). In this scenario, the use of HR-pQCT image data results in a statistically significant improvement (p-value$<$0.05) in fracture classification as measured using AUCs compared to the use of clinical data and BMD.

4.4 Discussion

In this study, we propose an automatic discriminative algorithm that uses tibial HR-pQCT image data to discriminate between people with and without previous fractures. The AUC obtained from HR-pQCT image data (0.73, 95% CI: 0.65-0.79) is higher than clinical data (0.60, 95% CI: 0.52-0.67) and DXA-measured BMD (0.63, 95% CI: 0.54-0.69);

volumetric texture analysis of image data significantly improves fracture discrimination compared to the use of clinical data and DXA-measured BMD (p-value$<$0.05). This suggests that valuable information regarding fracture risk is utilised by image processing methods which is not captured by DXA measurement and clinical risk factors. In addition, our approach shows higher accuracy than comparative methods for discriminating previous fractures in either men or women separately (see Table 4.2).

Imaging plays an important role in osteoporosis, with DXA BMD incorporated in the definition of the condition (Organization et al. (1994)) and with the advent of computer vision techniques, a body of image-related machine learning research has started to develop. This has largely centred on using deep learning to assess for osteoporosis on routine CT scans via automated vertebral body segmentation and then training an algorithm to predict aBMD or a measure of vBMD (Valentinitsch et al. (2019), Löffler et al. (2021), Valentinitsch et al. (2019)). A similar approach has also been used with hip radiographs to assess BMD (Yamamoto et al. (2020)). The only previous work using HR-pQCT utilised radial trabecular texture in 18 post-menopausal women with fragility fractures and 18 post-menopausal women without fragility fractures (a small number of participants compared to our study cohort) (Valentinitsch et al. (2013)), but did not investigate the discriminative performance of texture features for fracture classification. Our study also leverages recent computer vision developments in texture analysis and further exploits the random forest classifier to discriminate previous fractures.

We propose a 3D LBP method to capture bone microarchitectural information from HR-pQCT images, making full use of the three-dimensionality of the HR-pQCT images and the bone tissue they depict. 2D textural analysis is used in clinical practice in the form of trabecular bone scores (TBS) on lumbar spine DXA images (Silva et al. (2014)). However, these lack spatial information regarding bone geometry and microarchitecture. (Murala and Wu (2015), Abbasi and Tajeripour (2017)). To address this issue, we develop a method that constructs a 3D spatial cube for each voxel in the 3D images to calculate the feature descriptor. Statistical distributions of texture patterns in HR-pQCT images are used to quantify bone microarchitecture for fracture discrimination. The results presented in Table 4.1 and Table 4.2 confirm our hypothesis that fracture risk is determined in part by texture representations of HR-pQCT images.

This study has some limitations. Firstly, although random forest classifiers have advantages for analysing small datasets (Zhang and Ma (2012)) and similar sample sizes have been used in previous musculoskeletal research publications which have implemented this technique (Hussain and Han (2019), Mehta and Sebro (2019), Gornale et al. (2016)), a major limitation of our study is that the sample size is small. Secondly, our method relies on data analysis of HR-pQCT images rather than existing knowledge for

fracture classification. As a result, the generalisability and reproducibility of these findings may be limited. Thirdly, previous (rather than incident) fracture history is used as the outcome. Fourthly, further information on self-reported fractures, such as their location and type, is not available; at the 2011-2012 follow-up, participants were only asked whether they had broken any bones since age 45 years.

In addition, the HCS dataset is unbalanced which affects the discriminative performance of the random forest classifier (Bader-El-Den et al. (2018)). Therefore, undersampling techniques are applied to the group of participants without previous fractures to balance the data. In this study, 3D texture features are extracted from the entire CT scans. Samples with global texture information of CT scans lead to more statistically meaningful and stable outcomes. Oversampling techniques have also been attempted to balance the data (Lu et al. (2022a)). Specifically, tibial CT scans are assumed to be homogeneous. A sample of 22 consecutive slices is selected for each HR-pQCT image, and different numbers of samples are selected from the scans with and without previous fractures. However, in this scenario, samples with local texture information of images are used to train the classifier, and the outcomes are not stable (the best AUC result is 0.86 when tibial HR-pQCT image data is used as input data to the discriminative system).

4.5 Summary

Volumetric texture analysis combined with machine learning provides an exciting opportunity to utilise HR-pQCT imaging to identify individuals at high risk of fracture. This will potentially allow timely treatment and improved clinical outcomes; however, prior to deployment, this work needs to be applied to datasets associated with other cohorts.

Chapter 5

Decoupling Relative Contributions of Cortical and Trabecular Bone in Fracture Discrimination

5.1 Inspirations and Introduction

HR-pQCT measurements are volumetric, providing an assessment of physiological cross section of the bones and providing separate assessments of bone compartments and vBMD (Hansen et al. (2014), Engelke et al. (2013)). The accurate semantic segmentation and quantitative analysis of cortical and trabecular bone compartments in HR-pQCT images have received increasing attention in recent years (Whittier et al. (2020)). Initially, HR-pQCT slices were manually segmented by human operators (Laib et al. (1998)). Subsequently, a dual-thresholding algorithm emerged as the current gold standard for the automatic segmentation of cortical and trabecular compartments (Buie et al. (2007), Burghardt et al. (2010)). However, this approach proved to be less robust for specific populations such as osteoporotic patients, and still required manual inspection and corrections. Neeteson et al. (2023) recently developed an automated method based on U-Net for the segmentation of cortical and trabecular compartments in HR-pQCT images of the radius and tibia. However, they did not quantitatively analyse the bone microarchitecture of these two compartments in CT scans.

In Chapter 4, we reveal that texture information obtained from the entire HR-pQCT images is associated with fracture risk. However, the relative contributions of cortical and trabecular compartments are unknown. Based on our work in Chapter 4 and that of others (Fuggle et al. (2022), Valentinitsch et al. (2019)), we hypothesise that cortical and trabecular regions at the distal tibia possess important information separately regarding fracture risk. Therefore, we propose to extract texture features from the cortical

and trabecular regions in HR-pQCT images separately.

In this chapter, we develop an automatic approach that segments cortical and trabecular regions in HR-pQCT scans for texture feature extraction to discriminate previous fractures. First, we construct a dataset with annotated cortical and trabecular regions in HR-pQCT slices, and employ a deep CNN for automatic and accurate segmentation of these two regions. Second, we use the LBP model to characterise the texture patterns of cortical and trabecular regions separately to quantify bone microarchitecture. Further, we investigate the relative contributions of cortical and trabecular compartments in fracture discrimination. Lastly, we evaluate the noise tolerance of our discriminative system and propose an efficient strategy to enhance its robustness against noise in HR-pQCT images.

5.2 Method

5.2.1 Design and Overview

The present study here is designed to test the hypothesis that both cortical and trabecular regions in tibial HR-pQCT images possess important information about fracture risk. Therefore if features are extracted from cortical or trabecular regions, previous fractures can be discriminated.

The framework of our method is illustrated in Figure 5.1. First, we employ a 2D U-Net model to segment cortical and trabecular regions in tibial HR-pQCT slices. Subsequently, we utilise the 3D LBP model to characterise cortical and trabecular bone microarchitecture in CT images. Texture features extracted from cortical and trabecular regions are then separately used to train the random forest classifier for fracture discrimination.

The segmentation and HCS datasets, detailed in Chapter 3, are utilised to evaluate the performance of our method in image segmentation and fracture discrimination. The cortical and trabecular regions in tibial HR-pQCT scans of 167 subjects from the HCS dataset are automatically segmented by the U-Net model, which was trained on the segmentation dataset.

5.2.2 Image Segmentation

The automatic segmentation of cortical and trabecular regions in HR-pQCT scans is our method's first and crucial step because the subsequent task is based on the located regions. To achieve accurate and efficient segmentation of regions in tibial HR-pQCT

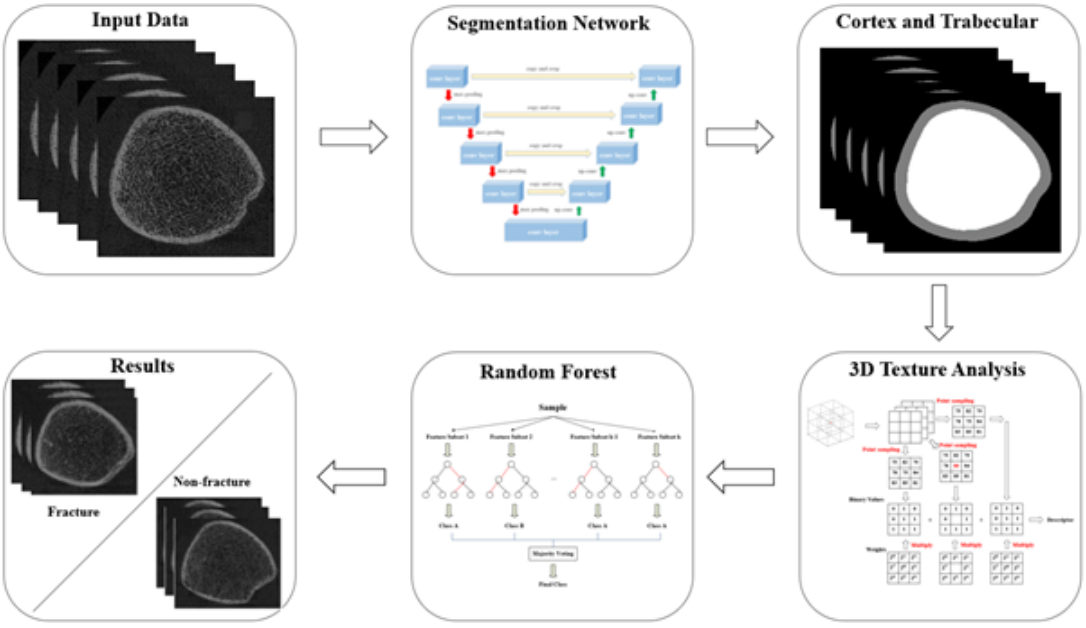

FIGURE 5.1: Automatic segmentation of cortical and trabecular regions for fracture classification. 3D texture analysis is separately conducted on segmented cortical and trabecular regions in tibial HR-pQCT scans.

images, referring to the success of deep CNNs in many tasks, we employ a U-Net model to perform this procedure (Fang et al. (2021)). Before training the model, we first resize the tibial CT transverse slices into 224×224 pixel resolutions.

The U-Net model contains an encoder and a decoder (Ronneberger et al. (2015)), as shown in Figure 5.2. The encoder extracts the features of input CT slices by repeatedly using 3×3 convolutional layers followed by rectified linear unit (ReLU) activation functions and 2×2 max pooling operations with stride 2. The decoder maps the high-level semantic features of images to the segmentation results through 2×2 up-convolution (up-conv) operations. In addition, there is a concatenation operation through skip connections between the feature map from the encoder to the decoder to reduce the information loss of border pixels in the convolution operations. In the final layer, each feature vector is mapped to three groups using 1×1 convolutions. There are 23 convolutional layers in the network.

The segmentation dataset comprises 3300 CT transverse slices with pixel-level annotations. Among these, 2310 tibial CT slices are used as the training set, 330 CT slices are used as the validation set, and the remaining CT slices are used as the testing set. The U-Net model, trained on this segmentation dataset, is subsequently employed to determine the position of the cortical and trabecular regions in other CT scans for further texture analysis.

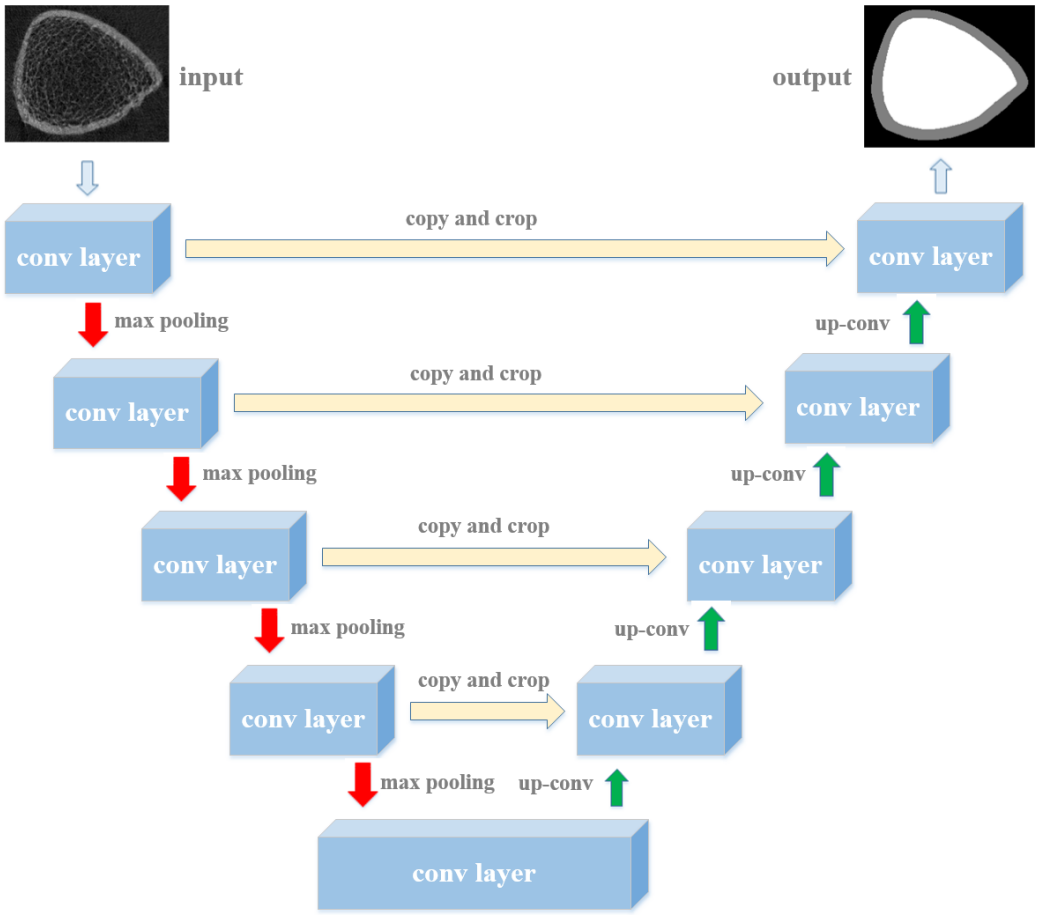

FIGURE 5.2: Architecture of the U-Net for image segmentation of tibial HR-pQCT slices. The skip connection involves cropping feature maps from the encoder, copying them and connecting them to the feature maps from the decoder.

5.2.3 3D Texture Analysis

Considering that microarchitectural deterioration of bone tissue is associated with fracture risk, we adopt the 3D LBP method to characterise the texture patterns of cortical and trabecular regions separately. Histograms are then constructed to quantify the statistical distributions for further classification.

Let $\Omega_v = \{v_{i,j,k} | 1 \leq i \leq W,\ 1 \leq j \leq H,\ 1 \leq k \leq Z\}$ represent a set of voxels in a solid image with $W \times H \times Z$. The neighborhood scheme here is that samples N voxels over a cube with $E \times E \times E$ (see Figure 4.2). We let r denote the Euclidean distance between the centre location v_c and sampled points v_i on the cube surface. The detailed procedure for 3D texture feature extraction is introduced in Section 4.2.2 of Chapter 4. We also adopt the multi-scale strategy that calculates a set of LBP descriptors from various scales and concatenates the corresponding texture feature vectors to quantify bone microarchitecture.

5.2.4 Classification

We use the random forest classifier to discriminate between subjects with and without previous fractures. After image segmentation of cortical and trabecular regions, there are 167 tibial CT scans with segmentation masks for fracture classification. However, the class distribution is unbalanced in our study. There are 121 images with non-fractures and 46 images with fractures. No participant has more than one tibial scan. The unbalanced data usually leads to a biasing problem that the classifier performs badly in the class with fewer samples. In order to overcome this issue, we use an under-sampling strategy for non-fracture images to balance the data. Then 80% of tibial HR-pQCT images from the balanced data are selected randomly as the training set, while the remaining data is equally divided into the validation and testing sets. We repeat the experiments 10 times and calculate the average of the classification results.

5.2.5 Statistical Analysis

Participant characteristics are described using summary statistics in Section 3.3.1 of Chapter 3. The Intersection over Union (IoU) is utilised to measure the segmentation performance of our U-Net model. It is the ratio of the overlapping area of ground truth and predicted area to the union area. In terms of fracture risk assessment, we use evaluation metrics such as TPR (sensitivity), FPR, TNR (specificity), AUC and accuracy to assess the discriminative performance (Singh et al. (2017)). These metrics are described in Chapter 4. In addition, a confidence interval (CI) which is a range with an upper and lower value is calculated from samples to take account of the uncertainty. A significant level of 0.05 is also used in this study, and we regard a statistically significant difference when the p-value is less than 0.05.

Model robustness analysis involves evaluating the noise tolerance of our approach on tibial CT scans. Zero mean Gaussian noise with various standard deviations is added to HR-pQCT images. We use the signal-to-noise ratio (SNR) to characterise the ratio of signal power to noise power (Ling and Bovik (2002)). It is defined as:

$$SNR = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}s(i,j,k)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}[s(i,j,k)-n(i,j,k)]^2} \tag{5.1}$$

where, $s(i,j,k)$ and $n(i,j,k)$ separately represent the grayscale value of noise-contaminated and noise-free images at the point (i,j,k).

We compare the discriminative performance of HR-pQCT-based measures (including the entire tibia, cortical region and trabecular region) with clinical risk factors and DXA-measured BMD for previous fracture. DXA measurement utilises femoral neck BMD

values for fracture classification. Clinical risk factors include age, sex, height, weight, BMI, dietary calcium, smoking history, alcohol consumption, physical activity, bisphosphonate usage, number of comorbidities and occupational social class. These two measurements are described in Chapter 4.

5.3 Results

5.3.1 Parameter Settings

U-Net is used as the segmentation network, and hyperparameters are set as follows: the Adam optimizer is utilised in the training process, with a learning rate of 0.0002 and 50 epochs, and the batch size is set to 4. The U-Net architecture is illustrated in Section 5.2.3, and the details are described in Table 5.1. 3D LBP is employed to extract texture features from tibial HR-pQCT images, and the details are described in Section 4.3.1 of Chapter 4.

5.3.2 Performance of Image Segmentation

Figure 5.3 illustrates the examples of original tibial HR-pQCT images, manual annotations and automated results of the U-Net model. IoU quantifies how well the manual marking matches automatic segmentation by dividing the intersection of two segments by their union. According to the results, a predicted segmentation and the corresponding ground-truth annotation correlate very well that the mean IoU of CT scans on the testing set is 0.96. Furthermore, manual and automatic segmentation of various regions in tibial HR-pQCT slices are used separately to measure the IoU. The segmentation results for cortical and trabecular regions, as well as surrounding soft tissue, are presented in Table 5.2. Therefore, our U-Net model can achieve state-of-the-art performance on automatically segmenting cortical and trabecular regions at the distal tibia.

5.3.3 Performance of Fracture Risk Assessment

We separate cortical and trabecular regions in tibial HR-pQCT scans and input different components into the LBP model to compare the discriminative performance of fracture risk assessment. Furthermore, we compare our method with traditional BMD measurement and participants' clinical covariates (Edwards et al. (2016)). The discriminative performance of HR-pQCT-based measures, clinical risk factors and DXA-measured BMD is summarised in Table 5.3. Specificity, sensitivity and accuracy are calculated using the predicted probability of fracture at the optimal threshold. The optimal threshold is determined by the Youden's Index (Youden (1950)). The classification accuracy and

TABLE 5.1: Detailed description of the U-Net architecture.

Layer name	Encoder	
	Output size	Operations
Conv1	224×224	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix}$
Conv2	112×112	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix}$
Conv3	56×56	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix}$
Conv4	28×28	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix}$
Conv5	14×14	$\begin{bmatrix} 3 \times 3, & 1024 \\ 3 \times 3, & 1024 \end{bmatrix}$
Layer name	Decoder	
	Output size	Operations
Up4	28×28	Upsample
Up-conv4	28×28	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix}$
Up3	56×56	Upsample
Up-conv3	56×56	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix}$
Up2	112×112	Upsample
Up-conv2	112×112	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix}$
Up1	224×224	Upsample
Up-conv1	224×224	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix}$
Output layer	224×224	$[1 \times 1, 3]$

TABLE 5.2: Segmentation performance of the U-Net model.

Regions	IoU (standard deviations)
Cortical region	0.84 ± 0.04
Trabecular region	0.96 ± 0.02
Surrounding soft tissue	0.97 ± 0.01

ROC curves of HR-pQCT-based measures, including the entire tibia, cortical region and trabecular region, are presented in Figure 5.4 and Figure 5.5.

As the classification results illustrated in Table 5.3, our volumetric texture analysis method applied to tibial HR-pQCT image data shows a stronger signal in identifying individuals with previous fractures than DXA measurement and clinical risk assessment. The standard DXA and clinical risk assessment approaches, used to discriminate between people with and without previous fractures, yield AUCs of 0.63 (95% CI: 0.54-0.69) and 0.60 (95% CI: 0.52-0.67) respectively. When the texture features of the entire

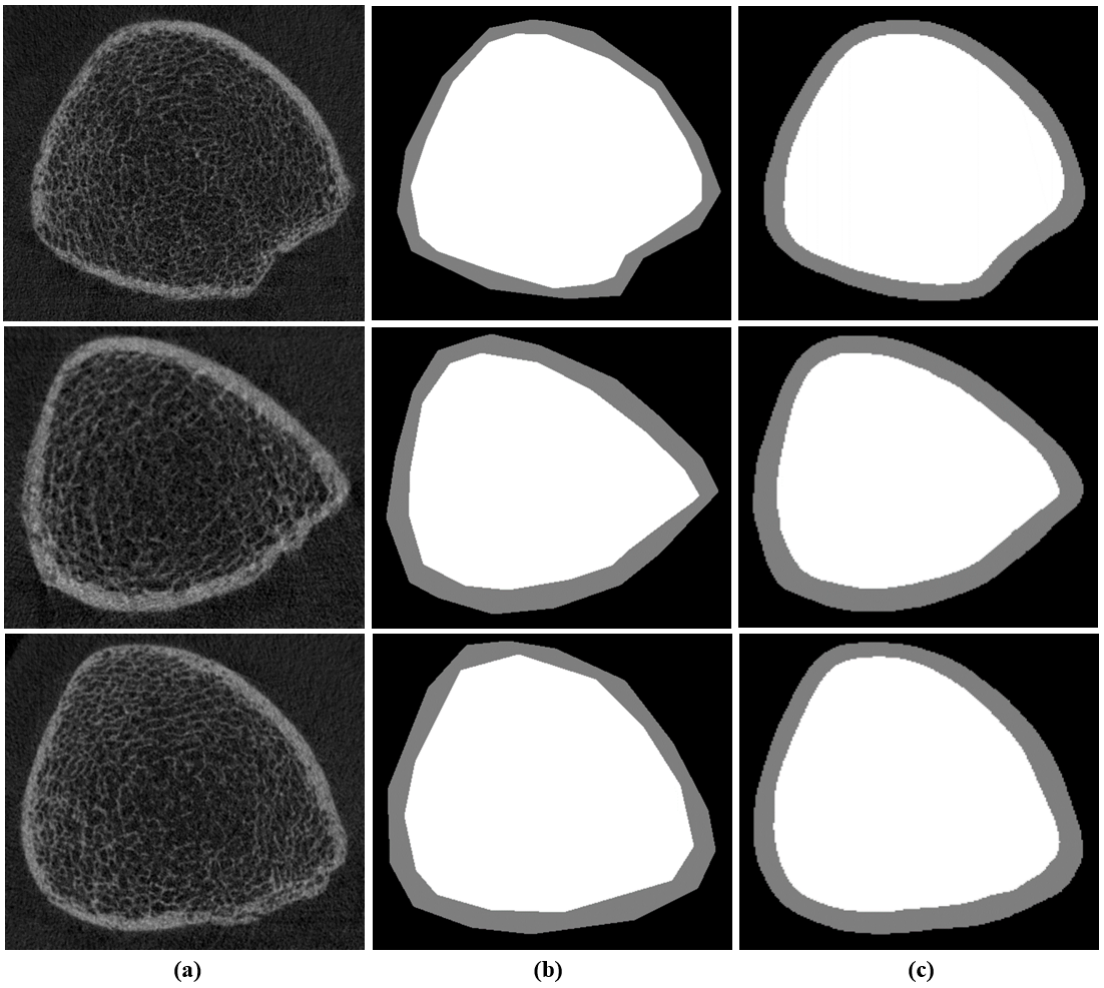

FIGURE 5.3: Examples of image segmentation: input CT slices from tibial scans (a), and manual (b) and automatic (c) segmentations.

tibia in HR-pQCT images (without segmentation) are extracted for fracture classification, the AUC improves to 0.73 (95% CI: 0.65-0.79). The sensitivity of this method is 0.46, and the specificity is 0.81. After image segmentation, the cortical and trabecular regions in HR-pQCT images are automatically localised, and texture features are extracted separately for fracture discrimination. The AUCs of the segmented compartments are 0.75 (95% CI: 0.67-0.81) for the cortical compartment and 0.66 (95% CI: 0.56-0.71) for the trabecular compartment. Therefore, both cortical and trabecular compartments contain valuable information regarding fracture risk. The cortical compartment significantly outperforms the trabecular compartment in terms of fracture discrimination (p-value <0.05).

