COMPUTATIONAL INTELLIGENCE METHODS FOR MEDICAL IMAGE UNDERSTANDING, VISUALIZATION, AND INTERACTION

TAY WEI LIANG
(B.ENG. (HONS.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2013

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which may have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Tay Wei Liang

15 August 2013

Acknowledgments

I would like to thank my supervisor, Prof Ong Sim Heng, for his supervision and guidance during the course of my Ph.D. study. Both this thesis and my research publications would not have been possible without his assistance, patience, and understanding.

I would also like to express my gratitude to my co-supervisor, Dr Chui Chee Kong, for his mentorship and advice in spite of his busy schedule. Dr Chui has provided constant support and direction throughout my study, and he deserves a huge share of credit for my success.

Several colleagues have contributed their invaluable knowledge and assistance during my studies. Thank you, Nguyen Phu Binh, Wen Rong, Cai Lile, and Li Bing Nan. I have enjoyed working alongside them. In particular, I would especially like to thank Nguyen Phu Binh for his help and advice which greatly aided my research during my candidature.

My thanks also goes to Dr Alvin Ng Choong Meng for allowing the use of his medical datasets which served to support most of my research work, and for contributing his domain knowledge towards my research.

His insights have kickstarted my early work and influenced my subsequent research direction.

Last but not least, I would like to thank my friends and family for their continued encouragement and support.

Contents

Summary

Machine learning technologies are excellent for medical data analysis and are particularly useful when applied to medical imaging, where imaging modalities such as computed tomography (CT) or magnetic resonance imaging (MRI) can generate large amounts of 3-D or 4-D image data which can be costly or difficult to manually analyze. While machine learning methods have achieved some success in computer-aided diagnosis for medicine, they can also be applied to non-diagnostic medical applications. Machine learning can be used to support clinicians in making medical decisions by analyzing medical data and focusing the clinician attention on important or relevant items, or to simplify or automate medical tasks for labor savings. This thesis explores the use of machine learning methods for medical image data analysis, such that the medical data can be more easily understood, visualized, and interacted with.

This thesis first describes an image-understanding approach using robust regression for opportunistic osteopenia screening. A new method modeling the methodology of DXA scans was applied to extract a CT-based areal bone mineral density (aBMD) equivalent of dual-energy X-ray absorptiometry (DXA) aBMD. The extracted information was then robustly correlated with DXA aBMD to obtain a calibration mapping from CT aBMD to DXA aBMD. Experimental results showed that the method of estimating aBMD from dCT is feasible, and that CT aBMD can be applied to accurately diagnose bone diseases such as osteopenia.

The second contribution of this thesis expands upon the screening of osteopenia by introducing two ensemble methods for classification and regression. For classification of osteoporosis, an algorithm automatically extracts a basket of grey-level and morphological features from CT scans of the lumbar vertebrae, and uses a genetic algorithm as a meta-learner to ensemble the outputs of several basic classifiers. The genetic algorithm ensemble improves upon the classification performance across multiple operating points and diagnoses osteopenia with high accuracy. An ensemble-based regression network was also developed to further improve the regression of CT and DXA aBMD by incorporating multimodal features obtained from non-CT modalities. A filtering-based metalearner scheme was employed to build feature-wise ensembles from multimodal medical data with a high relative dimensionality. These contributions allow for improved diagnostic accuracy, and increases the confidence and transparency in algorithmic screening.

The third contribution presented is a clustering-based method to design transfer functions for intelligent context-based visualization. Clustering is applied to a 2-D low-high histogram to group voxels into several clusters, where each cluster of voxels belong to the same object-object interface. The clustering-based method then automatically assigns optical properties to the each detected object boundary without extensive parameter tuning, or can be used to simplify the transfer function space into meaningful regions that are more intuitive for operators to manipulate. The visualization results obtained using the clustering-based method approach that of existing state-of-the-art transfer function design approaches, while requiring much less user interaction and parameter tuning.

Lastly, this thesis introduces a method for multi-user biometric recognition in a gesture-based surgical data access system, where palms are used to identify users and load the specific work environments specific to each user. Several novelties for one-class classifiers were introduced to correctly recognize and classify palms of previously registered users, while rejecting unknown and unregistered users. The results demonstrate that modified one-class classifier systems are useful for learning the properties of unknown distributions and discriminating against unknown classes. The biometric recognition system developed has potential to be deployed in several other data access interfaces.

The machine learning techniques presented in this work allow for the useful information contained within large medical image datasets to be extracted for diagnostic, exploration, or visualization purposes. These contributions may also be useful in the analysis of other types of large data, such as in scientific visualization or data mining.

List of Tables

List of Figures

List of Abbreviations

aBMD	areal bone mineral density
aBMD$_{CT}$	areal bone mineral density from dCT
aBMD$_{DXA}$	areal bone mineral density from DXA
AUROC	area under ROC
BMC	bone mineral content
BMD	bone mineral density
CT	computed tomography
dCT	diagnostic computed tomography
DXA	dual energy x-ray absorptiometry
EVWE	evolved weighted ensemble
GA	genetic algorithm
HU	Housfield units
K-NN	k-nearest neighbor

LH	low-high
LOOCV	leave-one-out cross-validation
MLP	multi-layer perceptron
MRI	magnetic resonance imaging
NN-d	nearest neighbor distances
NNE	nearest neighbor ensemble
QCT	quantative computed tomography
RFE	random forest ensemble
ROC	receiver operating characteristic
RSME	root mean square error
SVM	support vector machine
TF	transfer function
vBMD	volumetric bone mineral density
WHO	World Health Organization

CHAPTER 1

Introduction

1.1 Motivation

Medical imaging is an extremely important tool in the diagnosis and detection of diseases [1]. There are several medical imaging modalities available, varying from radiological scanning devices such as x-ray and computed tomography (CT) to non-radiological modalities such as magnetic resonance imaging (MRI) and ultrasound. All of the above techniques can generate copious amounts of medical data, especially modalities capable of three-dimensional (3-D) or even four-dimensional (3-D + time) data capture. Advances in medical imaging technology have also increased imaging resolutions and thus the size of medical datasets. Interpretation of volumetric datasets or dynamic/time-series datasets is extremely difficult, and hence experienced medical personnel are required to interpret the image data, which translates into increased time and cost in analyzing and studying the medical data. Furthermore, different physicians may give differing interpretations when presented with the same data (inter-observer variance), and the same physician may even propose a different result when presented with the same data on different occasions (intra-observer variance).