The statistical analysis shows that there is a significant difference between the AUCs obtained from HR-pQCT-based measures (including the entire tibia and cortical region) and both clinical risk factors and DXA-measured BMD (p-value <0.05). However, no statistically significant difference is found between the AUCs obtained from the trabecular region and both clinical risk factors and DXA-measured BMD (p-value >0.05).

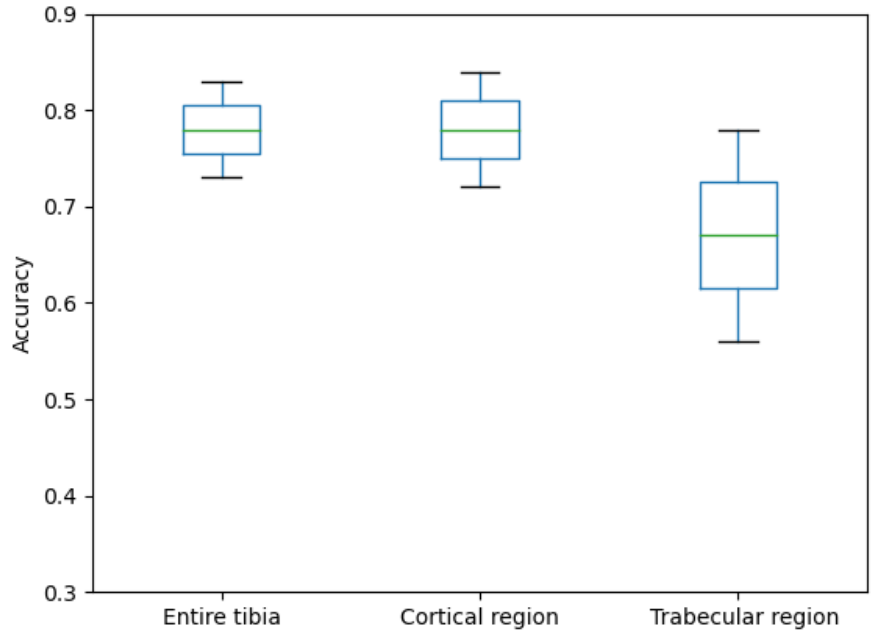

FIGURE 5.4: Classification accuracy of tibial HR-pQCT scans from automatic segmentation for previous fracture. 3D texture analysis is separately conducted on the entire tibia, cortical regions and trabecular regions in CT scans to discriminate previous fractures.

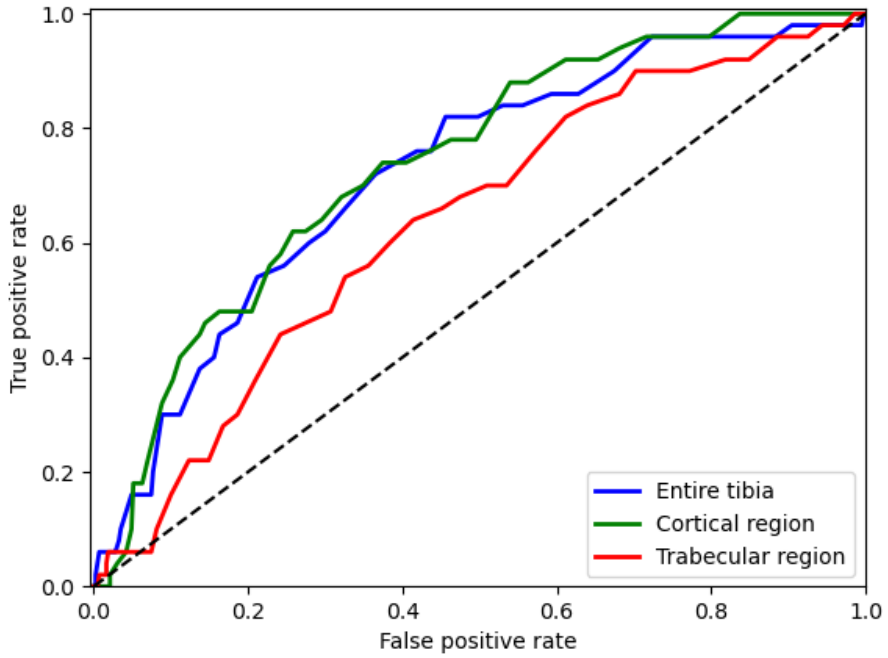

FIGURE 5.5: Receiver operating characteristic (ROC) curves for previous fracture from HR-pQCT-based measures. The entire tibia, cortical regions and trabecular regions in tibial HR-pQCT scans are used. Notably, at low false positives, the cortical region shows substantial improvement (higher positive predictive performance) compared to the trabecular region.

TABLE 5.3: Discriminative performance of HR-pQCT-based measures, DXA-measured BMD and clinical data for previous fracture.

Input data	AUC (95% CI)	Sensitivity	Specificity
Entire tibia	0.73 (0.65-0.79)	0.46	**0.81**
Cortical region	**0.75 (0.67-0.81)**	**0.56**	0.77
Trabecular region	0.66 (0.56-0.71)	0.48	0.69
DXA-measured BMD	0.63 (0.54-0.69)	0.42	0.73
Clinical data	0.60 (0.52-0.67)	0.40	0.66

The entire tibia, cortical regions and trabecular regions are from tibial HR-pQCT scans.
BMD: femoral neck BMD.
Clinical data includes 12 clinical covariates such as age and gender.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

5.3.4 Result Analysis

Here, we analyse the predictions when cortical regions in tibial HR-pQCT scans are segmented and used as input data for our algorithm. By inspecting the classification results in details, we find that most samples in the testing set are correctly identified, while a few samples are assigned incorrectly. Many samples have no apparent appearance of bone mass, and visual inspection to identify fractures is not feasible from images of the tibia as these may have occurred elsewhere in the skeleton.

Figure 5.6 and Figure 5.7 displays some examples of tibial HR-pQCT image classification results. Figure 5.6 (a) and (b) show that there is no apparent visual difference between two bone sections, especially in cortical bone. Our approach can analyse beyond the limitations of the standard image analysis method using the texture features extracted from cortical regions to achieve the correct identification of fracture and non-fracture cases. Figure 5.7 (a) and (b) show HR-pQCT images of fracture and non-fracture cases with incorrect predictions. These incorrect predictions may be an artefact of the fact that there is a substantial stochastic element to fracture occurrence which is independent of bone fragility and therefore will not be captured by the HR-pQCT image. By similar token, bone fragility (and therefore increased fracture risk) can be observed in the HR-pQCT image but may not have led to a fracture as the individual may not have fallen or been subject to any trauma.

5.3.5 Model Robustness Analysis

HR-pQCT images are, as with most medical images, subject to issues such as noise and movement artefact which may impact quality and cause issues with image processing methods. These uncertainties are likely to degrade the image quality and result in a lower discriminative performance for assessments, in this fracture classification. Here,

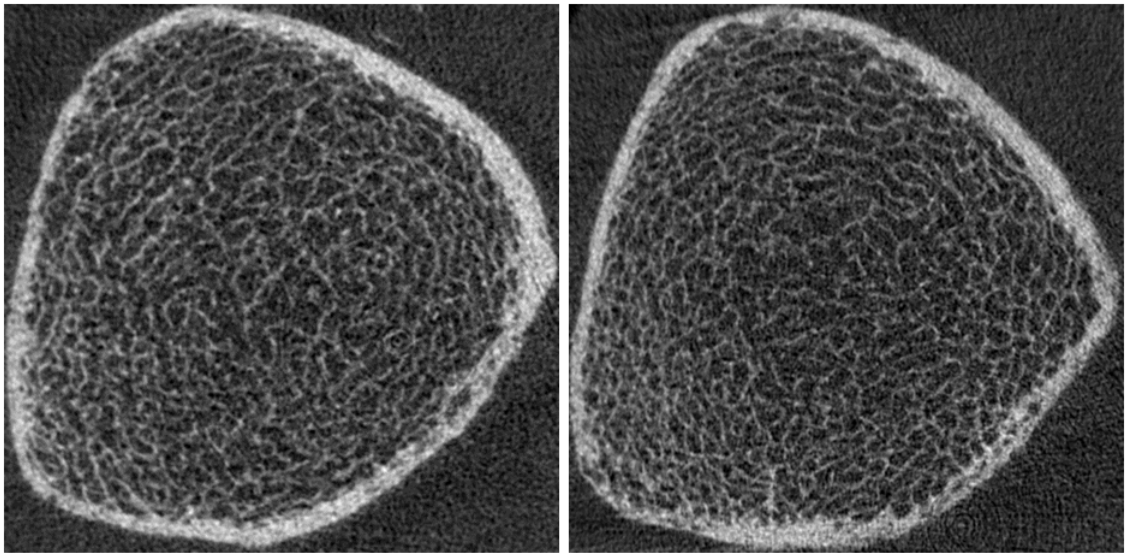

FIGURE 5.6: Examples of bone sections with correct classification results: fracture (a) and non-fracture (b) cases. Visual inspection does not discriminate between the two, while our discriminative system can achieve accurate predictions.

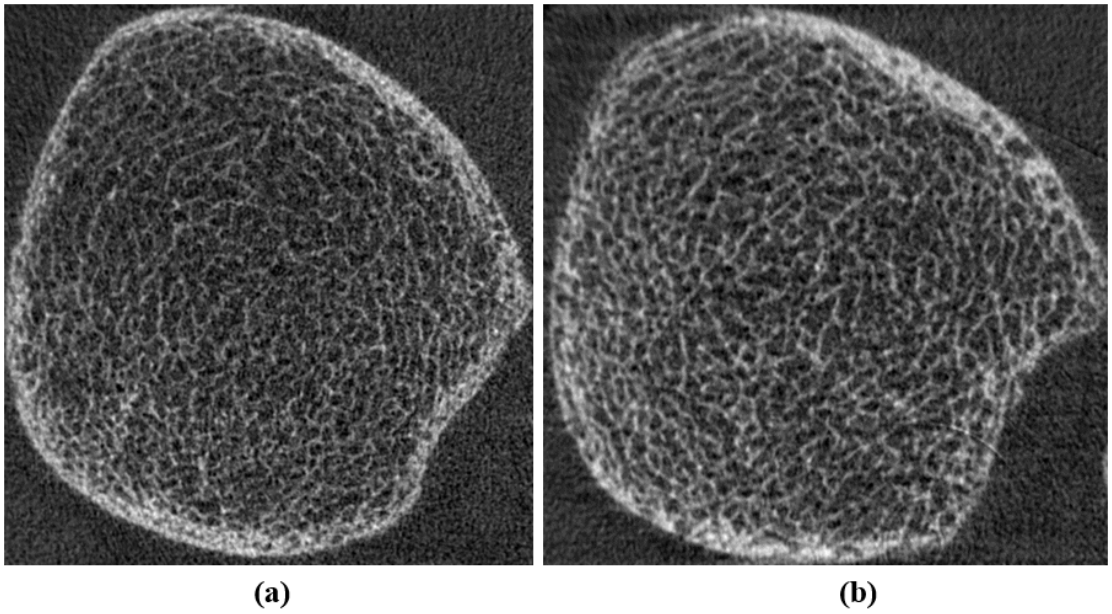

FIGURE 5.7: Examples of bone sections with incorrect classification results: fracture (a) and non-fracture (b) cases. The fracture in (a) may be caused by an accident, and the bone fragility in (b) may not have led to a fracture.

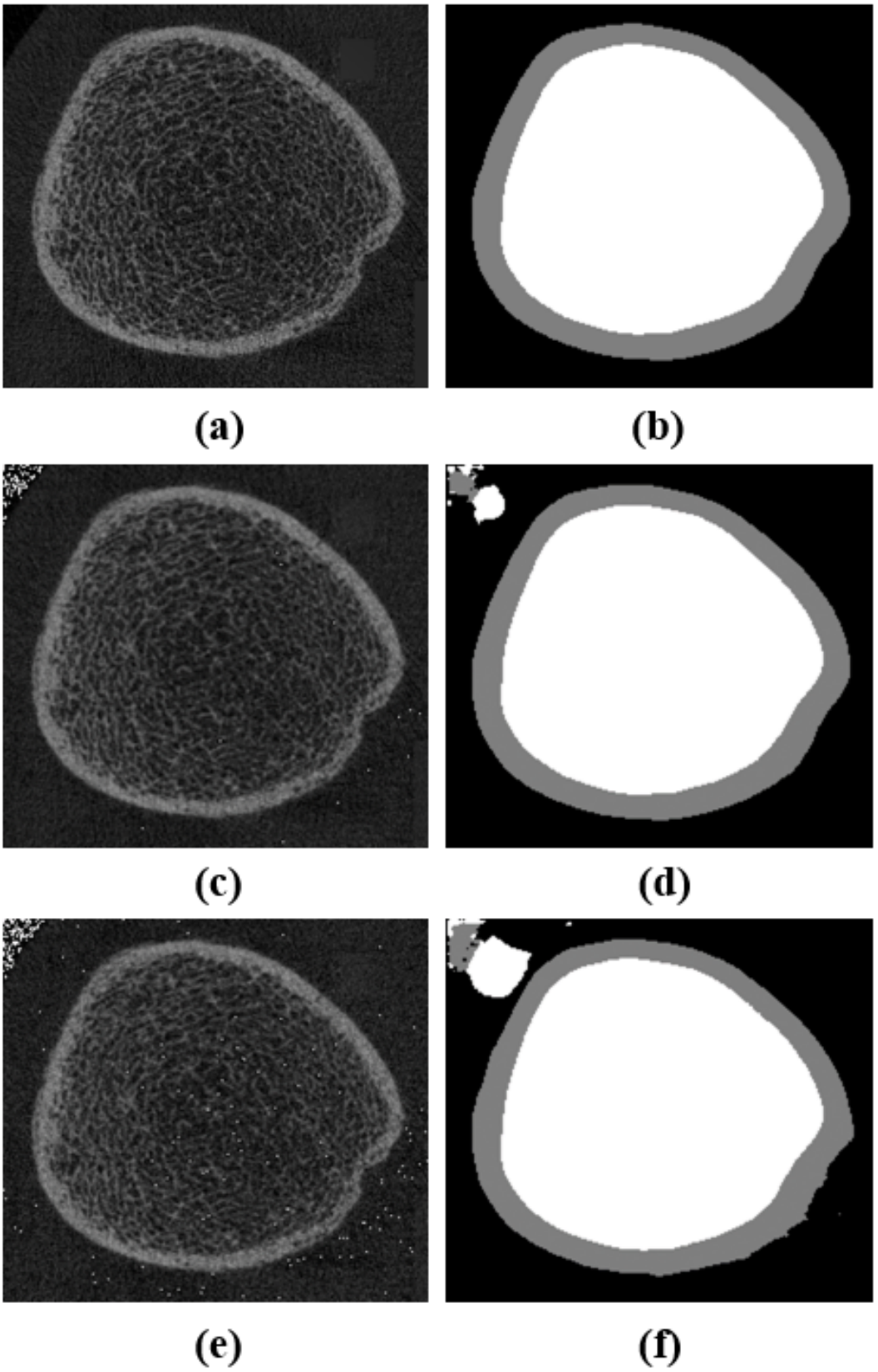

FIGURE 5.8: Performance of image segmentation against noise contamination. With noise-free images (a), U-Net produces good segmentation (b), while with additive Gaussian noise at SNR=20 (c) and SNR=15 (e), it shows gradual degradation in performance (d) and (f).

we present the performance of our discriminative system against image perturbations.

The training set consists of noise-free tibial HR-pQCT images, and zero mean Gaussian noise with various standard deviations is added to the images in the testing set. Gaussian noise is introduced to simulate noise scenarios in clinical practice and evaluate the noise-tolerant property of our approach. Noise contaminates HR-pQCT images and affects subsequent image analysis, including image segmentation and texture analysis.

First, the performance of cortical and trabecular region segmentation is influenced. Figure 5.8 (b), (d) and (f) illustrates the automatic segmentation results of normal and noisy CT slices. In noise-free HR-pQCT images, the U-Net model shows excellent performance (IoU, 0.96) in localising cortical and trabecular regions. However, when the image is contaminated with noise, segmentation performance gradually decreases, primarily in the voxel classification of boundary regions. The IoUs of image segmentation are 0.95 and 0.93 for noisy HR-pQCT images with SNR of 20 and 15 respectively. The results indicate that the U-Net model retains good performance in the automatic segmentation of cortical and trabecular regions even at the presence of noise in HR-pQCT images. The imperfect segmentation of boundaries between cortical and trabecular regions could also impact model performance. However, this aspect is not within the scope of robustness analysis in this study. Our robustness analysis focuses on evaluating the performance variations of our approach when Gaussian noise is introduced to HR-pQCT images.

Noise interference also changes the texture patterns of images (Lu et al. (2022b)) and has a negative influence on fracture classification. Although the LBP texture descriptor exhibits high discriminative power, it is susceptible to noise in the image, as described in Section 2.4.3 of Chapter 2. Figure 5.9 (a) depicts the classification performance of our approach with various Gaussian noise levels in HR-pQCT images. Cortical regions automatically segmented from CT images are used as input data for our approach. In noise-free HR-pQCT images, our proposed 3D LBP model achieves an AUC of 0.75 for fracture discrimination. However, when the images are contaminated with noise, the discriminative performance of our approach sharply decreases. To address this issue, we propose an efficient and robust texture descriptor to enhance the noise tolerance of our discriminative system. The details of our method are described in Appendix C. Figure 5.9 (b) demonstrates that the proposed robust completed local binary pattern (RCLBP) method can enhance the model robustness of our discriminative system against noise in HR-pQCT images.

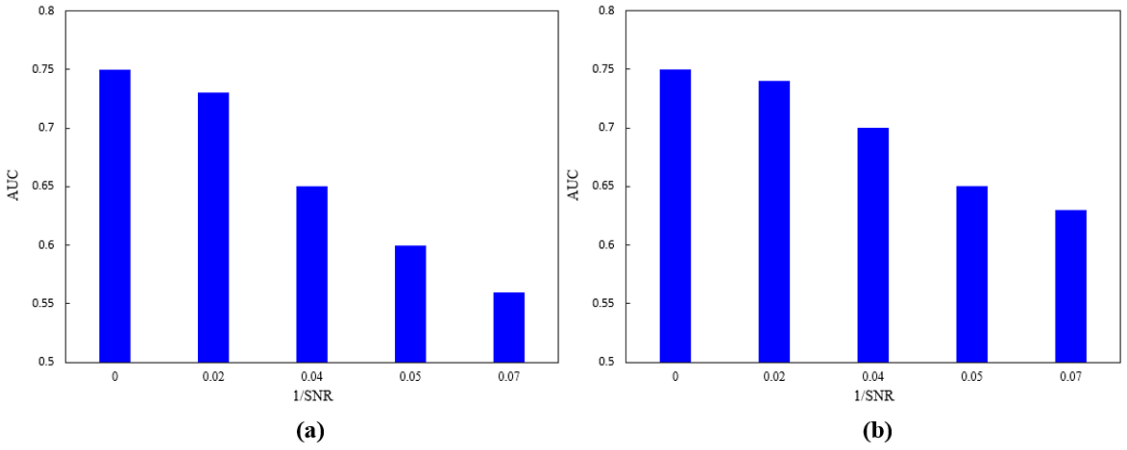

FIGURE 5.9: Classification performance with increasing levels of noise contamination. The proposed RCLBP method (b) in Appendix C shows a lower degradation in discriminative performance in comparison to the 3D LBP (a).

5.4 Discussion

This study presents an approach to automatically segment cortical and trabecular regions in tibial HR-pQCT images and separately quantify their bone microarchitecture for fracture discrimination. Specifically, we construct a segmentation dataset, and employ a deep neural network U-Net to extract semantic features of CT slices to tackle the image segmentation task. Further, a 3D LBP model is used to capture bone microarchitectural information from HR-pQCT scans to discriminate between subjects with and without previous fractures. Our study suggests that the cortical region (AUC: 0.75, 95% CI: 0.67-0.81) outperforms the trabecular one (AUC: 0.66, 95% CI: 0.56-0.71) in fracture discrimination. The results presented in Table 5.3 confirm our hypothesis that cortical and trabecular regions possess important information separately about fracture risk.

To the best of our knowledge, our work is the first to automatically segment cortical and trabecular regions at the distal tibia for fracture discrimination. In recent years, a combination of CT screening and computer-aided diagnosis has attracted extensive attention in fracture risk assessment, and some approaches have been proposed based on bone CT images such as spinal CT scans (Löffler et al. (2021)). Our research leverages the tibial HR-pQCT images from the HCS for fracture classification, and cortical and trabecular region localisation is the basis for our measurements. We employ a U-Net model for image segmentation that captures the context information of bone in CT slices and successfully trains the neural network with limited annotated training samples. Automatic segmentation of cortical and trabecular regions is highly correlated with manual annotation, and the average IoU is 0.96. Furthermore, we explore the individual contributions of cortical and trabecular compartments in fracture discrimination.

The 3D texture analysis method plays an important role in capturing bone fragility from HR-pQCT scans. It quantifies the regional variations and characterises the grayscale distribution. It is difficult to capture bone microarchitectural information from tibial CT images via visual inspection and interpret the clinical diagnosis. However, our proposed 3D LBP model provides an opportunity to quantify cortical and trabecular bone microarchitecture through statistical distributions. These distributions help identify patterns and variations in bone texture that are indicative of bone fragility, facilitating the differentiation between subjects with and without previous fractures. Compared to conventional deep learning methods that require massive training samples with manual annotations (Badgeley et al. (2019)), the 3D LBP model enables automated identification of previous fractures with only a small training set.

The occurrence of some uncertainties in practice such as noise and movement artefact may degrade the quality of CT images. Our segmentation model maintains excellent performance against noise present in HR-pQCT images. However, the 3D LBP model for image texture analysis is susceptible to noise. To address this, we develop an efficient and robust RCLBP descriptor (see Appendix C) to enhance the robustness of our discriminative system against noise present in HR-pQCT images.

As both cortical and trabecular regions in bone HR-pQCT images contribute to discriminating previous fractures, we also investigate the combination of texture features extracted from these two bone regions for fracture classification. Specifically, we propose two strategies to integrate the texture features of cortical and trabecular regions. The first one involves concatenating the feature vectors of cortical and trabecular regions and feeding them into the random forest classifier to discriminate previous fractures. The other strategy is to train two separate random forest classifiers using the texture features extracted from these two bone regions and then combining their outputs to produce the result. However, the highest AUC (0.75) for previous fracture results from using cortical regions as input data for our approach. Combining the texture features of cortical and trabecular regions does not improve the discriminative accuracy further. This may be because cortical features play a dominant role in fracture classification, overshadowing the contributions from trabecular features. In addition, there may be dependencies between cortical and trabecular features, potentially leading to redundancy rather than complementary information in the classification task.

There are also several limitations to this study. Firstly, the cortical regions in HR-pQCT scans are thin due to the distal site of measurement, and the manual annotations are not perfect with the current LabelMe tool. Especially around periosteal and endosteal surfaces in tibial CT slices, an accurate annotation is very difficult. As a result, this defect may affect the optimization of the U-Net model in the training process, potentially

lowering the discriminative performance of our approach for cortical regions. In the future, we aim to improve the automatic segmentation of cortical and trabecular regions and enhance model interpretability for fracture risk assessment. Also, this study only validates our approach in the elderly and British population. Future work should replicate our method in other cohorts and extend it to a larger and more diverse population, including young generations and subjects from other regions worldwide.

5.5 Summary

In this chapter, we propose an accurate and efficient method to segment cortical and trabecular regions of slices taken across CT scans of bones. The segmentation, coupled with 3D texture analysis performed separately on cortical and trabecular regions for feature extraction, shows the existence of discriminant information in both regions. Our work also shows the importance of significantly more information contained in the cortical-rich (compared to distal trabecular bone) sites. We also demonstrate the robustness of the segmentation and classification methods we proposed against significant levels of additive noise in HR-pQCT images.

Chapter 6

Multi-View Convolutional Neural Networks for Fracture Discrimination

6.1 Inspirations and Introduction

In Chapter 4, we demonstrate that volumetric texture analysis can be used to characterise bone microarchitecture based on a few of HR-pQT images for fracture discrimination. However, the proposed method may have limited generalisability to large populations. CNNs typically exhibit superior generalisability and robustness compared to statistical analysis methods when applied to large datasets (Krizhevsky et al. (2012)). Based on our work in Chapter 4 and that of others (Nishiyama et al. (2013), Mikolajewicz et al. (2020)), we investigate whether deep CNNs allow for characterising bone microarchitecture to identify previous fractures from HR-pQCT images in this chapter.

Over the decades, CNNs have been applied to clinical imaging to diagnose osteoporosis and assess fracture risk (Smets et al. (2021), Hsieh et al. (2021)). These algorithms can learn complex patterns and relationships in the image data that may not be easily discernible by human observers. By incorporating deep learning, it is possible to develop accurate and personalized fracture risk prediction models applied to HR-pQCT images, thereby reducing the burden of fractures on the healthcare system. Nevertheless, successful training of CNNs typically requires a large amount of annotated data (Tajbakhsh et al. (2016), Lu et al. (2022b)). Thus, automatic and accurate detection of bone fragility from HR-pQCT images remains challenging. Currently, CNNs have not yet been applied to bone HR-pQCT images for fracture risk prediction.

Transfer learning is an effective strategy to address the overfitting problem, especially

when the amount of labeled data is small. Pre-trained CNNs are generated from large-scale natural datasets with thousands of annotated samples, which have superior generalisation performance (Han et al. (2018)). Some studies suggest using pre-trained CNNs that learn features from large datasets and transfer them to the medical domain to enable automated diagnosis of diseases with limited data (Júnior et al. (2021), Huang et al. (2020)). Therefore, pre-trained CNNs have the potential to capture bone microarchitectural information from HR-pQCT images for fracture classification.

In this chapter, we propose an automatic method that employs multi-view CNNs to capture bone microarchitectural information from HR-pQCT images and then use the random forest classifier to identify previous fractures. Furthermore, we conduct numerical experiments to evaluate the discriminative performance of our method on the HCS dataset. The results of our study demonstrate that the image features extracted from tibial HR-pQCT scans via CNNs outperform DXA-measured BMD and clinical risk factors in fracture discrimination. This finding underscores the potential of our approach for application in clinical practice to enhance fracture risk prediction.

6.2 Method

6.2.1 Design and Overview

The present study here is designed to test the hypothesis that fracture risk is determined in part by feature representations of HR-pQCT images obtained from CNNs. Therefore, if image features are extracted from tibial HR-pQCT images of individuals using CNNs, those with previous fractures can be identified.

The framework of our method is illustrated in Figure 6.1. We employ a pre-trained CNN model on the ImageNet dataset with global average pooling across different feature maps to encode feature representations of CT slices. The high-level image features from multiple views of tibial HR-pQCT scans are integrated and fed into the random forest classifier to discriminate between subjects with and without previous fractures. The HCS dataset is used to evaluate the discriminative performance of our method. Data collection and data processing are detailed in Chapter 3.

6.2.2 Slice Selection

There are two common deep learning models for achieving 3D image classification (Su et al. (2015)). The first one is 3D CNNs, which can directly extract features from 3D volume data, but require high computational cost and large memory burden. In addition, not all parts of 3D data provide valuable signals for image classification tasks.

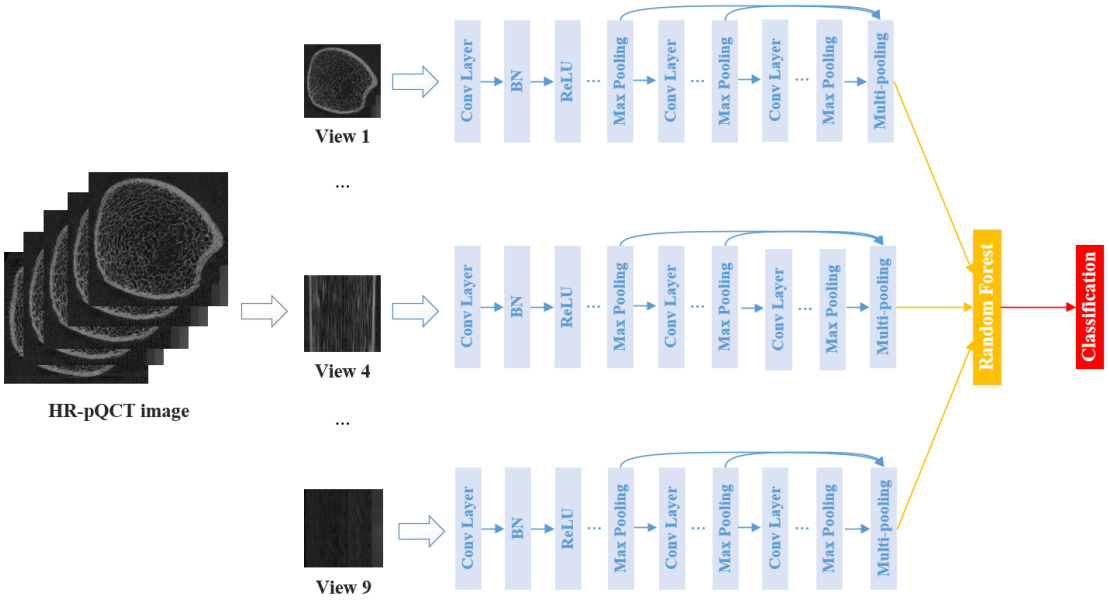

FIGURE 6.1: Multi-view convolutional neural networks for fracture classification. HR-pQCT slices taken from different views are encoded as feature representations and integrated to characterise bone microarchitecture.

The second model, multi-view CNNs, was proposed to address those issues by dividing 3D image data into multiple views. It uses the 2D CNN model to learn features from those slices for image classification. In this study, we adopt the latter model and propose to select nine CT slices of 3D volume data from various directions to extract image features and build a comprehensive feature representation. The procedure of slice selection is articulated in Algorithm 1.

Algorithm 1: Slice selection

Require: N: total number of 3D images having size $K \times W \times H$
Enquire: $Data[1{:}N, 1{:}M]$: selected slices from 3D image data, M: number of views
X_i is the i^{th} 3D image
for $i \leftarrow 1$ to N **do**:
  get X_i
  $Data[i, 1] \leftarrow X_i[1, :, :]$
  $Data[i, 2] \leftarrow X_i[K/2, :, :]$
  $Data[i, 3] \leftarrow X_i[K, :, :]$
  $Data[i, 4] \leftarrow X_i[:, 1, :]$
  $Data[i, 5] \leftarrow X_i[:, W/2, :]$
  $Data[i, 6] \leftarrow X_i[:, W, :]$
  $Data[i, 7] \leftarrow X_i[:, :, 1]$
  $Data[i, 8] \leftarrow X_i[:, :, H/2]$
  $Data[i, 9] \leftarrow X_i[:, :, H]$
end for
get $Data$

6.2.3 Feature Extraction

CNN acts as a feature extractor in image recognition, with low layers learning basic image features and high layers learning to identify the object in the image. Pre-trained CNNs based on transfer learning have been widely used and demonstrated excellent performance in many tasks (Han et al. (2018)). However, some studies indicate that the latter layers of pre-trained CNNs such as the softmax layer may exhibit inferior generalisation performance (Huang et al. (2020)). This problem is more serious in medical image analysis tasks due to the lack of sufficient labeled data to guide the network focusing on the local region of interest (Hu et al. (2021)). In addition, some valuable image information may be lost as the convolutional layers increase. To overcome these two obstacles, we adopt the ResNet-18 architecture pre-trained on the ImageNet dataset. Furthermore, we propose a multi-pooling strategy that combines image features from different feature maps, to effectively represent the features of CT slices, as shown in Figure 6.2.

For each selected slice of 3D images, we resize it into 224×224 and then use the 2D CNN model to extract image features. The CNN starts from a 7×7 convolutional layer, followed by batch normalization (BN), rectified linear unit (ReLU) and a max pooling layer. There are four basic modules in the network to extract image features, each of which comprises convolutional layers followed by BN and ReLUs. The output of each convolutional layer consists of a bank of feature maps. The feature map of the *l-th* convolutional layer is computed as:

$$F_l^c = \sigma\left(\sum_{m=1}^{M} (W_l^m \times F_{l-1}^c) + b_l^c\right) \tag{6.1}$$

where, F_{l-1}^c and F_l^c represent the input and output maps in the *c-th* channel respectively. M denotes the number of filters. W_l^m is the weight matrix of the *m-th* filter, and b_l^c is the bias. Function $\sigma(.)$ represents the ReLU function which is defined as:

$$\sigma(x) = \begin{cases} x & \mathrm{x} \geq 0 \\ 0 & \mathrm{x} < 0 \end{cases} \tag{6.2}$$

Then a max-pooling layer following the convolutional layer is introduced to select features. The output map of the *k-th* max-pooling layer is defined as follows:

$$Y_k^{(m,n),c} = \max h_\Omega^{(m,n),c} \tag{6.3}$$

where, $Y_k^{(m,n),c}$ represents the neuron at position (m, n) in the output map. h_Ω^c is the input map. Ω denotes the pooling region.

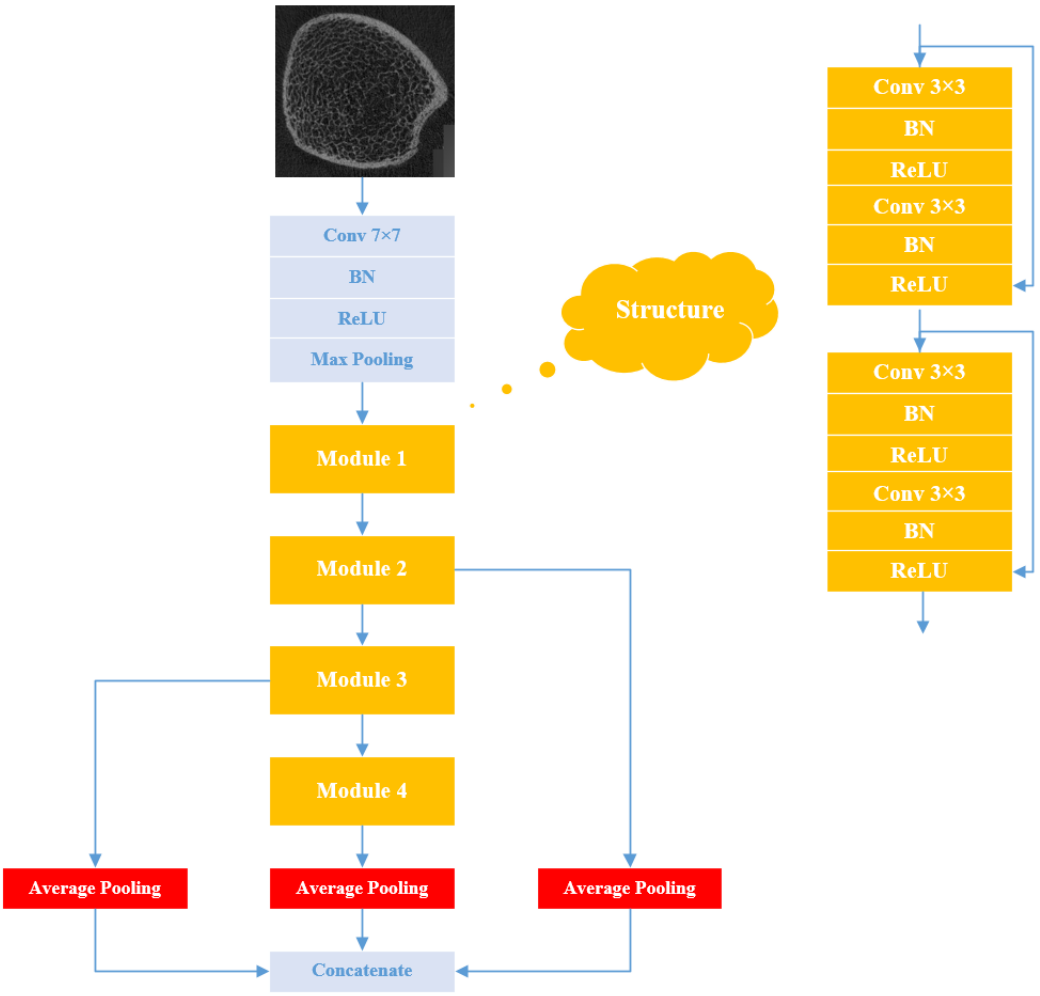

FIGURE 6.2: The architecture of the convolutional neural network for encoding feature representations of tibial HR-pQCT scans. Multi-pooling and concatenation operations are introduced to integrate feature maps from different convolutional layers.

In order to enrich global feature representations of 2D slices and reduce valuable information lost, we take full advantage of features extracted from different modules and integrate them to characterise the image. Here, we propose a multi-pooling strategy that uses the global average pooling operation to process the output feature maps of various modules. The global average pooling is defined as:

$$A_t^c = \frac{\sum_{(p,\,q)\in\Omega} s_{p,q}^c}{|\Omega|} \tag{6.4}$$

where, A_t^c denotes the output of the pooling operation in the *c-th* channel of the *t-th* module. $|\Omega|$ is the size of the pooling region. $s_{p,q}^c$ represents the element at position (p, q) in the input map.

We let A_t=[A_t^0, ..., A_t^c, ...] represent the image features derived from the *t-th* module

of the CNN model. As seen in Figure 6.2, we concatenate the feature vectors from different modules to form a vector f to characterise the 2D slice. Through extensive experimentation, we choose the image features derived from the latter three modules to produce f as follows:

$$f = A_2 \boxplus A_3 \boxplus A_4 \tag{6.5}$$

where, $\boxplus$ represents the concatenation operation.

6.2.4 Classification

In Figure 2.7 of Chapter 2, we illustrate the traditional CNN architecture, which utilises convolutional operations and the backpropagation algorithm to automatically learn image features for classification. When the training sample size is small, the softmax layer in the traditional CNN framework generally performs badly in identifying objects in the image. Some studies suggest utilising alternative strong classifiers, such as the SVM and conditional random field, instead of the softmax layer, to process the feature vectors from the fully connected layer to improve classification performance (Niu and Suen (2012)). In this study, we propose to use the ensemble learning classifier, random forest, for 3D image classification to reduce overfitting. Random forest can also reduce prediction variance by dividing the data into multiple subsets (Speiser et al. (2019)).

The detailed steps for dividing the training set, validation set and testing set are introduced in Section 4.2.3 of Chapter 4. Here, these feature vectors $\{V_1, ...V_i, ..., V_N\}$ produced by the CNN model are regarded as input data to the random forest classifier, where N is the number of samples. The feature vector V_i of 3D image S_i contains nine subsets $V_i^1, V_i^2, ..., V_i^9$, each of which characterises the image feature of the selected CT slice. During training, the feature vectors of all training samples are used to train the random forest classifier, and model parameters are updated. During testing, a participant is considered to have a fracture if the predicted fracture probability exceeds the threshold. The optimal threshold is determined by the Youden's Index (Youden (1950)) on the validation set.

6.2.5 Statistical Analysis

Participant characteristics are described using summary statistics in Section 3.3.1 of Chapter 3. The ROC curve and metrics such as AUC, sensitivity, specificity and accuracy, are utilised to evaluate the discriminative performance of our approach for previous fracture (Singh et al. (2017)). The 95% confidence interval (CI) for AUC is calculated to evaluate the uncertainty. The statistical significance difference in AUCs obtained from various methods is examined, and a p-value <0.05 is considered statistically significant.

We compare the discriminative performance of our approach with current fracture risk assessment techniques, including clinical risk assessment and DXA measurement. Our method, based on the HR-pQCT measurement, employs deep CNNs to extract features from tibial CT scans to characterise bone microarchitecture. The DXA measurement utilises femoral neck BMD values for fracture risk assessment. The clinical risk assessment uses 12 clinical covariates including age, sex, height, weight, BMI, dietary calcium, smoking history, alcohol consumption, physical activity, bisphosphonate usage, number of comorbidities and occupational social class. These two measurements are described in Chapter 4. Sensitivity analyses involve stratifying the analyses by sex for these three techniques.

6.3 Results

6.3.1 Parameter Settings

Our method and comparisons are implemented in Python (version 3.8) with Pytorch as the platform for deep learning models, running on a Windows 10 operating system with 8 GB RAM, Intel (R) Core (TM) i5-6600 3.30 GHz CPU. The architecture and parameter settings of our multi-view CNN model are described in Section 6.2.3 of this chapter. For each random forest, 50 decision trees using the Gini criterion are constructed.

6.3.2 Performance of Fracture Risk Assessment

Table 6.1 summarises the results of sensitivity, specificity and AUC (95% CI) from HR-pQCT measurement, clinical risk assessment and DXA measurement for fracture discrimination. The Youden's Index (Youden (1950)) is used to determine the optimal threshold and calculate specificity, sensitivity and accuracy using the predicted probability of fracture. Figure 6.3 and Figure 6.4 illustrate the classification accuracy and ROC curves of these three fracture risk assessment methods.

As shown in Table 6.1, compared to DXA measurement (AUC: 0.63, 95% CI: 0.54-0.69) and clinical risk assessment (AUC: 0.60, 95% CI: 0.52-0.67), HR-pQCT measurement demonstrates a higher discriminative performance (AUC: 0.75, 95% CI: 0.67-0.82) for previous fracture. The statistical analysis shows that there is a significant difference between the AUCs obtained from HR-pQCT measurement and both clinical risk assessment and DXA measurement (p-value <0.05). When the FPR is allowed to be 20%, clinical risk assessment and DXA measurement detect only 20% and 25% of individuals

TABLE 6.1: Discriminative performance of different fracture risk assessment methods for previous fracture.

Methods	AUC (95% CI)	Sensitivity	Specificity
DXA measurement	0.63 (0.54-0.69)	0.42	0.73
Clinical risk assessment	0.60 (0.52-0.67)	0.40	0.66
HR-pQCT measurement	**0.75 (0.67-0.82)**	**0.56**	**0.84**

DXA-measured femoral neck BMD is used.
Clinical risk assessment uses 12 clinical covariates such as age and gender.
HR-pQCT-measured tibial scans are used.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

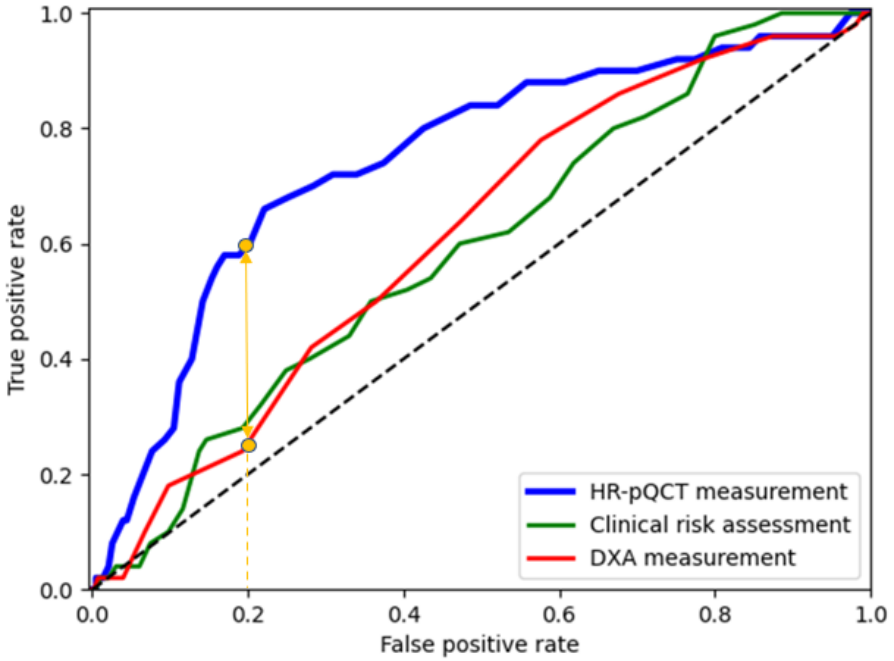

FIGURE 6.3: Receiver operating characteristic (ROC) curves for previous fracture from different fracture risk assessment methods. HR-pQCT-measured tibial scans are used. Notably, at a false positive rate of 20%, the true positive rate of the gold standard DXA is poor. HR-pQCT shows substantial improvement compared to DXA (depicted by the yellow line).

with previous fractures. However, HR-pQCT measurement significantly improves the TPR to 58%.

Figure 6.5 illustrates the classification results of our method compared to conventional deep learning architectures, including ResNet-18 (He et al. (2016)), VGG-19 (Simonyan and Zisserman (2014)) and 3D ResNet-18 (Hara et al. (2018)). These architectures are end-to-end deep neural networks that automatically learn image features using the backpropagation algorithm during training. It can be clearly seen that our model significantly outperforms other deep learning architectures in fracture discrimination. Conventional deep learning models, when trained on a few of HR-pQCT images, exhibit a severe overfitting problem, resulting in poor performance.

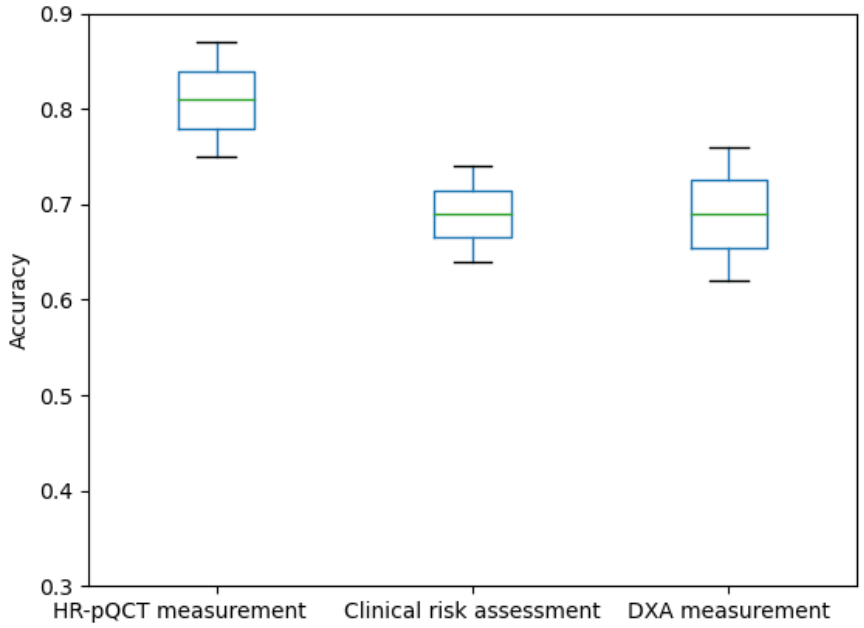

FIGURE 6.4: Classification accuracy of HR-pQCT measurement for previous fracture compared with clinical risk assessment and DXA measurement. HR-pQCT-measured tibial scans are used.

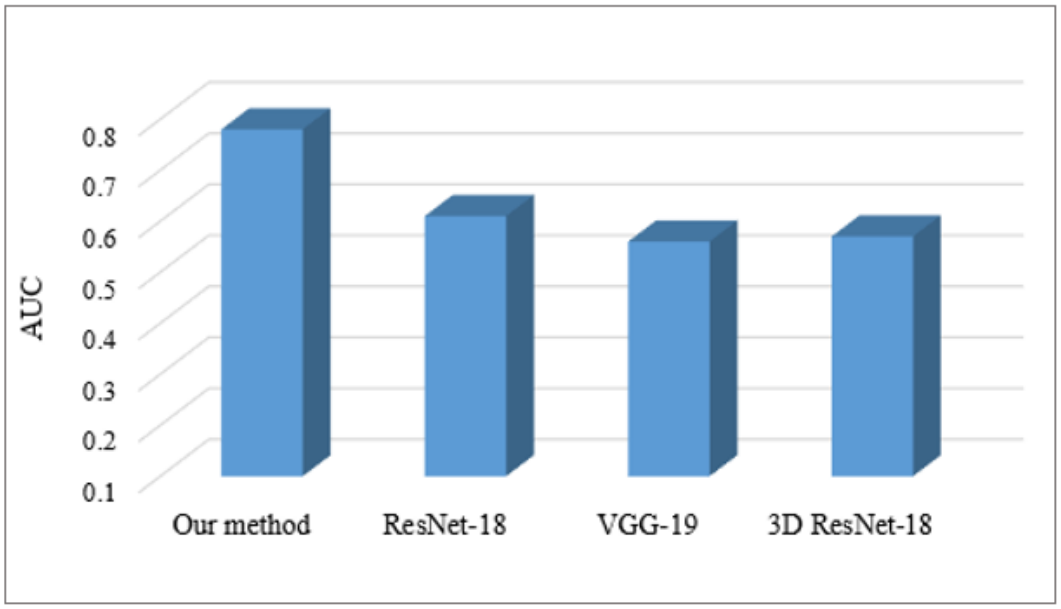

FIGURE 6.5: Discriminative performance of different deep learning models for previous fracture based on small samples. HR-pQCT-measured tibial scans are used.

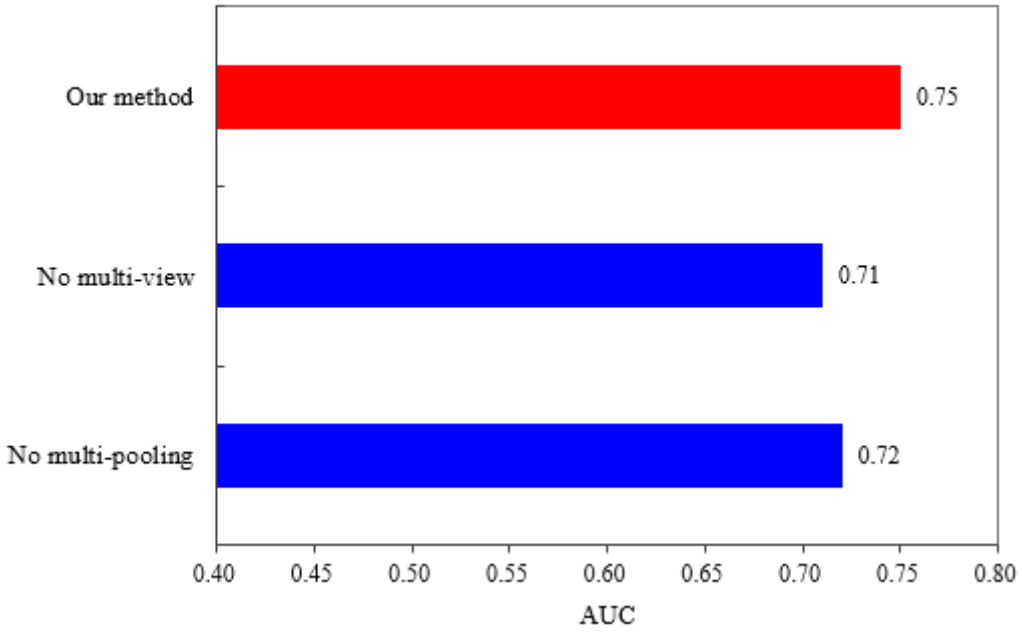

FIGURE 6.6: An ablation study of our method with different components for fracture classification. HR-pQCT-measured tibial scans are used.

Our proposed method contains two important components, and it is meaningful to analyse their contributions separately to the results. Therefore, we study the effects of replacing or removing components to provide additional insight into what makes the discriminative performance. In detail, we evaluate the effects of

- No multi-pooling: extracting image features of CT slices using only the feature maps from the final convolutional layer.
- No multi-view: removing the multiple views of 3D images and using the image feature extracted from the single view as input data to the random forest classifier.

Here we present an ablation study on fracture classification, and the discriminative performance of different methods is illustrated in Figure 6.6. We can find that each component of our method contributes to the overall performance. Our approach combines these two strategies to extract image features from HR-pQCT images, resulting in the highest AUC in fracture discrimination.

6.3.3 Sensitivity Analyses

Table 6.2 illustrates similar results in the stratification of analyses by sex. In this scenario, the use of HR-pQCT measurement improves discriminative accuracy for previous fracture compared to the use of DXA measurement and clinical risk assessment methods.

TABLE 6.2: Discriminative performance of different fracture risk assessment methods for previous fracture (sex-specific analyses).

Methods	Males		
	AUC (95% CI)	Sensitivity	Specificity
DXA measurement	0.64 (0.53-0.71)	0.13	0.78
Clinical risk assessment	0.58 (0.50-0.67)	0.23	0.79
HR-pQCT measurement	**0.69 (0.59-0.77)**	**0.30**	**0.90**
Methods	Females		
	AUC (95% CI)	Sensitivity	Specificity
DXA measurement	0.67 (0.55-0.75)	0.30	**0.83**
Clinical risk assessment	0.61 (0.50-0.71)	0.30	0.80
HR-pQCT measurement	**0.72 (0.61-0.82)**	**0.43**	**0.83**

DXA-measured femoral neck BMD is used.
Clinical risk assessment uses 12 clinical covariates such as age and gender.
HR-pQCT-measured tibial scans are used.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

6.4 Discussion

Accurate and reliable fracture risk assessment can facilitate the early diagnosis and treatment of bone fragility and reduce healthcare costs. In this study, we develop an automatic method for fracture discrimination from HR-pQCT images. Firstly, we employ a pre-trained CNN on the ImageNet dataset to extract multi-scale image features from CT slices. Then we integrate image features from multiple views of HR-pQCT scans and use the random forest classifier to discriminate between individuals with and without previous fractures. This study suggests that automated quantitative analysis of HR-pQCT imaging (AUC: 0.75, 95% CI: 0.67-0.82) improves discriminative accuracy for previous fracture compared to traditional methods of DXA measurement (AUC: 0.63, 95% CI: 0.54-0.69) and clinical risk assessment (AUC: 0.60, 95% CI: 0.52-0.67).

The discriminative performance of our approach is evaluated on the clinical dataset from the HCS. Numerical experiments demonstrate that our method which combines multi-view CNNs and random forest is feasible for automated fracture discrimination from HR-pQCT imaging. CNNs exhibit reliability and effectiveness in quantifying bone microarchitecture in HR-pQCT images. The results shown in Table 6.1 confirm our hypothesis that information obtained from HR-pQCT images through CNNs is associated with fracture risk.

An ablation study is conducted to evaluate the contribution of each component of our method, and it demonstrates that both of them contribute to the overall performance. Compared to traditional deep learning architectures (see Figure 6.5), our method demonstrates superior results in fracture classification based on a few of HR-pQCT images.

TABLE 6.3: Discriminative performance of different image feature extraction methods for previous fracture.

Methods	AUC (95% CI)	Sensitivity	Specificity
Multi-view CNNs	**0.75 (0.67-0.82)**	**0.56**	**0.84**
LBP-entire tibia	0.73 (0.65-0.79)	0.46	0.81
LBP-cortical regions	**0.75 (0.67-0.81)**	**0.56**	0.77
LBP-trabecular regions	0.66 (0.56-0.71)	0.48	0.69

The entire tibia, cortical regions and trabecular regions are from tibial HR-pQCT scans, and 3D LBP is conducted separately on each of them.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

Our model integrates multi-scale image features from multiple views of HR-pQCT scans to enhance feature representation of bone microarchitecture. In addition, our model adopts the random forest classifier to handle high-dimensional image features from the fully connected layer and performs better than the softmax in the few-shot image scenario.

In Chapter 4 and Chapter 5, we propose to use 3D LBP to extract features from the entire tibia, cortical regions and trabecular regions in HR-pQCT scans separately for fracture discrimination. Table 6.3 compares the discriminative performance of multi-view CNNs proposed in this chapter with LBP in the above three scenarios. Unlike LBP, which focuses on local texture characteristics, multi-view CNNs require contextual and global information of the image to encode feature representations. Therefore, we utilise the entire tibial HR-pQCT scans as input data for multi-view CNNs to avoid losing valuable information. The corresponding results are presented in Table 6.3. Multi-view CNNs demonstrate a similar AUC result to LBP when using cortical regions segmented from CT scans as input data. However, when the entire tibia or trabecular regions are used as input data for 3D LBP, multi-view CNNs yield a higher AUC for fracture classification. This comparison also highlights that besides bone texture features, other image features are associated with previous fractures. LBP is effective in capturing bone texture features, while multi-view CNNs capture comprehensive image features from HR-pQCT scans. As a result, multi-view CNNs can leverage diverse image features to enhance the feature representations of bone microarchitecture and have the potential to improve fracture classification accuracy. However, since incorrectly labeled samples in the data could also impact model performance, we further conduct a comparison analysis between multi-view CNNs and LBP in Section 7.4 of Chapter 7.

However, our study also has several limitations. One limitation is that the clinical dataset collected from the HCS is small, which may affect the generalisation and classification performance of our model. Another limitation is the imbalance in the number of participants with previous fractures (n=46) compared to those without (n=121). Therefore, we have to employ an under-sampling strategy for the non-fracture group to

balance the data. In addition, the dataset contains incorrectly labeled samples, such as instances where fracture occurrence is not caused by bone fragility, potentially leading the classifier to learn incorrect patterns between fracture and non-fracture groups.

6.5 Summary

In this chapter, we propose to use deep CNNs to characterise bone microarchitecture from multiple views of HR-pQCT images and then integrate these high-level image features to identify subjects with previous fractures. Our proposed automated method is able to capture richer information from HR-pQCT imaging beyond BMD and clinical risk factors for fracture discrimination. This approach has the potential to improve the accuracy and reliability of bone fragility detection in clinical practice and reduce the clinical workloads.

Chapter 7

An Enhanced and Robust Fracture Discrimination System

7.1 Inspirations and Introduction

Our work in Chapter 4 and Chapter 6 shows that computer vision approaches (where the entire HR-pQCT scan is 'read' by computer algorithms) to determine fracture discrimination provide benefits above the use of clinical risk factors or DXA-measured BMD. However, as introduced in Chapter 2, fracture is affected by multiple factors aside from bone fragility (Litwic (2020)). Our current datasets lack the ground truth to label non-fractured healthy and osteoporotic fracture subjects correctly. Fracture history was determined via self-report or vertebral fracture assessment, and HR-pQCT scans lacked manual review by clinicians. Two groups of individuals included in our datasets cannot be accurately identified through image analysis, potentially lowering the performance of our discriminative systems. One group comprises healthy subjects who have suffered from traumatic fractures. The other group includes subjects with bone fragility who have not experienced a fracture. Samples from these two groups are referred to as incorrectly labeled samples or incorrectly labeled data in this thesis. Therefore, there is still room for improvement in the accuracy of fracture discrimination.

We hypothesise that there are differences in bone microarchitecture in HR-pQCT images between fracture and non-fracture groups. Therefore in order to discriminate fragility fractures, in our dataset, we require two distinct groups: i) fractured patients due to fragility and ii) non-fractured healthy individuals. In this study, our objective is to filter out incorrectly labeled samples to enhance the accuracy and robustness of

our discriminative system in the quantitative analysis of HR-pQCT imaging. DXA-measured BMD is widely utilised for assessing bone health and provides an opportunity to label non-fractured healthy individuals and those with osteoporotic fractures within the population.

We observe from Chapter 4 that specific DXA-measured BMD and T-score values can be used for fracture discrimination. In the absence of clinical ground truth to form our training dataset containing two groups with correctly labeled samples of fractured patients with fragilities and non-fractured healthy individuals, we exploit T-scores to form our training dataset. Therefore, we propose to employ adaptive T-score thresholds to filter out incorrectly labeled samples from the raw dataset. Subsequently, we use multi-view CNNs to characterise bone microarchitecture in HR-pQCT images to discriminate between non-fractured healthy subjects and subjects with osteoporotic fractures. Our numerical experiments demonstrate that using DXA BMD to filter out incorrectly labeled samples significantly improves the accuracy and robustness of our discriminative system. When evaluated on an independent cohort, our model can maintain high discriminative performance. In the absence of clinical ground truth, our approach that exploits DXA BMD and HR-pQCT images introduces a learning system to improve fracture discrimination compared to HR-pQCT and DXA alone.

7.2 Method

7.2.1 Design and Overview

The framework of our method is illustrated in Figure 7.1. We quantitatively analyse bone microarchitecture in HR-pQCT images using multi-view CNNs (as described in Chapter 6) and generate adaptive optimal T-score thresholds to categorise individuals into non-fractured healthy and osteoporotic fracture groups. After filtering out incorrectly labeled data, image features are extracted from CT scans and fed into the random forest classifier to discriminate osteoporotic fractures.

The HCS dataset is used to evaluate the discriminative performance of our method. The GLOW dataset from an independent cohort is also used to evaluate the robustness of our discriminative system. The data collection and data processing of these two datasets are detailed in Chapter 3.

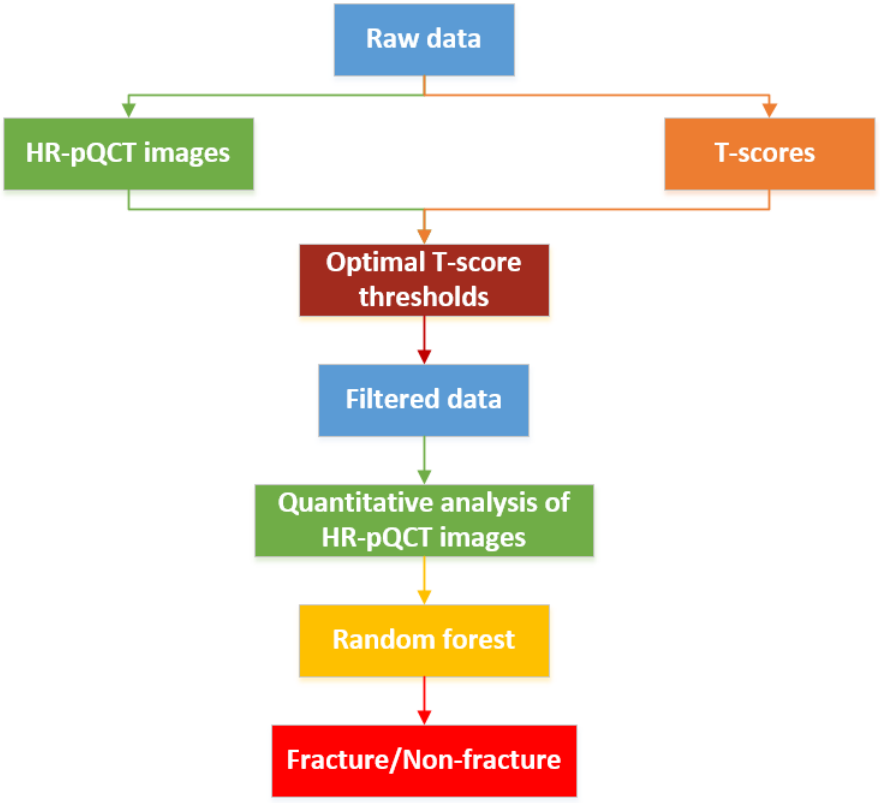

FIGURE 7.1: An adaptive threshold strategy to enhance fracture discrimination. DXA-measured T-scores are used as the ground truth to select samples in the raw data.

7.2.2 Image Feature Extraction

Building upon the correlation between microarchitectural deterioration of bone tissue and fracture risk, we propose to use volumetric texture analysis and deep CNNs techniques to characterise bone microarchitecture in HR-pQCT images for fracture classification. The details of these two approaches are described in Chapter 4 and Chapter 6. In this chapter, we adopt multi-view CNNs (as described in Chapter 6) to extract image features from HR-pQCT scans and use DXA-measured T-score information to select correctly labeled samples for a correct performance analysis of our system. In addition, we compare the discriminative performance of multi-view CNNs and LBP for fracture discrimination in both the HCS and GLOW datasets in Section 7.4.

7.2.3 Adaptive T-score Thresholds

Fracture occurrence is a binary random variable. Such a random variable would depend on how careful an individual is to avoid a traumatic accident. For example, the

discrimination of traumatic fractures through quantitative analysis of HR-pQCT imaging is not feasible. This is due to the fact that the traumatic fractures of healthy individuals are completely random. Moreover, even if bone fragility exists in the human body, it may not cause a fracture if the individual has not experienced a traumatic incident. Instances from these two scenarios are considered as incorrectly labeled data and cannot be accurately discriminated due to the random nature of fracture occurrence. Furthermore, including these incorrectly labeled samples in the training set may weaken the classifiers' capability to discriminate between fracture and non-fracture groups.

A high BMD indicates a healthy and strong bone structure, offering increased strength and resistance to compression, thereby reducing the risk of fractures. Conversely, low BMD suggests fragility in bone structure or loss of bone mass. As depicted in Figure 7.2, the raw data from the HCS and GLOW cohorts include some incorrectly labeled samples. Some individuals with high T-scores experienced fractures, and others with low T-scores did not. These instances are considered incorrectly labeled samples and would lower the performance of our discriminative systems. Therefore, filtering out incorrectly labeled data from the original cohorts has the potential to enhance the discriminative performance of our model.

T-score is widely used as the gold standard to distinguish between osteoporosis, osteopenia and healthy subjects in clinical practice, with reference to the diagnostic thresholds shown in Table 2.1 (Watts (2004)). However, to the best of our knowledge, there are no established standards to categorise individuals into healthy and osteoporotic fracture groups. Here, we propose to use the T-score as a tool to select non-fractured healthy individuals and osteoporotic fracture patients in the raw data. However, the specific thresholds for categorising are unknown.

Inspired by diagnostic criteria for osteoporosis, we propose two adaptive T-score thresholds denoted as α and β to categorise individuals into non-fractured healthy and osteoporotic fracture groups. The steps followed to generate optimal thresholds are listed in Algorithm 2. α and β are as hyperparameters added into our discriminative system to filter out incorrectly labeled samples from the raw data. Our proposed image feature extraction method, as described in Chapter 6, is then applied to the filtered data to discriminate between non-fractured healthy subjects and subjects with osteoporotic fractures. Starting from the lowest value for the T-score, the T-score thresholds increase incrementally to compute AUC for each value of the threshold. The T-score thresholds that correspond to the highest AUC in fracture discrimination are selected as the optimal thresholds.

Algorithm 2: Adaptive thresholds

Require: N: total number of participants in the training and validation sets
X_i is the DXA-measured T-score of participant i
Y_i is the fracture history of participant i (0: non-fracture, 1: fracture)
S_i is the HR-pQCT image data of participant i
m, n : the lowest and highest values for the T-score
Enquire: α, β: T-score thresholds
$Data$: selected samples
$best_{auc}$: the best AUC result for fracture discrimination

$best_{auc} \leftarrow 0.5$
for $\beta \leftarrow m$ to n **do**:
  for $\alpha \leftarrow \beta$ to n **do**:
    for $i \leftarrow 1$ to N **do**:
      get X_i, Y_i, S_i
      if Y_i=0:
        if $X_i > \alpha$:
          calculate image features for S_i
          add this sample into $Data$
        else:
          remove this sample
      if Y_i=1:
        if $X_i \leq \beta$:
          calculate image features for S_i
          add this sample into $Data$
        else:
          remove this sample
    end for
    divide $Data$ into a training set and a validation set
    train the classifier and calculate the AUC on the validation set
    if AUC $> best_{auc}$
      select α and β as the current optimal thresholds
      $best_{auc} \leftarrow$ AUC
  end for
end for
get optimal thresholds

7.2.4 Classification

In the HCS dataset, there are fewer participants with previous fractures (n=46) compared to those without (n=121). Each participant included in this dataset only has one tibial scan. Unbalanced data typically biases the classifier towards the class with larger samples. Therefore, an under-sampling strategy is utilised for participants without previous fractures to balance the data (Lin et al. (2017)). Then 80% of tibial HR-pQCT images from the balanced data are selected randomly as the training set, while the remaining data is equally divided into the validation and testing sets. The experiments are repeated ten times to evaluate performance. The T-score thresholds that enable our approach to produce the highest AUC in fracture classification on the validation set are

determined as the adaptive optimal thresholds, which are then used as the criteria to filter the HCS dataset. The discriminative performance of our approach for previous fracture is evaluated on the testing set.

The GLOW dataset comprises 84 participants with previous fractures and 297 without. Each participant included in this dataset only has one tibial scan. An under-sampling strategy is also utilised for the non-fracture group to balance the data. The same data partitioning method as that of the HCS dataset is used to divide the training set, validation set and testing set of the GLOW dataset. The adaptive optimal T-score thresholds for the GLOW cohort are determined on the validation set. Both internal and external tests are conducted on the testing set of the GLOW dataset to evaluate the discriminative performance of our approach for previous fracture. In the internal test, both the training and testing sets are drawn from the GLOW dataset. In the external test, the model is trained on the HCS dataset and then tested on the testing set of the GLOW.

7.2.5 Statistical Analysis

Participant characteristics are described using summary statistics in Section 3.3 of Chapter 3. The ROC curve and metrics such as AUC, sensitivity, specificity and accuracy are utilised to evaluate the discriminative performance of our approach for previous fracture. The 95% confidence interval (CI) for AUC is calculated to evaluate the uncertainty. The statistical significance difference in AUCs obtained from various fracture risk assessment methods is examined, and a p-value <0.05 is considered statistically significant.

In both the HCS and GLOW datasets, the discriminative performance of HR-pQCT measurement is compared with traditional methods of clinical risk assessment and DXA measurement. These two methods are described in Chapter 6. Clinical risk factors for the HCS dataset include age, sex, height, weight, BMI, dietary calcium, smoking history, alcohol consumption, physical activity, bisphosphonate usage, number of comorbidities and occupational social class. Similarly, clinical risk factors for the GLOW dataset comprise age, sex, BMI, current smoker, alcoholic drinks, cortisone or prednisone usage, rheumatoid history, colitis history, diabetes history, coeliac history and premature menopause.

In addition, the discriminative capacity of our approach with and without DXA BMD is compared. Of particular interest is whether the AUC for our approach with the adaptive threshold strategy is substantially greater than the AUC for that without.

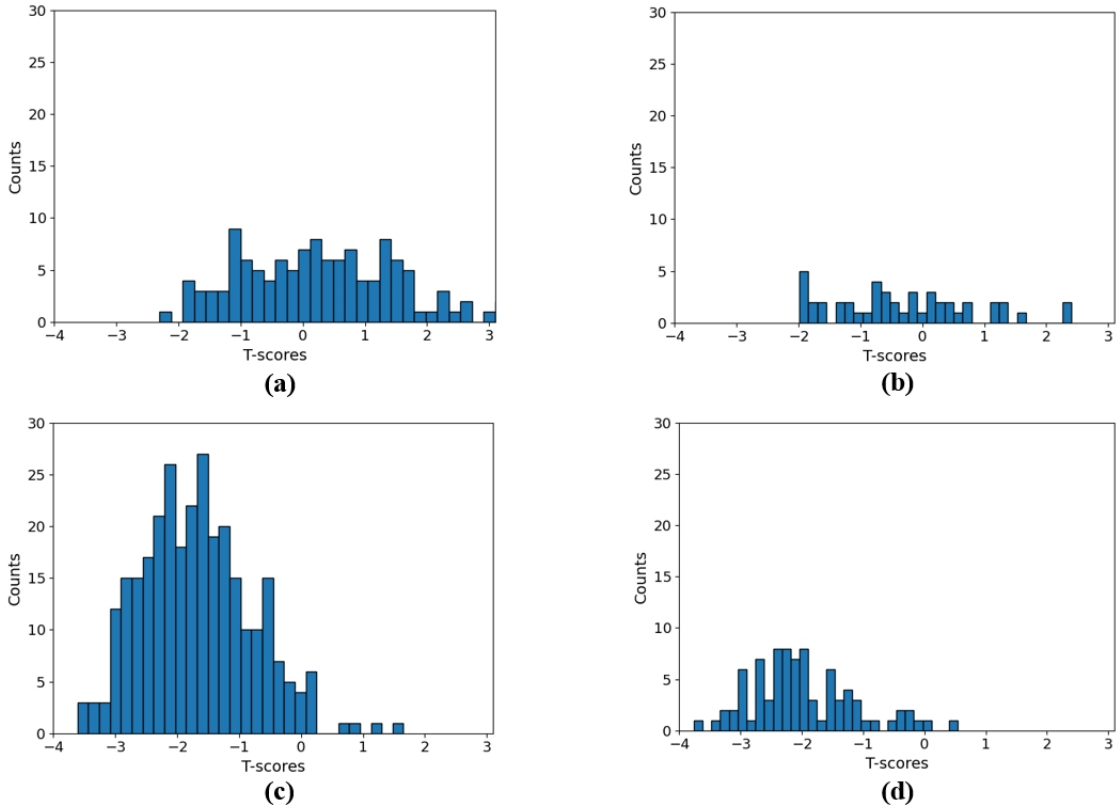

FIGURE 7.2: Comparative analysis of T-score distributions between the Hertfordshire Cohort Study and Global Longitudinal Study of Osteoporosis in Women. (a) and (b) are the non-fracture and fracture groups from the former dataset, while (c) and (d) are the non-fracture and fracture groups from the latter dataset.

7.3 Results

7.3.1 Comparative Analyses

Distributions of femoral neck BMD T-scores for the HCS and GLOW datasets are shown in Figure 7.2. The T-scores for participants in both the HCS and GLOW cohorts were derived using NHANES III data and Equation 2.1 in Chapter 2 (Ward et al. (2023)). DXA measurement is considered the gold standard for assessing BMD and diagnosing osteoporosis in clinical practice. In Figure 7.2, both the HCS and GLOW datasets reveal participants with previous fractures overall tend to have lower T-scores than those without fractures. However, there is a discernible difference between the T-score distributions of these two datasets. Participants in the HCS exhibit lower levels of bone loss or bone deterioration compared to those in the GLOW. As a result, a variance would exist in the feature distributions of HR-pQCT images between the HCS and GLOW datasets, potentially lowering the discriminative performance of our model. Domain adaption is used to address this issue, as described in Appendix D.

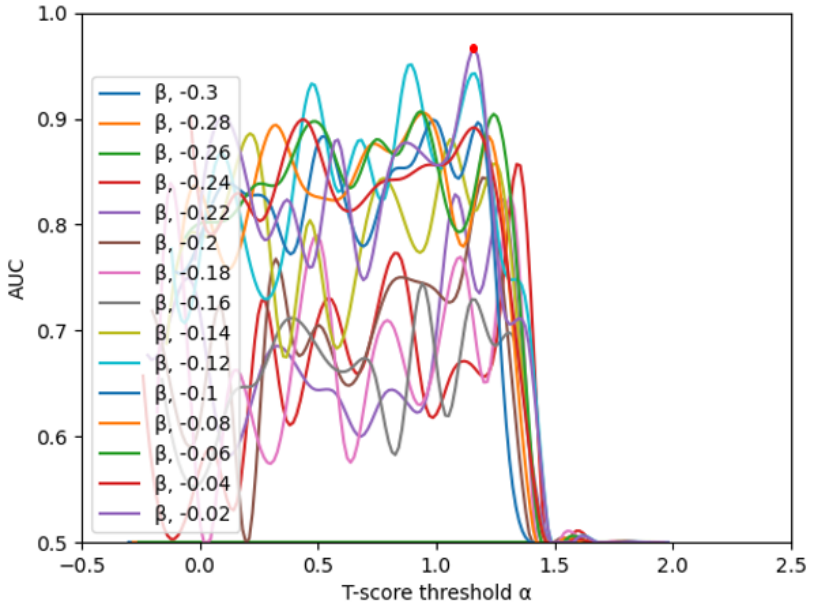

FIGURE 7.3: Determining optimal T-score thresholds through quantitative analysis of HR-pQCT images. The red point has the highest AUC on the validation set of the HCS dataset, and its corresponding T-score thresholds of α=1.18, β=-0.02 are selected.

7.3.2 Performance of Fracture Risk Assessment in the Hertfordshire Cohort Study

Our discriminative system employs T-score thresholds α and β (as described in Algorithm 2 in Section 7.2.3) to select samples from the raw data and quantitatively analyse HR-pQCT images to discriminate between non-fractured healthy subjects and subjects with osteoporotic fractures. Figure 7.3 illustrates the AUC results of our approach on the validation set according to specific T-score thresholds. When α=1.18 and β=-0.02, our approach produces the highest AUC in fracture discrimination; therefore, these values are considered the adaptive optimal T-score thresholds for the HCS dataset. We apply optimal thresholds to select non-fractured healthy and osteoporotic fracture subjects in the HCS dataset (see Figure 7.4), and present the discriminative performance of different methods.

The sensitivity, specificity and AUC (95% CI) results from various fracture risk assessment methods on the raw and filtered datasets from the HCS are summarised in Table 7.1. Specificity, sensitivity and accuracy are calculated using the predicted probability of fracture at the optimal threshold, determined by the Youden's Index (Youden (1950)). The ROC curves of HR-pQCT measurements with and without using adaptive T-score thresholds to filter out incorrectly labeled samples are illustrated in Figure 7.5. A comparison of classification accuracy between HR-pQCT measurement and clinical risk assessment using the filtered dataset is presented in Figure 7.6.

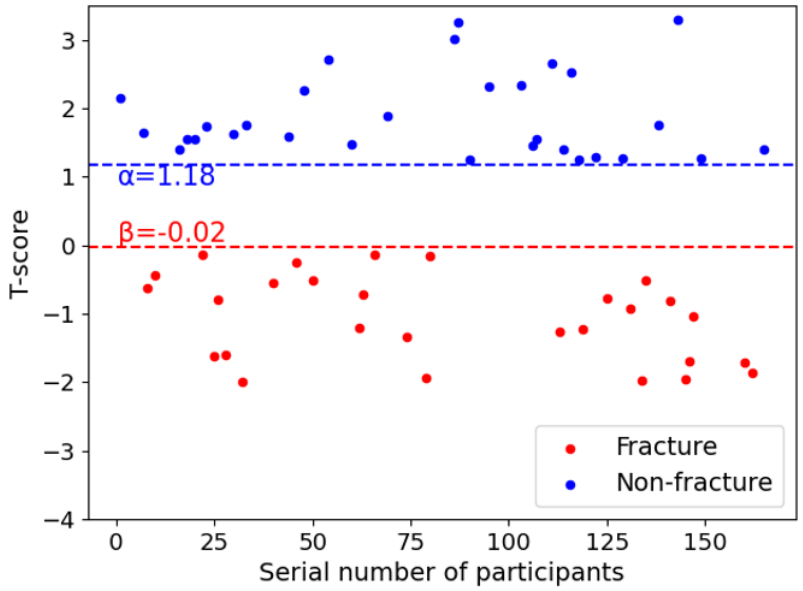

FIGURE 7.4: Adaptive T-score thresholds for selecting non-fractured healthy and osteoporotic fracture subjects in the Hertfordshire Cohort Study dataset.

In the raw dataset from the HCS, traditional methods of DXA measurement and clinical risk assessment approaches used to discriminate previous fractures yield AUCs of 0.63 (95% CI: 0.54-0.69) and 0.60 (95% CI: 0.52-0.67) respectively. The quantitative analysis of HR-pQCT images with our multi-view CNN model improves the AUC to 0.75 (95% CI: 0.67-0.82). Furthermore, after filtering the raw dataset with adaptive T-score thresholds, the AUC of HR-pQCT measurement improves to 0.90 (95% CI: 0.82-0.97), as shown in Figure 7.5. The sensitivity and specificity of this approach are 0.80 and 0.87. The statistical analysis reveals a significant difference between the AUCs obtained from HR-pQCT measurements with and without DXA BMD filtering (p-value $<$0.001). Samples in the filtered dataset are selected using DXA-measured T-scores as the ground truth; therefore, DXA measurement should not be assessed for previous fracture. Table 7.1 and Figure 7.6 demonstrate that HR-pQCT measurement has a higher discriminative performance than clinical risk assessment in identifying individuals with previous fractures in the filtered dataset.

7.3.3 Performance of Fracture Risk Assessment in the Global Longitudinal Study of Osteoporosis in Women

Similar to the HCS dataset, our discriminative system employs adaptive T-score thresholds α and β to analyse the GLOW dataset. When α=-0.95 and β=-2.35, our approach produces the highest AUC on the validation set for fracture classification; therefore, these values are considered the adaptive optimal T-score thresholds for the GLOW dataset. As depicted in Figure 7.7, we apply optimal thresholds to filter out incorrectly labeled data and retain non-fractured healthy and osteoporotic fracture subjects in the

TABLE 7.1: Discriminative performance of different methods for previous fracture using the raw and filtered datasets from the Hertfordshire Cohort Study.

Methods	Raw dataset		
	AUC (95% CI)	Sensitivity	Specificity
DXA measurement	0.63 (0.54-0.69)	0.42	0.73
Clinical risk assessment	0.60 (0.52-0.67)	0.40	0.66
HR-pQCT measurement	**0.75 (0.67-0.82)**	**0.56**	**0.84**
Methods	Filtered dataset		
	AUC (95% CI)	Sensitivity	Specificity
Clinical risk assessment	0.78 (0.68-0.91)	0.47	**0.97**
HR-pQCT measurement	**0.90 (0.82-0.97)**	**0.80**	0.87

DXA-measured femoral neck BMD is used.
Clinical risk assessment uses 12 clinical covariates such as age and gender.
HR-pQCT measurement uses multi-view CNNs to capture bone microarchitectural information from tibial scans.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

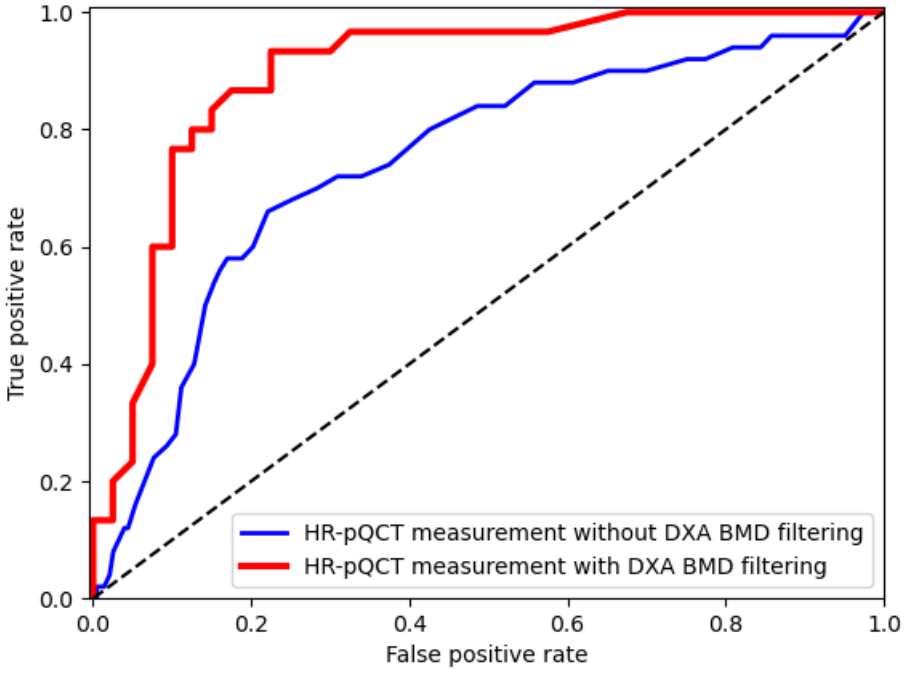

FIGURE 7.5: Comparative analysis of HR-pQCT measurements with and without DXA BMD filtering for fracture classification. HR-pQCT measurements use multi-view CNNs to capture bone microarchitectural information from tibial scans. Note, while the AUC is the usual performance metric (shown in Table 7.1), the difference in positive predictive performance at low false positives shows substantial improvement by incorporating DXA BMD filtering.

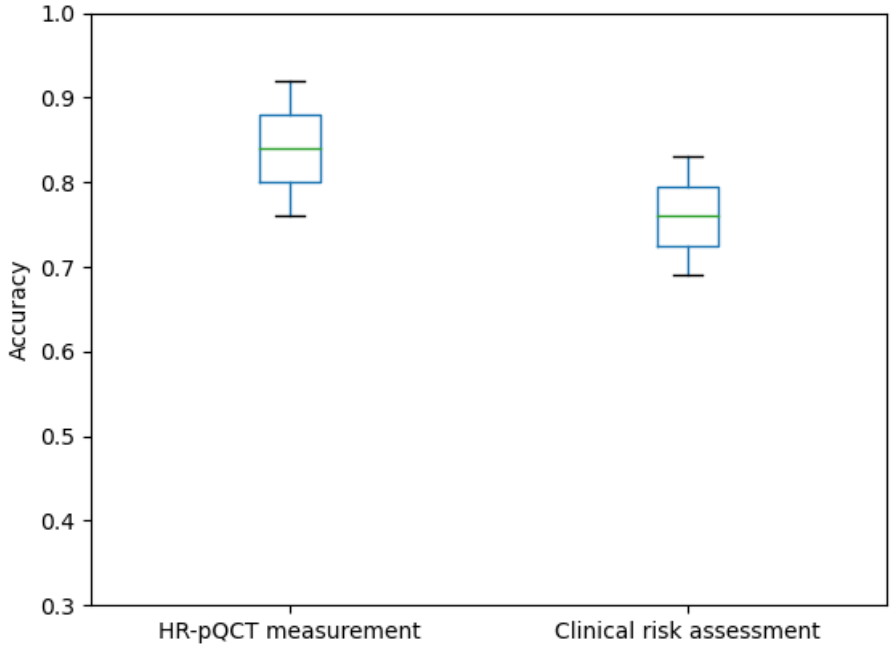

FIGURE 7.6: Classification accuracy of HR-pQCT measurement and clinical risk assessment for previous fracture using the filtered dataset from the Hertfordshire Cohort Study. HR-pQCT measurement uses multi-view CNNs to capture bone microarchitectural information from tibial scans.

GLOW dataset. Subsequently, we present the discriminative performance of different methods.

Both internal and external tests are conducted to evaluate the discriminative capability of our approach on the testing set of the GLOW dataset. During training, the internal test uses the training set of the GLOW, while the external test utilises the HCS dataset.

In the internal test, similar to the HCS, the sensitivity, specificity and AUC (95% CI) results from various fracture risk assessment methods on the raw and filtered datasets from the GLOW are summarised in Table 7.2. Specificity, sensitivity and accuracy are calculated using the predicted probability of fracture at the optimal threshold, determined by the Youden's Index (Youden (1950)).

Similar results to the HCS dataset are observed in Table 7.2. In the raw dataset from the GLOW, compared to DXA measurement (AUC: 0.60, 95% CI: 0.53-0.67) and clinical risk assessment (AUC: 0.57, 95% CI: 0.51-0.63), HR-pQCT measurement demonstrates a higher classification accuracy (AUC: 0.65, 95% CI: 0.58-0.70) for fracture discrimination. Furthermore, after filtering out incorrectly labeled samples with DXA BMD, the AUC of HR-pQCT measurement significantly improves to 0.94 (95% CI: 0.89-0.97).

The statistical analysis shows that there is a significant difference between the AUCs obtained from HR-pQCT measurements with and without DXA BMD filtering (p-value <0.001). In addition, when filtering out incorrectly labeled data from the original

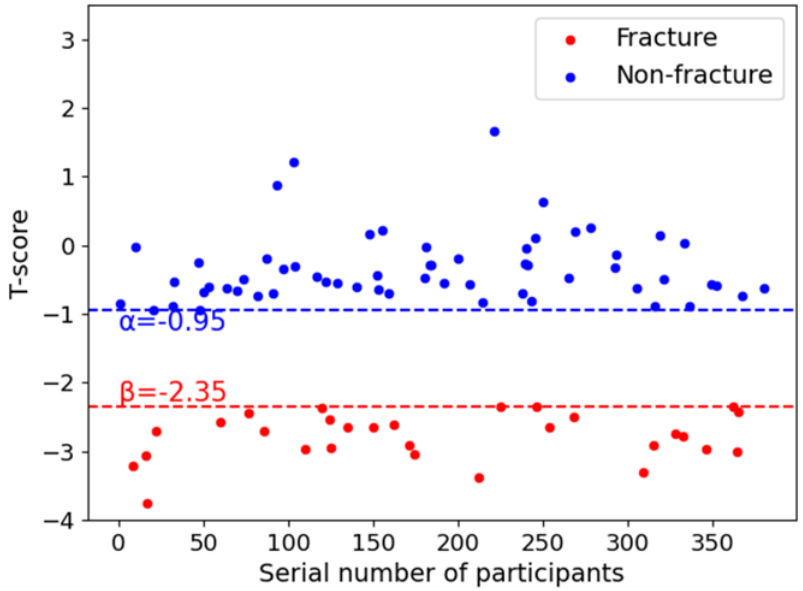

FIGURE 7.7: Adaptive T-score thresholds for selecting non-fractured healthy and osteoporotic fracture subjects in the Global Longitudinal Study of Osteoporosis in Women dataset.

TABLE 7.2: Discriminative performance of different methods for previous fracture using the raw and filtered datasets from the Global Longitudinal Study of Osteoporosis in Women dataset (internal test).

Methods	Raw dataset		
	AUC (95% CI)	Sensitivity	Specificity
DXA measurement	0.60 (0.53-0.67)	0.70	0.51
Clinical risk assessment	0.57 (0.51-0.63)	0.56	0.54
HR-pQCT measurement	**0.65 (0.58-0.70)**	**0.77**	**0.74**
Methods	Filtered dataset		
	AUC (95% CI)	Sensitivity	Specificity
Clinical risk assessment	0.77 (0.67-0.83)	0.63	0.64
HR-pQCT measurement	**0.94 (0.89-0.97)**	**0.77**	**0.86**

DXA-measured femoral neck BMD is used.
Clinical risk assessment uses 12 clinical covariates such as age and BMI.
HR-pQCT measurement uses multi-view CNNs to capture bone microarchitectural information from tibial scans.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

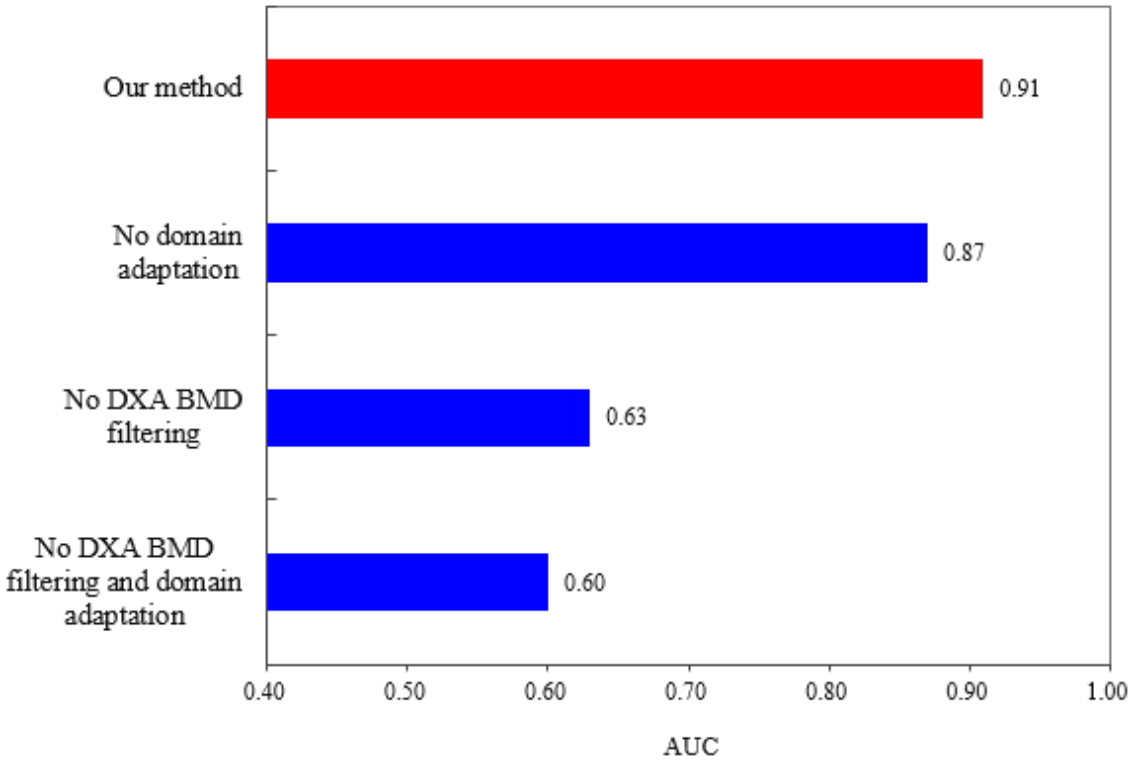

FIGURE 7.8: An ablation study of our method with different components for fracture classification in the external test scenario. HR-pQCT measurement uses multi-view CNNs to capture bone microarchitectural information from tibial scans.

GLOW dataset using adaptive T-score thresholds, a statistically significant difference is found between the AUCs obtained from HR-pQCT measurement and clinical risk assessment (p-value <0.05).

In the external test, the AUC results of our method under different scenarios are presented in Figure 7.8. The prediction model (random forest) is trained on the entire HCS dataset and subsequently evaluated for discriminative performance on the testing set of the GLOW dataset. Two important techniques (DXA BMD filtering and domain adaptation) are included in our discriminative system to enhance its accuracy and robustness for previous fracture. As described in Appendix D, domain adaptation involves retraining the prediction model on the validation set of the GLOW dataset to adapt it to the target domain. We further study the effects of removing techniques to provide additional insight into what makes the discriminative performance. In detail, we evaluate the effects of

- Our method: using DXA BMD filtering for sample selection in the HCS and GLOW datasets; training the prediction model on the filtered HCS dataset; retraining the model on the validation set of the GLOW dataset.

- No domain adaptation: using DXA BMD filtering for sample selection in the HCS and GLOW datasets, and then training the prediction model only on the filtered HCS dataset.
- No DXA BMD filtering: training the prediction model on the raw HCS dataset, and then retraining the model on the validation set of the raw GLOW dataset.
- No DXA BMD filtering and domain adaptation: training the prediction model only on the raw HCS dataset, and then testing the model on the raw GLOW dataset.

Figure 7.8 illustrates that each technique within our method contributes to the overall performance. Notably, the adaptive threshold strategy, which eliminates incorrectly labeled samples from the raw data, significantly enhances the classification performance of our model. In the best-case scenario where all incorrectly labeled samples are removed from the HCS and GLOW datasets and also domain adaptation is used, our discriminative system achieves an AUC of 0.91 when evaluated on an independent cohort. However, the presence of incorrectly labeled instances in clinical practice would diminish the discriminative accuracy of our approach. When considering all incorrectly labeled samples, the AUC of our approach (no DXA BMD filtering) is only 0.63.

In incorrectly labeled samples, non-fractured healthy subjects who have experienced traumatic fractures would be misclassified as the non-fracture group through quantitative analysis of HR-pQCT images, hindering physicians in providing fracture prevention and intervention for these subjects. In contrast, individuals with bone fragility who have not experienced a fracture would be misclassified as the fracture group by our model. It is important to note that these instances, where non-fracture patients with bone fragility are misclassified as the fracture group, are not the errors of our system. Instead, these misclassified samples can be considered as future predictions for such patients to indicate that a fracture in the future is likely to happen, although such a fracture has not occurred yet. Therefore, without filtering out such samples in the testing set of the GLOW dataset, we evaluate the performance of our discriminative system. The results demonstrate that our method achieves an AUC of 0.85, and the sensitivity and specificity are 0.76 and 0.85.

7.4 Discussion

In this study, we propose an enhanced and robust discriminative algorithm that quantitatively analyses HR-pQCT images to discriminate between non-fractured healthy subjects and subjects with osteoporotic fractures. The highest AUCs were obtained for

both the HCS (0.90, 95% CI: 0.82-0.97) and GLOW (0.94, 95% CI: 0.89-0.97) datasets using a combination of HR-pQCT measurement and DXA BMD filtering; HR-pQCT measurement has a higher discriminative accuracy for previous fracture than clinical risk assessment and DXA measurement; inclusion of DXA BMD filtering significantly improves fracture discrimination compared to HR-pQCT alone. This suggests that valuable information regarding fracture risk is utilised by combining HR-pQCT imaging and DXA BMD, which provides benefits beyond performance with clinical risk factors, BMD or HR-pQCT alone.

The results summarised in Table 7.1 and Table 7.2 indicate that using DXA-measured T-scores as the ground truth for sample selection in the raw data significantly enhances the discriminative capability of our model. To the best of our knowledge, there are currently no established standards for categorising individuals into non-fractured healthy and osteoporotic fracture groups. Through quantitative analysis of HR-pQCT images using our approach, optimal T-score thresholds can be determined. This reveals a correlation exists between BMD and bone microarchitecture assessments. These findings offer valuable references for clinicians in identifying individuals at high risk of osteoporotic fracture.

The adaptive threshold strategy plays an important role in our enhanced discriminative system, which provides crucial training datasets in the absence of clinical ground truth. Such a filtering system generates optimal T-score thresholds tailored to specific cohorts, effectively separating non-fractured healthy subjects from the fracture group and also removing osteoporotic non-fractured patients from the healthy individual group. These incorrectly labeled instances cannot be accurately discriminated by quantitative analysis of clinical imaging. Therefore, we propose to filter out incorrectly labeled samples from the raw data to enhance the discriminative capability and interpretability of our model.

In Chapter 4, Chapter 5 and Chapter 6, we propose two methods based on i) 3D LBP and ii) multi-view CNNs to capture bone microarchitectural information from HR-pQCT images and then feed the extracted features into the random forest classifier to discriminate previous fractures. Table 7.3 summarises the discriminative performance of LBP and multi-view CNNs for previous fracture using the HCS and GLOW datasets. We compare the discriminative performance of 3D LBP for fracture classification when the entire tibia, cortical regions and trabecular regions in HR-pQCT images are used separately as input data. The use of cortical regions segmented from CT scans yields the highest AUC performance, and the corresponding results are presented in Table 7.3. In the raw dataset of the HCS, multi-view CNNs and LBP demonstrate similar AUC results for fracture discrimination. However, in the raw dataset of the GLOW, multi-view CNNs demonstrate a higher AUC than LBP. In addition, in both the HCS and GLOW

TABLE 7.3: Comparison analysis of image feature extraction methods for fracture classification using the raw and filtered datasets.

Methods	AUC (95% CI)	
	Raw HCS dataset	Filtered HCS dataset
Multi-view CNNs	**0.75 (0.67-0.82)**	**0.90 (0.82-0.97)**
LBP	0.75 (0.67-0.81)	0.86 (0.72-0.89)
Methods	**AUC (95% CI)**	
	Raw GLOW dataset	Filtered GLOW dataset
Multi-view CNNs	**0.65 (0.58-0.70)**	**0.94 (0.89-0.97)**
LBP	0.62 (0.55-0.67)	0.85 (0.76-0.91)

HR-pQCT image data from tibial scans are used.
The highest values in each column are highlighted in bold.

datasets, when using DXA BMD to filter out incorrectly labeled samples, multi-view CNNs significantly outperform 3D LBP in distinguishing between subjects with and without previous fractures. This also highlights that multi-view CNNs can capture more comprehensive bone microarchitecture information from HR-pQCT scans compared to LBP, thereby improving fracture classification accuracy.

Our discriminative system is tested on an independent cohort to evaluate its robustness. The model is trained using the HCS dataset (source domain), and its discriminative performance is then evaluated on the GLOW dataset (target domain). Using DXA BMD filtering, our approach effectively mitigates the impact of incorrectly labeled samples and maintains robustness. When tested on the independent GLOW dataset, our model demonstrates high classification accuracy. Therefore, our discriminative system provides an opportunity to identify individuals at high risk of fracture across diverse populations.

This study has some limitations. Firstly, BMD measured by 2D DXA imaging lacks bone microarchitecture and vBMD information of participants. However, T-scores derived from BMD are used as the ground truth for sample selection. As a result, some samples in the raw data may be mistaken for incorrectly labeled samples and subsequently filtered out. Secondly, the generalisability of our model is limited. The proposed enhanced discriminative system here is designed for identifying osteoporotic fractures and may not accurately identify other types of fractures. Thirdly, both the HCS and GLOW datasets are collected from older adult populations. The robustness of our discriminative system has not yet been evaluated in cohorts from different age groups, including young and middle-aged generations. Fourthly, the sample size for both the HCS (n=167) and GLOW (n=381) datasets is small. There are differences in the optimal T-score thresholds between them. It is worth mentioning that if the model is evaluated in larger populations, the optimal thresholds may converge towards -2.5 and -1 (Watts (2004)).

7.5 Summary

A combination of HR-pQCT imaging and DXA BMD demonstrates high accuracy in identifying individuals with previous fractures. In the absence of clinical ground truth, we propose to use adaptive T-score thresholds to filter out incorrectly labeled samples from the raw data. This strategy not only significantly improves the discriminative accuracy of our approach using HR-pQCT images for fracture discrimination but also enhances the model's robustness across diverse populations. Our work also contributes insights toward understanding the relationship between BMD and bone microarchitecture assessments.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Bone microarchitecture is an important risk factor for assessing bone health and identifying individuals at high risk of fracture. HR-pQCT, providing detailed cortical and trabecular bone microarchitecture, has the potential to capture richer information beyond BMD and improve the accuracy of fracture risk prediction. In this study, we present automated fracture risk assessment tools applied to HR-pQCT images based on volumetric texture analysis and deep learning techniques. We evaluate the proposed discriminative systems here on the HCS and GLOW datasets. Numerical results demonstrate that our approaches applied to HR-pQCT images improve fracture discrimination compared to DXA-measured BMD and clinical risk factors.

In Chapter 4, considering that microarchitectural deterioration of bone tissue is associated with fracture risk, we propose a method based on 3D LBP to characterise bone microarchitecture in HR-pQCT images. Further, we present a discriminative system that combines volumetric texture analysis and machine learning classifiers to distinguish between people with and without previous fractures. Our proposed approach applied to HR-pQCT images yields a higher discriminative performance for previous fracture compared to traditional methods of clinical risk factors and femoral neck DXA.

Chapter 5 focuses on developing an automatic approach to segment cortical and trabecular regions in HR-pQCT images and investigating their relative contributions to fracture discrimination. Our method includes a deep CNN for image segmentation, enabling texture features to be extracted separately and their statistical distributions quantified for further classification. Our results demonstrate that both cortical and trabecular compartments possess important information regarding fracture risk. Notably, the cortical compartment outperforms the trabecular compartment in fracture discrimination.

An automatic method based on deep learning techniques is presented in Chapter 6 to discriminate between participants with and without previous fractures. Our CNN model employs a multi-pooling strategy to extract high-level image features from multiple views of HR-pQCT scans. These features are then integrated and fed into the random forest classifier to discriminate previous fractures. Compared to the traditional deep learning framework, our model encodes richer feature representations of CT scans and better handles high-dimensional image features, resulting in a higher accuracy based on a few of bone HR-pQCT images. This approach also outperforms traditional methods of clinical risk factors and femoral neck DXA in fracture discrimination.

We propose an adaptive threshold strategy in Chapter 7 to filter out incorrectly labeled data from the original cohorts to further improve fracture discrimination. The presence of incorrectly labeled samples in the original cohorts diminishes the classifier's capability to discriminate between fracture and non-fracture groups. Our method employs multi-view CNNs to quantitatively analyse bone microarchitecture in HR-pQCT images and generates optimal T-score thresholds to categorise individuals into non-fractured healthy and osteoporotic fracture groups. The incorporation of DXA BMD filtering significantly improves the accuracy and robustness of our discriminative system, and the evaluation in an independent cohort further supports this.

Our research facilitates the development of quantitative analysis of bone HR-pQCT images and provides an opportunity to assist clinicians in fracture prevention and medical decisions. We also expect that the proposed methods can provide motivation for bone health measurement and be generalised to other domains such as osteoarthritis and osteomyelitis diagnosis.

8.2 Future Work

The current work introduces novel computer-aided diagnosis methods for HR-pQCT imaging, which contribute to improving fracture discrimination compared to BMD and clinical risk factors. However, it also has several limitations. Looking to the future of this work, we expect to advance the research and explore larger datasets in relation to fracture risk assessment with potential ideas. Our future work not only focuses on methodological improvements but also aims to address critical applications in clinical settings. These efforts are poised to contribute to improving diagnostic accuracy, treatment planning and enhancing robustness in fracture risk assessment using medical imaging techniques.

8.2.1 Osteoporotic Fracture Prediction

HR-pQCT scans collected from both the HCS and GLOW cohorts lacked manual review by clinicians. Consequently, healthy subjects who have experienced traumatic fractures and those with bone fragility who have not experienced a fracture were erroneously classified into the fracture group and the non-fracture group respectively. These samples introduce noise and diminish the discriminative capacity of our approaches. Although T-score values provide an opportunity for sample selection, DXA imaging lacks 3D bone microarchitectural information and has limitations in labeling. Manual exclusion of such samples through clinical observations can label the data correctly and guide the model to distinguish between non-fractured healthy individuals and those with osteoporotic fractures. This not only improves the classification performance and robustness of the prediction model but also enhances its interpretability. A clinically correctly labeled dataset would also enable us to evaluate the performance of our filtering system as an interesting future work.

8.2.2 Future Fracture Prediction

Although discriminating previous fractures based on medical records provides valuable insights into patients' risk profiles, anticipating future fractures offers a broader prospect for early intervention and bone health management. Future fracture prediction not only identifies high-risk populations before bone deterioration but also assists healthcare professionals in formulating personalized intervention plans, thereby minimizing the likelihood of patients experiencing fractures.

8.2.3 Fracture Location Prediction

Our current approaches quantitatively analyse bone HR-pQCT images to discriminate previous fractures from all parts of the patients' body. However, the specific location of fragility fractures cannot be detected, and overall preventive measures are needed for those at high risk of fracture. We plan to extend our prediction model to localise fragility fractures. By predicting the specific location of potential fractures in patients, healthcare professionals can gain insights into patients' bone microarchitecture and precisely formulate prevention or treatment plans. This also contributes to optimizing the allocation of medical resources and improving healthcare efficiency.

8.2.4 Semi-Supervised Learning

Our proposed approaches here use supervised-learning methods that only select labeled data to train the risk model, while unlabeled data cannot be exploited to enhance

the discriminative performance. In practice, collecting massive labeled data is very expensive and time-consuming. Therefore, semi-supervised methods are desirable to simultaneously reduce the costs of labeled data and leverage the unlabeled data to improve the accuracy of fracture discrimination (Berthelot et al. (2019)). One common approach is self-training, where the model initially trained on labeled data iteratively uses its predictions (known as pseudo labels) on unlabeled data to expand the training set and refine its discriminative capability.

8.2.5 Subspace Representation

In few-shot 3D image classification tasks, CNNs may ignore the spatial structure and diversity of an image due to insufficient supervision signals guiding the model focusing on the region of interest. Motivated by subspace representation learning for handling high-dimensional features (Hu et al. (2021)), our plan is to employ principal component analysis (PCA) or singular value decomposition (SVD) to reduce the dimensionality of high-dimensional CNN features for further fracture classification.

8.2.6 Scale Invariant Local Binary Pattern

Our current 3D LBP model adopts a fixed scale to encode texture patterns, which depends on prior knowledge. The statistical distribution of volumetric images may be altered by the various resolutions of HR-pQCT imaging. To the best of our knowledge, the scale invariant property of the 3D LBP framework has not been proposed. We plan to develop a scale invariant 3D LBP descriptor to overcome this issue and apply it to enhance the robustness of our discriminative system.

Appendix A

Case Index of Subjects

TABLE A.1: The case index of the Hertfordshire Cohort Study.

Case Index	Subject ID	Fracture Status	Case Index	Subject ID	Fracture Status
Case 1	15014	Fracture	Case 2	15163	Fracture
Case 3	55041	Non-fracture	Case 4	55059	Non-fracture
Case 5	55256	Fracture	Case 6	55274	Non-fracture
Case 7	55289	Non-fracture	Case 8	55311	Fracture
Case 9	55422	Unknown	Case 10	55509	Fracture
Case 11	55534	Non-fracture	Case 12	55559	Fracture
Case 13	55588	Fracture	Case 14	55608	Unknown
Case 15	55631	Non-fracture	Case 16	55677	Non-fracture
Case 17	55712	Non-fracture	Case 18	55745	Fracture
Case 19	55810	Non-fracture	Case 20	55875	Non-fracture
Case 21	55909	Unknown	Case 22	55917	Fracture
Case 23	55938	Non-fracture	Case 24	55956	Unknown
Case 25	56011	Fracture	Case 26	56031	Non-fracture
Case 27	56044	Fracture	Case 28	56085	Non-fracture
Case 29	56092	Fracture	Case 30	56094	Fracture
Case 31	56133	Non-fracture	Case 32	56204	Unknown
Case 33	56210	Non-fracture	Case 34	56242	Non-fracture
Case 35	56250	Non-fracture	Case 36	56270	Unknown
Case 37	56273	Non-fracture	Case 38	56310	Fracture
Case 39	65039	Non-fracture	Case 40	65045	Non-fracture
Case 41	75003	Fracture	Case 42	75058	Fracture
Case 43	75078	Non-fracture	Case 44	75084	Unknown
Case 45	95239	Non-fracture	Case 46	95256	Non-fracture
Case 47	95277	Fracture	Case 48	95301	Fracture
Case 49	95309	Unknown	Case 50	95315	Non-fracture

Case 51	95323	Non-fracture	Case 52	95328	Non-fracture
Case 53	95352	Non-fracture	Case 54	105039	Fracture
Case 55	105043	Non-fracture	Case 56	105071	Non-fracture
Case 57	105123	Fracture	Case 58	105146	Non-fracture
Case 59	105190	Non-fracture	Case 60	105202	Non-fracture
Case 61	115050	Non-fracture	Case 62	115051	Non-fracture
Case 63	125063	Fracture	Case 64	135066	Fracture
Case 65	135070	Non-fracture	Case 66	135083	Non-fracture
Case 67	135257	Non-fracture	Case 68	135378	Fracture
Case 69	135434	Fracture	Case 70	135455	Non-fracture
Case 71	135482	Non-fracture	Case 72	135485	Fracture
Case 73	135524	Fracture	Case 74	135529	Fracture
Case 75	135687	Non-fracture	Case 76	135718	Unknown
Case 77	135784	Non-fracture	Case 78	135837	Fracture
Case 79	135851	Non-fracture	Case 80	135877	Unknown
Case 81	135892	Non-fracture	Case 82	135896	Non-fracture
Case 83	135929	Non-fracture	Case 84	135955	Non-fracture
Case 85	135960	Non-fracture	Case 86	135986	Fracture
Case 87	135993	Non-fracture	Case 88	145011	Unknown
Case 89	145045	Non-fracture	Case 90	155016	Non-fracture
Case 91	155059	Fracture	Case 92	155084	Non-fracture
Case 93	155116	Fracture	Case 94	155123	Non-fracture
Case 95	175035	Non-fracture	Case 96	175069	Fracture
Case 97	175239	Non-fracture	Case 98	175275	Non-fracture
Case 99	175297	Unknown	Case 100	175409	Non-fracture
Case 101	175442	Non-fracture	Case 102	175455	Non-fracture
Case 103	175501	Unknown	Case 104	175553	Non-fracture
Case 105	175858	Non-fracture	Case 106	175890	Non-fracture
Case 107	175911	Fracture	Case 108	176028	Non-fracture
Case 109	176029	Non-fracture	Case 110	176051	Fracture
Case 111	176091	Unknown	Case 112	176206	Unknown
Case 113	176328	Fracture	Case 114	176530	Fracture
Case 115	176544	Non-fracture	Case 116	176563	Fracture
Case 117	176631	Non-fracture	Case 118	176691	Fracture
Case 119	185007	Fracture	Case 120	185015	Non-fracture
Case 121	185018	Non-fracture	Case 122	195016	Non-fracture
Case 123	225014	Non-fracture	Case 124	225019	Fracture
Case 125	225115	Fracture	Case 126	225162	Fracture
Case 127	235022	Fracture	Case 128	235044	Fracture
Case 129	245230	Non-fracture	Case 130	245252	Non-fracture
Case 131	245267	Non-fracture	Case 132	245314	Unknown

Case 133	255022	Non-fracture	Case 134	255049	Non-fracture
Case 135	255058	Non-fracture	Case 136	265045	Non-fracture
Case 137	295011	Non-fracture	Case 138	295072	Fracture
Case 139	315019	Non-fracture	Case 140	315032	Non-fracture
Case 141	325017	Fracture	Case 142	335029	Non-fracture
Case 143	335034	Non-fracture	Case 144	355008	Fracture
Case 145	355022	Non-fracture	Case 146	355081	Non-fracture
Case 147	365008	Non-fracture	Case 148	365018	Non-fracture
Case 149	415040	Non-fracture	Case 150	445048	Non-fracture
Case 151	475013	Non-fracture	Case 152	515017	Non-fracture
Case 153	525357	Non-fracture	Case 154	565020	Fracture
Case 155	565079	Non-fracture	Case 156	565080	Fracture
Case 157	565209	Non-fracture	Case 158	575033	Non-fracture
Case 159	595564	Non-fracture	Case 160	596737	Non-fracture
Case 161	597269	Unknown	Case 162	597350	Non-fracture
Case 163	598134	Unknown	Case 164	635009	Fracture
Case 165	635031	Non-fracture	Case 166	635097	Fracture
Case 167	655209	Non-fracture	Case 168	665380	Fracture
Case 169	665491	Non-fracture	Case 170	666215	Non-fracture
Case 171	705075	Non-fracture	Case 172	705130	Fracture
Case 173	735674	Non-fracture	Case 174	785040	Non-fracture
Case 175	785091	Unknown	Case 176	785097	Fracture
Case 177	785122	Non-fracture	Case 178	785132	Non-fracture
Case 179	785134	Non-fracture	Case 180	795028	Non-fracture
Case 181	855021	Fracture	Case 182	855027	Unknown
Case 183	855038	Non-fracture	Case 184	855046	Fracture
Case 185	855072	Non-fracture	Case 186	855110	Fracture
Case 187	865008	Non-fracture	Case 188	865011	Non-fracture
Case 189	865055	Fracture	Case 190	865125	Unknown
Case 191	865198	Unknown	Case 192	865261	Fracture
Case 193	865290	Non-fracture	Case 194	865308	Non-fracture
Case 195	865337	Non-fracture	Case 196	865355	Non-fracture
Case 197	865359	Non-fracture	Case 198	865379	Non-fracture
Case 199	865427	Non-fracture	Case 200	865430	Non-fracture
Case 201	865433	Non-fracture	Case 202	865512	Fracture
Case 203	864547	Fracture	Case 204	86550	Non-fracture
Case 205	865608	Fracture	Case 206	865648	Non-fracture
Case 207	865668	Fracture	Case 208	865671	Non-fracture
Case 209	865704	Non-fracture	Case 210	865784	Unknown
Case 211	865885	Non-fracture	Case 212	865983	Non-fracture
Case 213	866028	Non-fracture	Case 214	866042	Fracture

Case 215	866058	Non-fracture	Case 216	866096	Fracture
Case 217	866109	Non-fracture	Case 218	866185	Non-fracture
Case 219	866235	Fracture	Case 220	866260	Non-fracture
Case 221	866325	Non-fracture	Case 222	866400	Fracture
Case 223	866405	Non-fracture	Case 224	866416	Fracture
Case 225	866448	Non-fracture	Case 226	866457	Non-fracture
Case 227	866472	Non-fracture	Case 228	866587	Non-fracture
Case 229	866615	Non-fracture	Case 230	866617	Non-fracture
Case 231	866685	Non-fracture	Case 232	866693	Non-fracture
Case 233	875068	Fracture	Case 234	885020	Non-fracture
Case 235	895006	Non-fracture	Case 236	898025	Non-fracture
Case 237	905024	Non-fracture	Case 238	905035	Fracture
Case 239	905023	Non-fracture	Case 240	925033	Non-fracture
Case 241	925044	Fracture	Case 242	925055	Non-fracture
Case 243	935192	Fracture	Case 244	935276	Non-fracture
Case 245	935377	Fracture	Case 246	935393	Non-fracture
Case 247	935487	Fracture	Case 248	935489	Non-fracture
Case 249	935492	Non-fracture	Case 250	935612	Non-fracture
Case 251	935659	Non-fracture	Case 252	935715	Non-fracture
Case 253	935723	Non-fracture	Case 254	735749	Non-fracture
Case 255	935755	Non-fracture	Case 256	95765	Non-fracture
Case 257	935794	Fracture	Case 258	935819	Non-fracture
Case 259	935825	Fracture	Case 260	935886	Fracture
Case 261	935923	Non-fracture	Case 262	935934	Non-fracture
Case 263	945003	Non-fracture	Case 264	945075	Non-fracture
Case 265	945123	Non-fracture	Case 266	945128	Non-fracture
Case 267	945130	Non-fracture	Case 268	955106	Non-fracture
Case 269	965016	Fracture	Case 270	965037	Non-fracture
Case 271	965081	Non-fracture	Case 272	965174	Non-fracture
Case 273	985014	Non-fracture	Case 274	1005012	Non-fracture
Case 275	1005025	Non-fracture	Case 276	1005041	Non-fracture
Case 277	1005060	Fracture	Case 278	1005094	Non-fracture
Case 279	1005151	Fracture	Case 280	1005159	Non-fracture
Case 281	1005205	Non-fracture	Case 282	1005222	Non-fracture
Case 283	1006091	Fracture	Case 284	855046	Unknown
Case 285	1006184	Fracture	Case 286	1006408	Unknown
Case 287	1075646	Fracture	Case 288	1125309	Non-fracture
Case 289	1206896	Non-fracture	Case 290	1207395	Non-fracture
Case 291	1215522	Fracture	Case 292	1235629	Unknown
Case 293	1335091	Fracture	Case 294	1335105	Non-fracture
Case 295	1345422	Non-fracture	Case 296	1345508	Non-fracture

Case 297	1345781	Non-fracture	Case 298	1347041	Fracture
Case 299	1347102	Non-fracture	Case 300	1347474	Non-fracture
Case 301	1348431	Non-fracture	Case 302	1348578	Non-fracture
Case 303	1348675	Non-fracture	Case 304	1355061	Non-fracture
Case 305	1355935	Non-fracture	Case 306	1356198	Non-fracture
Case 307	1356255	Non-fracture	Case 308	1356471	Non-fracture
Case 309	1356511	Fracture	Case 310	1356514	Non-fracture
Case 311	1356649	Non-fracture	Case 312	1356863	Non-fracture
Case 313	1357081	Non-fracture	Case 314	1357262	Non-fracture
Case 315	1347426	Non-fracture	Case 316	1357702	Non-fracture
Case 317	1357995	Non-fracture	Case 318	1358009	Fracture
Case 319	1358100	Non-fracture	Case 320	1358357	Fracture
Case 321	1358408	Unknown	Case 322	1406063	Fracture
Case 323	1406069	Non-fracture	Case 324	1406089	Unknown
Case 325	1406124	Non-fracture	Case 326	1406168	Unknown
Case 327	1406238	Non-fracture	Case 328	1406319	Non-fracture
Case 329	1406337	Non-fracture	Case 330	1406353	Non-fracture
Case 331	1406362	Fracture	Case 332	1406368	Non-fracture
Case 333	1406493	Non-fracture	Case 334	1406553	Unknown
Case 335	1406647	Non-fracture	Case 336	1406699	Fracture
Case 337	1406930	Fracture	Case 338	1406931	Non-fracture
Case 339	1406943	Non-fracture	Case 340	1407019	Unknown
Case 341	1407069	Non-fracture	Case 342	1407110	Fracture
Case 343	1407119	Non-fracture	Case 344	1407149	Non-fracture
Case 345	1407150	Fracture	Case 346	1407159	Non-fracture
Case 347	1407187	Non-fracture	Case 348	1407208	Non-fracture
Case 349	1407389	Fracture	Case 350	1407449	Non-fracture
Case 351	1407535	Non-fracture	Case 352	1407552	Fracture
Case 353	1407561	Fracture	Case 354	1407605	Fracture
Case 355	1407657	Non-fracture	Case 356	1407659	Non-fracture
Case 357	1407685	Non-fracture	Case 358	1407727	Non-fracture
Case 359	1407759	Fracture	Case 360	1407771	Non-fracture
Case 361	1407827	Non-fracture	Case 362	1407849	Non-fracture
Case 363	1407862	Non-fracture	Case 364	1407877	Non-fracture
Case 365	1407902	Non-fracture	Case 366	1407925	Fracture
Case 367	1407929	Non-fracture	Case 368	1408063	Non-fracture
Case 369	1408086	Non-fracture	Case 370	1408095	Non-fracture
Case 371	1408124	Non-fracture	Case 372	1408215	Non-fracture
Case 373	1408217	Fracture	Case 374	1408245	Non-fracture
Case 375	1408256	Fracture	Case 376	1408409	Non-fracture

Fracture: the participant had a vertebral fracture or a self-reported fracture.
Non-fracture: no fractures occurred in the participant.
Unknown: the fracture status of the participant was not collected.

TABLE A.2: The case index of the Global Longitudinal Study of Osteoporosis in Women.

Case Index	Subject ID	Fracture Status	Case Index	Subject ID	Fracture Status
Case 1	10001	Non-fracture	Case 2	10003	Non-fracture
Case 3	10007	Non-fracture	Case 4	10020	Non-fracture
Case 5	10025	Fracture	Case 6	10040	Non-fracture
Case 7	10049	Non-fracture	Case 8	10050	Non-fracture
Case 9	10061	Non-fracture	Case 10	10067	Non-fracture
Case 11	10075	Non-fracture	Case 12	10077	Fracture
Case 13	10080	Non-fracture	Case 14	10081	Fracture
Case 15	10083	Non-fracture	Case 16	10088	Fracture
Case 17	10092	Non-fracture	Case 18	10095	Non-fracture
Case 19	10096	Non-fracture	Case 20	10101	Non-fracture
Case 21	10108	Non-fracture	Case 22	10109	Non-fracture
Case 23	10115	Non-fracture	Case 24	10124	Non-fracture
Case 25	10134	Unknown	Case 26	10138	Non-fracture
Case 27	10139	Fracture	Case 28	10141	Non-fracture
Case 29	10156	Fracture	Case 30	10158	Non-fracture
Case 31	10159	Fracture	Case 32	10160	Non-fracture
Case 33	10166	Fracture	Case 34	10168	Non-fracture
Case 35	10170	Fracture	Case 36	10171	Non-fracture
Case 37	10172	Non-fracture	Case 38	10181	Non-fracture
Case 39	10188	Non-fracture	Case 40	10196	Non-fracture
Case 41	10197	Non-fracture	Case 42	10200	Non-fracture
Case 43	10210	Fracture	Case 44	10214	Non-fracture
Case 45	10218	Non-fracture	Case 46	10222	Fracture
Case 47	10225	Fracture	Case 48	10226	Non-fracture
Case 49	10239	Non-fracture	Case 50	10240	Non-fracture
Case 51	10248	Non-fracture	Case 52	10260	Non-fracture
Case 53	10289	Fracture	Case 54	10291	Fracture
Case 55	10309	Non-fracture	Case 56	10321	Non-fracture
Case 57	10326	Non-fracture	Case 58	10344	Unknown
Case 59	10345	Fracture	Case 60	10347	Fracture
Case 61	10355	Non-fracture	Case 62	10356	Fracture
Case 63	10369	Unknown	Case 64	10392	Fracture
Case 65	10394	Non-fracture	Case 66	10396	Non-fracture
Case 67	10398	Unknown	Case 68	10401	Non-fracture

Case 69	10411	Fracture	Case 70	10424	Fracture
Case 71	10433	Non-fracture	Case 72	10442	Non-fracture
Case 73	10446	Non-fracture	Case 74	10456	Fracture
Case 75	10483	Non-fracture	Case 76	10484	Non-fracture
Case 77	10502	Non-fracture	Case 78	10506	Non-fracture
Case 79	10762	Non-fracture	Case 80	10767	Fracture
Case 81	10768	Non-fracture	Case 82	10774	Unknown
Case 83	10776	Fracture	Case 84	10777	Non-fracture
Case 85	10778	Non-fracture	Case 86	10779	Fracture
Case 87	10780	Non-fracture	Case 88	10781	Non-fracture
Case 89	10786	Non-fracture	Case 90	10799	Fracture
Case 91	10803	Non-fracture	Case 92	10807	Non-fracture
Case 93	10808	Non-fracture	Case 94	10808	Fracture
Case 95	10812	Non-fracture	Case 96	10815	Fracture
Case 97	10818	Non-fracture	Case 98	10820	Fracture
Case 99	10822	Non-fracture	Case 100	10828	Non-fracture
Case 101	10829	Non-fracture	Case 102	10830	Fracture
Case 103	10839	Non-fracture	Case 104	10841	Unknown
Case 105	10846	Non-fracture	Case 106	10851	Fracture
Case 107	10853	Non-fracture	Case 108	10863	Unknown
Case 109	10866	Non-fracture	Case 110	10868	Non-fracture
Case 111	10872	Non-fracture	Case 112	10879	Non-fracture
Case 113	10880	Fracture	Case 114	10889	Non-fracture
Case 115	10900	Fracture	Case 116	10901	Fracture
Case 117	10902	Non-fracture	Case 118	10904	Non-fracture
Case 119	10905	Non-fracture	Case 120	10906	Non-fracture
Case 121	10912	Non-fracture	Case 122	10919	Non-fracture
Case 123	10921	Non-fracture	Case 124	10922	Non-fracture
Case 125	10932	Non-fracture	Case 126	10934	Non-fracture
Case 127	10938	Non-fracture	Case 128	10944	Non-fracture
Case 129	10945	Non-fracture	Case 130	10949	Non-fracture
Case 131	10950	Fracture	Case 132	10952	Fracture
Case 133	10955	Fracture	Case 134	10970	Fracture
Case 135	10972	Non-fracture	Case 136	10977	Non-fracture
Case 137	10978	Unknown	Case 138	10981	Non-fracture
Case 139	10985	Non-fracture	Case 140	10989	Non-fracture
Case 141	10992	Non-fracture	Case 142	10996	Fracture
Case 143	10997	Non-fracture	Case 144	11002	Non-fracture
Case 145	11004	Fracture	Case 146	11010	Fracture
Case 147	11013	Fracture	Case 148	11014	Fracture
Case 149	11015	Non-fracture	Case 150	11019	Non-fracture

Case 151	11020	Non-fracture	Case 152	11021	Non-fracture
Case 153	11023	Fracture	Case 154	11027	Non-fracture
Case 155	11036	Fracture	Case 156	11038	Non-fracture
Case 157	11039	Non-fracture	Case 158	11042	Non-fracture
Case 159	11059	Non-fracture	Case 160	11062	Non-fracture
Case 161	11064	Non-fracture	Case 162	11067	Non-fracture
Case 163	11072	Fracture	Case 164	11073	Non-fracture
Case 165	11076	Non-fracture	Case 166	11077	Non-fracture
Case 167	11080	Fracture	Case 168	11083	Non-fracture
Case 169	11084	Non-fracture	Case 170	11085	Fracture
Case 171	11088	Non-fracture	Case 172	11090	Non-fracture
Case 173	11091	Fracture	Case 174	11092	Non-fracture
Case 175	11095	Non-fracture	Case 176	11096	Non-fracture
Case 177	11099	Unknown	Case 178	11111	Unknown
Case 179	11115	Non-fracture	Case 180	11118	Non-fracture
Case 181	11121	Non-fracture	Case 182	11126	Non-fracture
Case 183	11134	Non-fracture	Case 184	11135	Non-fracture
Case 185	11141	Non-fracture	Case 186	11145	Fracture
Case 187	11147	Non-fracture	Case 188	11148	Non-fracture
Case 189	11149	Non-fracture	Case 190	11153	Non-fracture
Case 191	11155	Fracture	Case 192	11156	Non-fracture
Case 193	11159	Non-fracture	Case 194	11164	Non-fracture
Case 195	11170	Unknown	Case 196	11176	Non-fracture
Case 197	11178	Non-fracture	Case 198	11180	Non-fracture
Case 199	11185	Fracture	Case 200	11187	Non-fracture
Case 201	11190	Unknown	Case 202	11193	Non-fracture
Case 203	11195	Non-fracture	Case 204	11198	Non-fracture
Case 205	11199	Non-fracture	Case 206	11210	Non-fracture
Case 207	11217	Non-fracture	Case 208	11218	Non-fracture
Case 209	11223	Non-fracture	Case 210	11225	Non-fracture
Case 211	11226	Fracture	Case 212	11236	Unknown
Case 213	11242	Non-fracture	Case 214	11246	Fracture
Case 215	11248	Non-fracture	Case 216	11259	Non-fracture
Case 217	11261	Non-fracture	Case 218	11264	Non-fracture
Case 219	11268	Non-fracture	Case 220	11272	Non-fracture
Case 221	11280	Non-fracture	Case 222	11285	Fracture
Case 223	11286	Non-fracture	Case 224	11290	Non-fracture
Case 225	11291	Non-fracture	Case 226	11318	Non-fracture
Case 227	11322	Fracture	Case 228	11328	Fracture
Case 229	11334	Fracture	Case 230	11646	Non-fracture
Case 231	11647	Non-fracture	Case 232	11650	Non-fracture

Case 233	11653	Fracture	Case 234	11654	Non-fracture
Case 235	11655	Non-fracture	Case 236	11656	Non-fracture
Case 237	11658	Non-fracture	Case 238	11660	Non-fracture
Case 239	11662	Fracture	Case 240	11664	Non-fracture
Case 241	11667	Fracture	Case 242	11669	Non-fracture
Case 243	11672	Fracture	Case 244	11676	Non-fracture
Case 245	11677	Fracture	Case 246	11681	Non-fracture
Case 247	11683	Non-fracture	Case 248	11687	Non-fracture
Case 249	11692	Non-fracture	Case 250	11699	Non-fracture
Case 251	11702	Non-fracture	Case 252	11705	Unknown
Case 253	11708	Non-fracture	Case 254	11709	Non-fracture
Case 255	11712	Non-fracture	Case 256	11713	Unknown
Case 257	11716	Fracture	Case 258	11732	Non-fracture
Case 259	11734	Non-fracture	Case 260	11737	Non-fracture
Case 261	11741	Non-fracture	Case 262	11744	Non-fracture
Case 263	11753	Fracture	Case 264	11756	Fracture
Case 265	11760	Non-fracture	Case 266	11762	Fracture
Case 267	11773	Non-fracture	Case 268	11777	Non-fracture
Case 269	11782	Non-fracture	Case 270	11785	Non-fracture
Case 271	11786	Non-fracture	Case 272	11787	Non-fracture
Case 273	11792	Non-fracture	Case 274	11805	Non-fracture
Case 275	11809	Non-fracture	Case 276	11810	Non-fracture
Case 277	11813	Non-fracture	Case 278	11822	Non-fracture
Case 279	11823	Non-fracture	Case 280	11826	Non-fracture
Case 281	11833	Non-fracture	Case 282	11836	Non-fracture
Case 283	11838	Unknown	Case 284	11841	Non-fracture
Case 285	11846	Non-fracture	Case 286	11848	Non-fracture
Case 287	11851	Non-fracture	Case 288	11853	Non-fracture
Case 289	11860	Non-fracture	Case 290	11868	Non-fracture
Case 291	11869	Unknown	Case 292	11870	Non-fracture
Case 293	11872	Unknown	Case 294	11875	Non-fracture
Case 295	11876	Non-fracture	Case 296	11880	Fracture
Case 297	11883	Non-fracture	Case 298	12437	Non-fracture
Case 299	12512	Non-fracture	Case 300	12515	Non-fracture
Case 301	12537	Non-fracture	Case 302	12539	Non-fracture
Case 303	12555	Non-fracture	Case 304	12571	Non-fracture
Case 305	12577	Non-fracture	Case 306	12580	Fracture
Case 307	12588	Non-fracture	Case 308	12597	Unknown
Case 309	12598	Non-fracture	Case 310	12607	Non-fracture
Case 311	12614	Non-fracture	Case 312	12627	Non-fracture
Case 313	12629	Non-fracture	Case 314	12637	Non-fracture

Case 315	12640	Non-fracture	Case 316	12646	Unknown
Case 317	12655	Fracture	Case 318	12656	Non-fracture
Case 319	12657	Unknown	Case 320	12661	Fracture
Case 321	12662	Non-fracture	Case 322	12663	Non-fracture
Case 323	12670	Non-fracture	Case 324	12678	Non-fracture
Case 325	12679	Non-fracture	Case 326	12685	Non-fracture
Case 327	12693	Non-fracture	Case 328	12707	Fracture
Case 329	12711	Unknown	Case 330	12719	Non-fracture
Case 331	12730	Non-fracture	Case 332	12742	Non-fracture
Case 333	12742	Fracture	Case 334	12746	Non-fracture
Case 335	12755	Non-fracture	Case 336	12757	Non-fracture
Case 337	12767	Non-fracture	Case 338	12774	Non-fracture
Case 339	12777	Fracture	Case 340	12781	Fracture
Case 341	12785	Non-fracture	Case 342	12788	Non-fracture
Case 343	12795	Fracture	Case 344	12802	Fracture
Case 345	12809	Non-fracture	Case 346	12820	Non-fracture
Case 347	12834	Fracture	Case 348	12845	Non-fracture
Case 349	12850	Non-fracture	Case 350	12859	Non-fracture
Case 351	12869	Fracture	Case 352	12883	Fracture
Case 353	12888	Fracture	Case 354	12889	Non-fracture
Case 355	12892	Non-fracture	Case 356	12904	Non-fracture
Case 357	12908	Fracture	Case 358	12913	Non-fracture
Case 359	12915	Non-fracture	Case 360	12922	Fracture
Case 361	12925	Non-fracture	Case 362	12936	Non-fracture
Case 363	12938	Non-fracture	Case 364	12941	Non-fracture
Case 365	12942	Non-fracture	Case 366	12947	Unknown
Case 367	12948	Non-fracture	Case 368	12959	Non-fracture
Case 369	12964	Non-fracture	Case 370	12977	Non-fracture
Case 371	12991	Non-fracture	Case 372	13004	Non-fracture
Case 373	13019	Non-fracture	Case 374	13023	Non-fracture
Case 375	13028	Fracture	Case 376	13036	Non-fracture
Case 377	13045	Unknown	Case 378	13047	Non-fracture
Case 379	13055	Fracture	Case 380	13077	Non-fracture
Case 381	13092	Fracture	Case 382	13097	Non-fracture
Case 383	13109	Unknown	Case 384	13112	Non-fracture
Case 385	13120	Non-fracture	Case 386	13126	Non-fracture
Case 387	13127	Non-fracture	Case 388	13137	Non-fracture
Case 389	13138	Non-fracture	Case 390	13140	Unknown
Case 391	13141	Non-fracture	Case 392	13146	Non-fracture
Case 393	13152	Non-fracture	Case 394	13153	Non-fracture
Case 395	13161	Non-fracture	Case 396	13172	Non-fracture

Case 397	13174	Non-fracture	Case 398	13176	Fracture
Case 399	13177	Non-fracture	Case 400	13178	Non-fracture
Case 401	13182	Non-fracture	Case 402	13183	Unknown
Case 403	13197	Fracture	Case 404	13201	Fracture
Case 405	13211	Unknown	Case 406	13215	Non-fracture
Case 407	13225	Non-fracture	Case 408	13227	Fracture
Case 409	13234	Non-fracture	Case 410	13240	Non-fracture
Case 411	13244	Unknown	Case 412	13250	Non-fracture
Case 413	13253	Non-fracture	Case 414	13254	Fracture
Case 415	13259	Unknown	Case 416	13270	Non-fracture
Case 417	13297	Fracture	Case 418	13303	Unknown
Case 419	13311	Non-fracture	Case 420	13328	Non-fracture
Case 421	13334	Fracture	Case 422	13335	Non-fracture
Case 423	13392	Fracture	Case 424	13397	Non-fracture
Case 425	13402	Non-fracture	Case 426	13404	Fracture
Case 427	13410	Non-fracture	Case 428	13423	Fracture
Case 429	13431	Non-fracture	Case 430	13438	Non-fracture
Case 431	13442	Non-fracture	Case 432	13450	Non-fracture
Case 433	13556	Non-fracture	Case 434	13637	Non-fracture
Case 435	13642	Non-fracture	Case 436	13762	Non-fracture
Case 437	13886	Non-fracture	Case 438	14034	Non-fracture
Case 439	14097	Non-fracture	Case 440	14100	Non-fracture
Case 441	14156	Non-fracture	Case 442	14166	Non-fracture
Case 443	14291	Non-fracture	Case 444	14559	Unknown
Case 445	14570	Fracture	Case 446	14583	Fracture
Case 447	14601	Fracture	Case 448	14615	Non-fracture
Case 449	14633	Unknown	Case 450	14636	Non-fracture
Case 451	14637	Unknown	Case 452	14646	Non-fracture
Case 453	14653	Non-fracture	Case 454	14660	Non-fracture
Case 455	14678	Non-fracture	Case 456	14680	Non-fracture
Case 457	14684	Non-fracture	Case 458	14690	Non-fracture
Case 459	14693	Fracture	Case 460	14695	Non-fracture
Case 461	14703	Fracture	Case 462	14705	Non-fracture
Case 463	14709	Non-fracture	Case 464	14713	Fracture
Case 465	14717	Non-fracture	Case 466	14724	Non-fracture
Case 467	14726	Non-fracture	Case 468	14731	Unknown
Case 469	14733	Non-fracture	Case 470	14735	Non-fracture
Case 471	14741	Non-fracture	Case 472	14742	Non-fracture
Case 473	14743	Non-fracture	Case 474	14757	Fracture
Case 475	14760	Non-fracture	Case 476	14764	Non-fracture
Case 477	14766	Fracture	Case 478	14771	Non-fracture

Case 479	14792	Non-fracture	Case 480	14793	Non-fracture
Case 481	14794	Non-fracture	Case 482	14797	Non-fracture
Case 483	14802	Fracture	Case 484	14803	Non-fracture
Case 485	14811	Non-fracture	Case 486	14822	Fracture
Case 487	14831	Non-fracture	Case 488	14832	Non-fracture
Case 489	14833	Unknown	Case 490	14834	Non-fracture
Case 491	14836	Fracture	Case 492	14840	Non-fracture
Case 493	14852	Non-fracture	Case 494	14853	Non-fracture
Case 495	14857	Fracture	Case 496	14858	Non-fracture
Case 497	14862	Unknown	Case 498	14877	Non-fracture
Case 499	14880	Non-fracture	Case 500	14887	Non-fracture
Case 501	14890	Non-fracture			

Fracture: the participant had a self-reported fracture..
Non-fracture: no fractures occurred in the participant.
Unknown: the fracture status of the participant was not collected.

Appendix B

Comparison Studies

TABLE B.1: Comparison of various machine learning classifiers for fracture discrimination.

Classifiers	AUC	Sensitivity	Specificity
Random forest	**0.73**	0.46	0.81
Gaussian Naive Bayes	0.69	**0.52**	0.74
Multilayer perceptron	0.72	0.36	**0.89**
Linear regression	0.70	0.43	0.78

Input data: tibial HR-pQCT scans.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

TABLE B.2: Discriminative performance of the random forest classifier for previous fracture according to thresholds.

Thresholds	AUC	Sensitivity	Specificity
0.38	**0.73**	**0.50**	0.75
0.40	**0.73**	0.46	0.81
0.42	**0.73**	0.40	0.84
0.44	**0.73**	0.30	0.89
0.50	**0.73**	0.12	**0.96**

Input data: tibial HR-pQCT scans.
The Youden's Index is used to determine the optimal threshold.
The highest values in each column are highlighted in bold.

Appendix C

Robust Completed Local Binary Pattern Descriptor for Fracture Discrimination

3D LBP shows significant performance in many domains such as solid textures analysis, face recognition and tumor detection (Citraro et al. (2017)). Due to its outstanding advantages of easy implementation and low computational complexity, LBP has received increasing attention in the past decades, and various variants for LBP have been put forward to improve its model capabilities (Zhao and Pietikainen (2007), Fehr and Burkhardt (2008)). Although 3D LBP has high discriminative power, its sensitivity to image noise remains a challenge.

Encoding high discriminant descriptors is essential to address specific tasks, but feature extraction may be affected by image perturbations such as noise (Maani et al. (2013)). Over the decades, some strategies have been proposed to improve the noise tolerance of LBP methods. For example, Fathi and Naghsh-Nilchi (2012) presented a noise-tolerant LBP descriptor that used a circular majority voting filter and labeling scheme to improve model robustness and discrimination ability. Based on the uniform LBP, Chen et al. (2013) proposed a robust texture descriptor that changed the coding of the three-bit substring to make the model more robust against noise. Nonetheless, those approaches are only limited to 2D texture analysis.

Here, we propose an efficient strategy to improve the noise tolerance of our 3D LBP descriptor, as presented in Chapter 4 and Chapter 5. For each voxel in the 3D image, we employ a 3D weighted average filter that replaces its original voxel value with the average local voxel value based on weights to reduce the influence of noise. In addition, inspired by the CLBP variant (as described in Section 2.4.2 of Chapter 2), we combine

the texture features of sign, magnitude and centre pixel operators to enhance the robustness and classification performance of 3D LBP (Liu et al. (2012)). Our experimental results demonstrate that the proposed robust completed local binary pattern (RCLBP) method can be applied to characterise bone microarchitecture in HR-pQCT images for fracture discrimination and to enhance the noise tolerance of our discriminative system (see Figure 5.9 of Chapter 5).

The details of 3D LBP are described in Section 4.3 of Chapter 4. p represents the number of sampled points around the centre voxel. r is the Euclidean distance between the centre voxel v_c and its sampled neighboring points v_i in the cube. In order to reduce the influence of noise present in images, we consider a weighted average function $\phi(v_c|w,q)$ for each voxel v_c in the image to make the model more robust against noise (Zhao et al. (2013)) as presented in Equation (C.1):

$$\phi(v_c|w,q) = \frac{\sum_{i=0}^{q-1} v_i + v_c \times w}{w+q} \tag{C.1}$$

where, q denotes the number of neighboring points around the centre voxel, and w represents the weight of the centre voxel in the neighborhood. Such voxel contains more valuable information than its neighbors. By calculating the average grayscale value based on weights, instead of the original grayscale value of the central voxel, we can reduce the influence of noise.

We use individual central voxel with regional representation in LBP value calculations and consider three robust texture operators.

The robust centre pixel operator (RCLBP_C) encodes the contrast information between the centre voxel and the global threshold. It is defined as follows:

$$RCLBP\_C_{p,r} = f\left(\phi(v_c|w,q) - \propto\right) \tag{C.2}$$

and

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else } x < 0 \end{cases} \tag{C.3}$$

where, $\propto$ represents the global threshold calculated by the mean voxel of the entire image processed using the weighted average function $\phi(v_c|w,q)$.

The robust sign operator RCLBP_S of our method is encoded as presented in Equation (C.4). The introduced threshold ξ instead of the grayscale value of the centre voxel

to encode texture patterns can enhance model robustness against noise.

$$RCLBP\_S_{p,r} = \sum_{i=0}^{p-1} f(\phi(v_i\,|w,q) - \xi) \times 2^{i-1} \tag{C.4}$$

and

$$\xi = \frac{\sum_{i=0}^{p-1} \phi(v_i|w,q) + \phi(v_c|w,q) \times m}{m+p} \tag{C.5}$$

where, m represents the weight of the centre voxel in the sampling space.

Similarly, the robust magnitude operator RCLBP_M uses the global threshold τ which is set as the mean magnitude value of the entire image to encode the local structure, i.e.:

$$RCLBP\_M_{p,r} = \sum_{i=0}^{p-1} f\left(\ell(v_c, v_i) - \tau\right) \times 2^{i-1} \tag{C.6}$$

and

$$\ell(v_c, v_i) = |\phi(v_i|w,q) - \phi(v_c|w,q)| \tag{C.7}$$

After the LBP value of each voxel $v_{i,j,k}$ in the image is calculated, we compute a histogram $H(b)$, $b \in [0, B]$ to represent the texture descriptor.

$$H(b) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \psi(LBP_{i,j,k}, b) \tag{C.8}$$

and

$$\psi(\alpha, \beta) = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{else} \end{cases} \tag{C.9}$$

where, B denotes the maximum LBP value in a texture image.

We calculate the histograms of the RCLBP_C, RCLBP_S and RCLBP_M descriptors respectively, and then concatenate these three histograms to construct a feature vector of a texture, as follows:

$$H = H_{RCLBP\_C} \oplus H_{RCLBP\_S} \oplus H_{RCLBP\_M} \tag{C.10}$$

where, $\oplus$ represents the concatenation operation.

Appendix D

Domain Adaptation

Domain adaptation has become a pivotal approach in the machine learning field to address the challenges posed by variations in data distributions (Guan and Liu (2021)). It aims to enhance the adaptability and generalisation capabilities of models across diverse domains. When the target domain varies from the source domain, it typically leads to a potential degradation in model performance. Domain adaptation has been proposed to solve this problem and improve model performance in the target domain.

As illustrated in Figure 7.2 of Chapter 7, a discernible difference exists in the T-score distributions of the HCS and GLOW datasets. This indicates that participants in these two cohorts had different levels of bone loss or bone deterioration. As a result, there would be a variance in the feature distributions of HR-pQCT images between the HCS

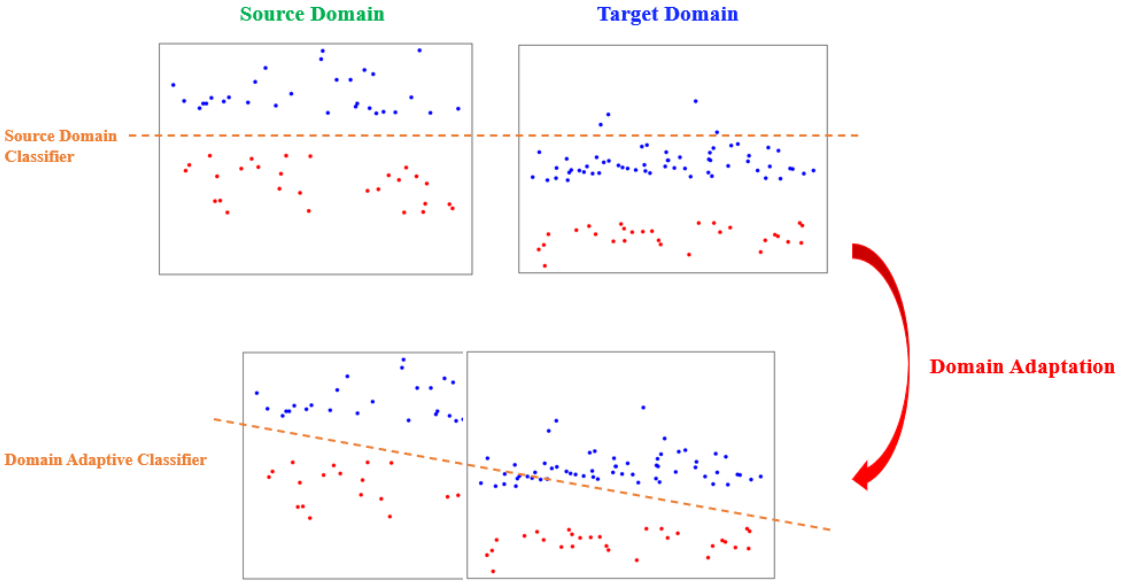

FIGURE D.1: Schematic diagram of our model for domain adaption, where the source domain and target domain exhibit different data distributions. The domain adaptive classifier is more effective than the source domain classifier in the target domain.

and GLOW, which may diminish the discriminative accuracy of our approach (see Figure D.1). To address this issue, we propose to use domain adaptation to enhance the generalisation performance of our model. Specifically, we initially train our discriminative system for previous fracture on the HCS dataset (source domain). Subsequently, we adapt our model to the target domain by retraining it on the validation set of the GLOW dataset. Finally, we evaluate the discriminative performance of our approach on the testing set of the GLOW.

References

Solmaz Abbasi and Farshad Tajeripour. Detection of brain tumor in 3d mri images using local binary patterns and histogram orientation gradient. *Neurocomputing*, 219: 526–535, 2017.

Saleh Albelwi and Ausif Mahmood. A framework for designing the architectures of deep convolutional neural networks. *Entropy*, 19(6):242, 2017.

Manar Aljabri, Manal AlAmir, Manal AlGhamdi, Mohamed Abdel-Mottaleb, and Fernando Collado-Mesa. Towards a better understanding of annotation tools for medical imaging: a survey. *Multimedia tools and applications*, 81(18):25877–25911, 2022.

Anu Shaju Areeckal, Michel Kocher, et al. Current and emerging diagnostic imaging-based techniques for assessment of osteoporosis and fracture risk. *IEEE reviews in biomedical engineering*, 12:254–268, 2018.

Mohammed Bader-El-Den, Eleman Teitei, and Todd Perry. Biased random forest for dealing with the class imbalance problem. *IEEE transactions on neural networks and learning systems*, 30(7):2163–2172, 2018.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):1–10, 2019.

Charles R Bardeen. Studies of the development of the human skeleton. *Am J Anat*, 4: 265–302, 1905.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mark J Bolland, Amanda TY Siu, Barbara H Mason, Anne M Horne, Ruth W Ames, Andrew B Grey, Greg D Gamble, and Ian R Reid. Evaluation of the frax and garvan fracture risk calculators in older women. *Journal of Bone and Mineral Research*, 26(2): 420–427, 2011.

Stephanie Boutroy, Mary L Bouxsein, Francoise Munoz, and Pierre D Delmas. In vivo assessment of trabecular bone microarchitecture by high-resolution peripheral quantitative computed tomography. *The Journal of Clinical Endocrinology & Metabolism*, 90 (12):6508–6515, 2005.

Helen R Buie, Graeme M Campbell, R Joshua Klinck, Joshua A MacNeil, and Steven K Boyd. Automatic segmentation of cortical and trabecular compartments based on a dual threshold technique for in vivo micro-ct bone analysis. *Bone*, 41(4):505–515, 2007.

Andrew J Burghardt, Helen R Buie, Andres Laib, Sharmila Majumdar, and Steven K Boyd. Reproducibility of direct quantitative measures of cortical bone microarchitecture of the distal radius and tibia by hr-pqct. *Bone*, 47(3):519–528, 2010.

Sylvie Isabelle Cappelle, Michel Moreau, Rafik Karmali, Laura Iconaru, Felicia Baleanu, V Kinnard, Marianne Paesmans, Serge Rozenberg, Michel Rubinstein, Murielle Surquin, et al. Discriminating value of hr-pqct for fractures in women with similar frax scores: A substudy of the frisbee cohort. *Bone*, 143:115613, 2021.

Jie Chen, Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Rlbp: Robust local binary pattern. In *BMVC*, 2013.

C Christodoulou and C Cooper. What is osteoporosis? *Postgraduate medical journal*, 79 (929):133–138, 2003.

Leonardo Citraro, Sasan Mahmoodi, Angela Darekar, and Brigitte Vollmer. Extended three-dimensional rotation invariant local binary patterns. *Image and vision Computing*, 62:8–18, 2017.

Michael A Clynes, Camille Parsons, Mark H Edwards, Karen A Jameson, Nicholas C Harvey, A Aihie Sayer, Cyrus Cooper, and Elaine M Dennison. Further evidence of the developmental origins of osteoarthritis: results from the hertfordshire cohort study. *Journal of developmental origins of health and disease*, 5(6):453–458, 2014.

Antonio Criminisi. Machine learning for medical images analysis, 2016.

Agnaldo S Cruz, Hertz C Lins, Ricardo VA Medeiros, MF José Filho, and Sandro G da Silva. Artificial intelligence on the identification of risk groups for osteoporosis, a general review. *Biomedical engineering online*, 17(1):1–17, 2018.

Niamh M Cummins, EK Poku, Mark R Towler, OM O'Driscoll, and SH Ralston. Clinical risk factors for osteoporosis in ireland and the uk: a comparison of frax and qfracturescores. *Calcified tissue international*, 89:172–177, 2011.

Marleen De Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, 2016.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Elaine M Dennison, Holly E Syddall, A Aihie Sayer, Helen J Gilbody, and Cyrus Cooper. Birth weight and weight at 1 year are independent determinants of bone mass in the seventh decade: the hertfordshire cohort study. *Pediatric research*, 57(4):582–586, 2005.

MH Edwards, DE Robinson, KA Ward, MK Javaid, K Walker-Bone, C Cooper, and EM Dennison. Cluster analysis of bone microarchitecture from high resolution peripheral quantitative computed tomography demonstrates two separate phenotypes associated with high fracture risk in men and women. *Bone*, 88:131–137, 2016.

K Engelke, C Libanati, T Fuerst, Philippe Zysset, and HK Genant. Advanced ct based in vivo methods for the assessment of bone density, structure, and strength. *Current osteoporosis reports*, 11(3):246–255, 2013.

Yijie Fang, Wei Li, Xiaojun Chen, Keming Chen, Han Kang, Pengxin Yu, Rongguo Zhang, Jianwei Liao, Guobin Hong, and Shaolin Li. Opportunistic osteoporosis screening in multi-detector ct images using deep convolutional neural networks. *European Radiology*, 31(4):1831–1842, 2021.

Abdolhossein Fathi and Ahmad Reza Naghsh-Nilchi. Noise tolerant local binary pattern operator for efficient texture analysis. *Pattern Recognition Letters*, 33(9):1093–1100, 2012.

Janis Fehr and Hans Burkhardt. 3d rotation invariant local binary patterns. In *2008 19th International conference on pattern recognition*, pages 1–4. IEEE, 2008.

Uran Ferizi, Harrison Besser, Pirro Hysi, Joseph Jacobs, Chamith S Rajapakse, Cheng Chen, Punam K Saha, Stephen Honig, and Gregory Chang. Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from mri data. *Journal of Magnetic Resonance Imaging*, 49(4):1029–1038, 2019.

Joel S Finkelstein, Sarah E Brockwell, Vinay Mehta, Gail A Greendale, MaryFran R Sowers, Bruce Ettinger, Joan C Lo, Janet M Johnston, Jane A Cauley, Michelle E Danielson, et al. Bone mineral density changes during the menopause transition in a multiethnic cohort of women. *The Journal of Clinical Endocrinology & Metabolism*, 93(3):861–868, 2008.

Nicholas R Fuggle, Leo D Westbury, Gregorio Bevilacqua, Philip Titcombe, Mícheál Ó Breasail, Nicholas C Harvey, Elaine M Dennison, Cyrus Cooper, and Kate A Ward. Level and change in bone microarchitectural parameters and their relationship with previous fracture and established bone mineral density loci. *Bone*, 147:115937, 2021.

Nicholas R Fuggle, Shengyu Lu, Mícheál Ó Breasail, Leo D Westbury, Kate A Ward, Elaine Dennison, Sasan Mahmoodi, Mahesan Niranjan, and Cyrus Cooper. Oa22 machine learning and computer vision of bone microarchitecture can improve the fracture risk prediction provided by dxa and clinical risk factors. *Rheumatology*, 61 (Supplement_1):keac132–022, 2022.

Harry K Genant, Chun Y Wu, Cornelis Van Kuijk, and Michael C Nevitt. Vertebral fracture assessment using a semiquantitative technique. *Journal of bone and mineral research*, 8(9):1137–1148, 1993.

Shivanand S Gornale, Pooja U Patravali, and Ramesh R Manza. Detection of osteoarthritis using knee x-ray image analyses: a machine vision based approach. *Int. J. Comput. Appl*, 145(1):20–26, 2016.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing*, 19(6): 1657–1663, 2010.

Zhenhua Guo, Xingzheng Wang, Jie Zhou, and Jane You. Robust texture image representation by scale selective local binary patterns. *IEEE Transactions on Image Processing*, 25(2):687–699, 2015.

Patrick Haentjens, Jay Magaziner, Cathleen S Colón-Emeric, Dirk Vanderschueren, Koen Milisen, Brigitte Velkeniers, and Steven Boonen. Meta-analysis: excess mortality after hip fracture among older women and men. *Annals of internal medicine*, 152 (6):380–390, 2010.

Dongmei Han, Qigang Liu, and Weiguo Fan. A new image classification method using cnn transfer learning and web data augmentation. *Expert Systems with Applications*, 95:43–56, 2018.

Stinus Hansen, Claire Gudex, Fabian Åhrberg, Kim Brixen, and Anne Voss. Bone geometry, volumetric bone mineral density, microarchitecture and estimated bone strength in caucasian females with systemic lupus erythematosus. a cross-sectional study using hr-pqct. *Calcified tissue international*, 95(6):530–539, 2014.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Julia Hippisley-Cox and Carol Coupland. Predicting risk of osteoporotic fracture in men and women in england and wales: prospective derivation and validation of qfracturescores. *Bmj*, 339, 2009.

Le Phuong Thao Ho. *Development and evaluation of personalized risk assessments for osteoporotic patients*. PhD thesis, 2018.

Mari Hoff, Eva Skovlund, Haakon E Meyer, Arnulf Langhammer, Anne-Johanne Søgaard, Unni Syversen, Kristin Holvik, Bo Abrahamsen, and Berit Schei. Does treatment with bisphosphonates protect against fractures in real life? the hunt study, norway. *Osteoporosis International*, 32(7):1395–1404, 2021.

Frederick H Hooven, Jonathan D Adachi, Silvano Adami, Steven Boonen, J Compston, Cyrus Cooper, Pierre Delmas, Adolfo Diez-Perez, S Gehlbach, Susan L Greenspan, et al. The global longitudinal study of osteoporosis in women (glow): rationale and study design. *Osteoporosis international*, 20:1107–1116, 2009.

Chen-I Hsieh, Kang Zheng, Chihung Lin, Ling Mei, Le Lu, Weijian Li, Fang-Ping Chen, Yirui Wang, Xiaoyun Zhou, Fakai Wang, et al. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nature communications*, 12(1):1–9, 2021.

Ting-Yao Hu, Zhi-Qi Cheng, and Alexander G Hauptmann. Subspace representation learning for few-shot image classification. *arXiv preprint arXiv:2105.00379*, 2021.

Xufeng Huang, Qiang Lei, Tingli Xie, Yahui Zhang, Zhen Hu, and Qi Zhou. Deep transfer convolutional neural network and extreme learning machine for lung nodule diagnosis on ct images. *Knowledge-Based Systems*, 204:106230, 2020.

Dildar Hussain and Seung-Moo Han. Computer-aided osteoporosis detection from dxa imaging. *Computer methods and programs in biomedicine*, 173:87–107, 2019.

Helena Johansson, Kristín Siggeirsdóttir, Nicholas C Harvey, Anders Odén, Vilmundur Gudnason, Eugene McCloskey, Gunnar Sigurdsson, and John A Kanis. Imminent risk of fracture after fracture. *Osteoporosis International*, 28:775–780, 2017.

Domingos Alves Dias Júnior, Luana Batista da Cruz, João Otávio Bandeira Diniz, Giovanni Lucca França da Silva, Geraldo Braz Junior, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. Automatic method for classifying covid-19 patients based on chest x-ray images, using deep features and pso-optimized xgboost. *Expert Systems with Applications*, 183: 115452, 2021.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

JA Kanis, Olof Johnell, Anders Odén, Helena Johansson, and EFRAX McCloskey. Frax™ and the assessment of fracture probability in men and women from the uk. *Osteoporosis international*, 19(4):385–397, 2008.

John A Kanis, L Joseph Melton, Claus Christiansen, Conrad C Johnston, Nikolai Khaltaev, et al. The diagnosis of osteoporosis. *J Bone Miner Res*, 9(8):1137–1141, 1994.

M Katsoulis, V Benetou, T Karapetyan, D Feskanich, F Grodstein, Ulrika Pettersson-Kymmer, Sture Eriksson, Tom Wilsgaard, L Jørgensen, LA Ahmed, et al. Excess mortality after hip fracture in elderly persons from europe and the usa: the chances project. *Journal of internal medicine*, 281(3):300–310, 2017.

Niyazi Kilic and Erkan Hosgormez. Automatic estimation of osteoporotic fracture cases by using ensemble learning approaches. *Journal of medical systems*, 40(3):1–10, 2016.

Anne Klibanski, Lucile Adams-Campbell, Tamsen Bassford, Steven N Blair, Scott D Boden, Kay Dickersin, David R Gifford, Lou Glasse, Steven R Goldring, Keith Hruska, et al. Osteoporosis prevention, diagnosis, and therapy. *Journal of the American Medical Association*, 285(6):785–795, 2001.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Christian Kruse, Pia Eiken, and Peter Vestergaard. Machine learning principles can improve hip fracture prediction. *Calcified tissue international*, 100(4):348–360, 2017.

Andres Laib, Hans Jörg Häuselmann, and Peter Rüegsegger. In vivo high resolution 3d-qct of the human forearm. *Technology and health care*, 6(5-6):329–337, 1998.

Lisa Langsetmo, Katherine W Peters, Andrew J Burghardt, Kristine E Ensrud, Howard A Fink, Peggy M Cawthon, Jane A Cauley, John T Schousboe, Elizabeth Barrett-Connor, Eric S Orwoll, et al. Volumetric bone mineral density and failure load of distal limbs predict incident clinical fracture independent of frax and clinical risk factors among older men. *Journal of Bone and Mineral Research*, 33(7):1302–1311, 2018.

T Le Corroller, J Halgrin, M Pithioux, D Guenoun, P Chabrand, and P Champsaur. Combination of texture analysis and bone mineral density improves the prediction of fracture load in human femurs. *Osteoporosis International*, 23(1):163–169, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ki-Sun Lee, Seok-Ki Jung, Jae-Jun Ryu, Sang-Wan Shin, and Jinwook Choi. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. *Journal of clinical medicine*, 9(2):392, 2020.

Zhi Li, Guizhong Liu, Yang Yang, and Junyong You. Scale-and rotation-invariant local binary pattern using scale-adaptive texton and subuniform-based circular shift. *IEEE Transactions on Image Processing*, 21(4):2130–2140, 2011.

Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

H Ling and Alan C Bovik. Smoothing low-snr molecular images via anisotropic median-diffusion. *IEEE transactions on medical imaging*, 21(4):377–384, 2002.

Mirjam A Lips, Holly E Syddall, Tom R Gaunt, Santiago Rodriguez, Ian NM Day, Cyrus Cooper, Elaine M Dennison, Southampton Genetic Epidemiology Research Group, et al. Interaction between birthweight and polymorphism in the calcium-sensing receptor gene in determination of adult bone mass: the hertfordshire cohort study. *The Journal of rheumatology*, 34(4):769–775, 2007.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

AE Litwic, LD Westbury, Kathryn Ward, Cyrus Cooper, and EM Dennison. Adiposity and bone microarchitecture in the glow study. *Osteoporosis International*, 32:689–698, 2021.

Anna Ewa Litwic. *Bone microstructure and self-perception of fracture risk among women participating in the UK arm of the GLOW study.* PhD thesis, University of Southampton, 2020.

Li Liu, Lingjun Zhao, Yunli Long, Gangyao Kuang, and Paul Fieguth. Extended local binary patterns for texture classification. *Image and Vision Computing*, 30(2):86–99, 2012.

Li Liu, Songyang Lao, Paul W Fieguth, Yulan Guo, Xiaogang Wang, and Matti Pietikäinen. Median robust extended local binary pattern for texture classification. *IEEE Transactions on Image Processing*, 25(3):1368–1381, 2016.

Maximilian T Löffler, Alina Jacob, Andreas Scharr, Nico Sollmann, Egon Burian, Malek El Husseini, Anjany Sekuboyina, Giles Tetteh, Claus Zimmer, Jens Gempt, et al. Automatic opportunistic osteoporosis screening in routine ct: improved prediction of patients with prevalent vertebral fractures compared to dxa. *European Radiology*, pages 1–9, 2021.

Shengyu Lu, Nicholas R Fuggle, Leo D Westbury, Mícheál Ó Breasail, Gregorio Bevilacqua, Kate A Ward, Elaine M Dennison, Sasan Mahmoodi, Mahesan Niranjan, and Cyrus Cooper. Machine learning applied to hr-pqct images improves fracture discrimination provided by dxa and clinical risk factors. *Bone*, page 116653, 2022a.

Shengyu Lu, Sasan Mahmoodi, and Mahesan Niranjan. Robust 3d rotation invariant local binary pattern for volumetric texture classification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 578–584. IEEE, 2022b.

Rouzbeh Maani, Sanjay Kalra, and Yee-Hong Yang. Noise robust rotation invariant features for texture classification. *Pattern Recognition*, 46(8):2103–2116, 2013.

Samir D Mehta and Ronnie Sebro. Random forest classifiers aid in the detection of incidental osteoblastic osseous metastases in dexa studies. *International Journal of Computer Assisted Radiology and Surgery*, 14:903–909, 2019.

Nicholas Mikolajewicz, Nick Bishop, Andrew J Burghardt, Lars Folkestad, Anthony Hall, Kenneth M Kozloff, Pauline T Lukey, Michael Molloy-Bland, Suzanne N Morin, Amaka C Offiah, et al. Hr-pqct measures of bone microarchitecture predict fracture: systematic review and meta-analysis. *Journal of Bone and Mineral Research*, 35(3):446–459, 2020.

Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

MM Moinuddin, KA Jameson, HE Syddall, A Aihie Sayer, HJ Martin, S Robinson, C Cooper, and EM Dennison. Cigarette smoking, birthweight and osteoporosis in adulthood: results from the hertfordshire cohort study. *The Open Rheumatology Journal*, 2:33, 2008.

Urs J Muehlematter, Manoj Mannil, Anton S Becker, Kerstin N Vokinger, Tim Finkenstaedt, Georg Osterhoff, Michael A Fischer, and Roman Guggenberger. Vertebral body insufficiency fractures: detection of vertebrae at risk on standard ct images using texture analysis and machine learning. *European radiology*, 29(5):2207–2217, 2019.

Subrahmanyam Murala and QM Jonathan Wu. Spherical symmetric 3d local ternary patterns for natural, texture and biomedical image indexing and retrieval. *Neurocomputing*, 149:1502–1514, 2015.

Nathan J Neeteson, Bryce A Besler, Danielle E Whittier, and Steven K Boyd. Automatic segmentation of trabecular and cortical compartments in hr-pqct images using an embedding-predicting u-net and morphological post-processing. *Scientific Reports*, 13(1):252, 2023.

Nguyen D Nguyen, Steven A Frost, JR Center, John A Eisman, and Tuan V Nguyen. Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks. *Osteoporosis International*, 19(10):1431–1444, 2008.

KK Nishiyama, HM Macdonald, DA Hanley, and SK Boyd. Women with previous fragility fractures can be classified based on bone microarchitecture and finite element analysis measured with hr-pqct. *Osteoporosis international*, 24:1733–1740, 2013.

Kyle K Nishiyama and Elizabeth Shane. Clinical imaging of bone microarchitecture with hr-pqct. *Current osteoporosis reports*, 11:147–155, 2013.

Kyle K Nishiyama, Masako Ito, Atsushi Harada, and Steven K Boyd. Classification of women with and without hip fracture based on quantitative computed tomography and finite element analysis. *Osteoporosis International*, 25(2):619–626, 2014.

Xiao-Xiao Niu and Ching Y Suen. A novel hybrid cnn–svm classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4):1318–1325, 2012.

Shunjiro Noguchi, Mizuho Nishio, Masahiro Yakami, Keita Nakagomi, and Kaori Togashi. Bone segmentation on whole-body ct using convolutional neural network with novel data augmentation techniques. *Computers in biology and medicine*, 121: 103767, 2020.

Frank Q Nuttall. Body mass index: obesity, bmi, and health: a critical review. *Nutrition today*, 50(3):117, 2015.

Anders Oden, Eugene V McCloskey, John A Kanis, Nicholas C Harvey, and Helena Johansson. Burden of high fracture probability worldwide: secular increases 2010–2040. *Osteoporosis International*, 26(9):2243–2248, 2015.

Great Britain. Office of Population Censuses and Surveys. *Standard Occupational Classification*, volume 2. HM Stationery Office, 1995.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

World Health Organization et al. *Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: report of a WHO study group [meeting held in Rome from 22 to 25 June 1992]*. World Health Organization, 1994.

Zhibin Pan, Shiqi Hu, Xiuquan Wu, and Ping Wang. Adaptive center pixel selection strategy in local binary pattern for texture classification. *Expert Systems with Applications*, 180:115123, 2021.

Yves Pauchard, Anna-Maria Liphardt, Heather M Macdonald, David A Hanley, and Steven K Boyd. Quality control for bone quality parameters affected by subject motion in high-resolution peripheral quantitative computed tomography. *Bone*, 50(6): 1304–1310, 2012.

Angshuman Paul, Dipti Prasad Mukherjee, Prasun Das, Abhinandan Gangopadhyay, Appa Rao Chintha, and Saurabh Kundu. Improved random forest for classification. *IEEE Transactions on Image Processing*, 27(8):4012–4024, 2018.

Matti Pietikäinen, Timo Ojala, and Zelin Xu. Rotation-invariant texture classification using feature distributions. *Pattern recognition*, 33(1):43–52, 2000.

Albrecht W Popp, Helene Buffat, Ursula Eberli, Kurt Lippuner, Manuela Ernst, R Geoff Richards, Vincent A Stadelmann, and Markus Windolf. Microstructural parameters of bone evaluated using hr-pqct correlate with the dxa-derived cortical index and the trabecular bone score in a cohort of randomly selected premenopausal women. *PLoS One*, 9(2):e88946, 2014.

Zeng Qiang, Adu Jianhua, Sun Xiaoya, and Hong Sunyan. Extended complete local binary pattern for texture classification. *Multimedia Tools and Applications*, pages 1–17, 2021.

Farhan Riaz, Ali Hassan, Rida Nisar, Mario Dinis-Ribeiro, and Miguel Tavares Coimbra. Content-adaptive region-based color texture descriptors for medical images. *IEEE journal of biomedical and health informatics*, 21(1):162–171, 2015.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008.

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

Elizabeth J Samelson, Kerry E Broe, Hanfei Xu, Laiji Yang, Steven Boyd, Emmanuel Biver, Pawel Szulc, Jonathan Adachi, Shreyasee Amin, Elizabeth Atkinson, et al. Cortical and trabecular bone microarchitecture as an independent predictor of incident fracture risk in older women and men in the bone microarchitecture international consortium (bomic): a prospective study. *The lancet Diabetes & endocrinology*, 7(1): 34–43, 2019.

Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 2021.

Barbara C Silva, William D Leslie, Heinrich Resch, Olivier Lamy, Olga Lesnyak, Neil Binkley, Eugene V McCloskey, John A Kanis, and John P Bilezikian. Trabecular bone score: a noninvasive analytical method based upon the dxa image. *Journal of Bone and Mineral Research*, 29(3):518–530, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Anushikha Singh, Malay Kishore Dutta, Rachid Jennane, and Eric Lespessailles. Classification of the trabecular bone structure of osteoporotic patients using machine vision. *Computers in biology and medicine*, 91:148–158, 2017.

Julien Smets, Enisa Shevroja, Thomas Hügle, William D Leslie, and Didier Hans. Machine learning solutions for osteoporosis—a review. *Journal of Bone and Mineral Research*, 36(5):833–851, 2021.

Elisabeth Sornay-Rendu, Stephanie Boutroy, François Duboeuf, and Roland D Chapurlat. Bone microarchitecture assessed by hr-pqct as predictor of fracture risk in postmenopausal women: the ofely study. *Journal of Bone and Mineral Research*, 32(6): 1243–1251, 2017.

Jaime Lynn Speiser, Michael E Miller, Janet Tooze, and Edward Ip. A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, 134:93–101, 2019.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

D Sukumar, Y Schlussel, CS Riedt, C Gordon, T Stahl, and SA Shapses. Obesity alters cortical and trabecular bone density and geometry in women. *Osteoporosis International*, 22(2):635–645, 2011.

HE Syddall, A Aihie Sayer, EM Dennison, HJ Martin, DJP Barker, and C Cooper. Cohort profile: the hertfordshire cohort study. *International journal of epidemiology*, 34(6): 1234–1242, 2005.

Holly E Syddall, Shirley J Simmonds, Sarah A Carter, Sian M Robinson, Elaine M Dennison, Cyrus Cooper, Hertfordshire Cohort Study Research Group, et al. The hertfordshire cohort study: an overview. *F1000Research*, 8, 2019.

Sayed Mohamad Tabatabaei and Abdolah Chalechale. Noise-tolerant texture feature extraction through directional thresholded local binary pattern. *The Visual Computer*, 36(5):967–987, 2020.

Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

Hengliang Tang, Baocai Yin, Yanfeng Sun, and Yongli Hu. 3d face recognition using local binary patterns. *Signal Processing*, 93(8):2190–2198, 2013.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11): 4793–4813, 2020.

Akira Tsujii, Norimasa Nakamura, and Shuji Horibe. Age-related changes in the knee meniscus. *The Knee*, 24(6):1262–1270, 2017.

A Valentinitsch, S Trebeschi, Johannes Kaesmacher, C Lorenz, MT Löffler, C Zimmer, T Baum, and JS Kirschke. Opportunistic osteoporosis screening in multi-detector ct images via local classification of textures. *Osteoporosis international*, 30(6):1275–1285, 2019.

Alexander Valentinitsch, Janina M Patsch, Andrew J Burghardt, Thomas M Link, Sharmila Majumdar, Lukas Fischer, Claudia Schueller-Weidekamm, Heinrich Resch, Franz Kainberger, and Georg Langs. Computational identification and quantification of trabecular microarchitecture classes by 3-d texture analysis-based clustering. *Bone*, 54(1):133–140, 2013.

Insha Majeed Wani and Sakshi Arora. Computer-aided diagnosis systems for osteoporosis detection: A comprehensive survey. *Medical & Biological Engineering & Computing*, pages 1–45, 2020.

Kate A Ward, Camille M Pearse, Tafadzwa Madanhire, Alisha N Wade, June Fabian, Lisa K Micklesfield, and Celia L Gregson. Disparities in the prevalence of osteoporosis and osteopenia in men and women living in sub-saharan africa, the uk, and the usa. *Current Osteoporosis Reports*, pages 1–12, 2023.

Nelson B Watts. Fundamentals and pitfalls of bone densitometry using dual-energy x-ray absorptiometry (dxa). *Osteoporosis international*, 15:847–854, 2004.

Leo D Westbury, Clare Shere, Mark H Edwards, Cyrus Cooper, Elaine M Dennison, and Kate A Ward. Cluster analysis of finite element analysis and bone microarchitectural parameters identifies phenotypes with high fracture risk. *Calcified tissue international*, 105(3):252–262, 2019.

Danielle E Whittier, Steven K Boyd, Andrew J Burghardt, Julien Paccou, Ali Ghasem-Zadeh, Roland Chapurlat, Klaus Engelke, and Mary L Bouxsein. Guidelines for the assessment of bone density and microarchitecture in vivo using high-resolution peripheral quantitative computed tomography. *Osteoporosis International*, 31:1607–1627, 2020.

Norio Yamamoto, Shintaro Sukegawa, Akira Kitamura, Ryosuke Goto, Tomoyuki Noda, Keisuke Nakano, Kiyofumi Takabatake, Hotaka Kawai, Hitoshi Nagatsuka,

Keisuke Kawasaki, et al. Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates. *Biomolecules*, 10(11):1534, 2020.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Bin Zhang, Keyan Yu, Zhenyuan Ning, Ke Wang, Yuhao Dong, Xian Liu, Shuxue Liu, Jian Wang, Cuiling Zhu, Qinqin Yu, et al. Deep learning of lumbar spine x-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. *Bone*, 140:115561, 2020.

Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using volume local binary patterns. In *Dynamical vision*, pages 165–177. Springer, 2006.

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

Yang Zhao, Wei Jia, Rong-Xiang Hu, and Hai Min. Completed robust local binary pattern for texture classification. *Neurocomputing*, 106:68–76, 2013.

Keni Zheng and Sokratis Makrogiannis. Bone texture characterization for osteoporosis diagnosis using digital radiography. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1034–1037. IEEE, 2016.