Machine learning can play an important roles in the analysis and visualization of medical data. Machine learning algorithms can efficiently and effectively handle the large volumes of medical data, thus reducing the dependence on expert labor [1]. In particular, the increased amount of medical data ceases to be a weakness and instead becomes an advantage as machine learning is better able to uncover subtle and hidden relationships to disease conditions with larger databases. Machine learning is therefore especially helpful for screening applications, where computer-aided analysis can reduce the cost of mass screening and draw the experts attention onto more difficult clinical cases or onto image regions that may contain malignant elements [2].

Machine learning also lends itself to automated medical image understanding, which extends upon computer-aided diagnosis. The aim of image understanding is to build a system which can analyze images to draw conclusions about the nature of the observed disease process and the way in which this pathology can be overcome using various therapeutic methods. Image understanding constructs a semantic understanding of the underlying medical condition, therefore improving the reliability and comprehensibility of the computed results [2, 3]. Image understanding can be used to study medical conditions for diagnosis, or even to assist in the visualization of medical volumes [3].

It is clear that machine learning can provide the means for efficient processing, management, and reasoning for problems in medicine and healthcare. Therefore, the objective of this thesis is to explore the ways in which machine learning can address new issues in medicine, and to develop new machine learning solutions for tackling these problems.

1.2 Thesis Contributions

This thesis attempts to apply and develop machine learning techniques to handle several different problems faced in medicine.

1. *How can a relationship between dual energy X-ray absorptiometry (DXA) and CT be established such that a result from one modality can be converted into an equivalent result in the other modality?* In the medical diagnosis of osteoporosis, the golden clinical standard is typically established using DXA imaging. Results from other imaging modalities, such as CT, are not accepted for osteoporosis diagnosis despite the existence of strong similarities between the imaging modalities. In such situations, an algorithm to map the results from CT to DXA would be useful in allowing opportunistic screening of osteoporosis.

2. *How can structural and morphological features of bone be estimated from diagnostic computed tomography, and how can the estimated features diagnose osteopenia?* Osteoporosis is diagnosed based on the bone mineral density, but this measure does not include the structural or morphological information that is also contained in medical images. Additional information can be extracted from medical images to improve accuracy of osteoporosis diagnosis.

3. *In osteopenia screening, how can multimodal medical data be used to predict bone mineral density, and what insights into the disease condition can be obtained from the prediction?* During medical examinations, besides medical imaging, it is not unusual for several other tests to be conducted. The results from these other tests forms an additional source of information that may be useful for disease diagnosis, or for obtaining further insights into the disease condition.

4. *In direct volume rendering of medical volume data, how can transfer functions be automatically designed while allowing for important structures to be visualized?* The appearance of a rendered volume is dependent on the transfer function used to assign the optical properties. Transfer function design is difficult as it requires the understanding of the structures in the volume, and the transfer function domain. An automatic or semi-automatic transfer function design

greatly reduces the amount of expert intervention required in medical visualization.

5. *How can multiple surgeons/clinicians quickly access personalized data and interfaces in an aseptic surgical environment?* For human-computer interaction in surgical environments, a touch-free computer interface is required for asepsis. Gesture-based approaches allow for touch-free interaction, but typical interaction interfaces are not streamlined to cater to a wide and varied user group with different interaction objectives. A biometric recognition system can automatically recognize the user and immediately customize the interface to match that user's requirements, thus offering faster access to data and functions.

1.3 Thesis Organization

This thesis is organized as follows. Chapter 2 provides the medical context for the subsequent chapters by introducing the condition of osteoporosis and describing the existing clinical techniques used in its diagnosis. Then, it describes an image-understanding approach using robust regression for opportunistic osteopenia screening, and reports on the results and findings after experimental evaluation.

Chapter 3 expands upon the screening of osteopenia by presenting an ensemble method for osteopenia classification. The chapter also introduces a genetic algorithm optimization scheme, and describes the features designed to quantify spinal bone properties.

Chapter 4 first compares several methods of multivariate linear regression. The chapter then presents an ensemble-based regression network that improves the regression of CT and DXA aBMD by incorporating multimodal features obtained from non-CT modalities.

Chapter 5 is devoted to a clustering-based method to design transfer functions for intelligent context-based visualization, where clustering is used to detect material boundaries in order to automatically assign optical properties to each surface.

Chapter 6 introduces a method for multi-user gesture recognition and interaction for surgical augmented reality. The chapter also introduces a biometric user-recognition system for a gesture-based surgical augmented reality application that uses one-class classifiers for user identification based on hand profiles.

Lastly, the conclusions of this thesis and the proposals for future work are given in Chapter 7.

CHAPTER 2

Robust Regression for Areal Bone Mineral Density Estimation from Diagnostic CT Images

The aim of traditional medical image analysis is to extract useful information from medical data, whereas the aim of medical image understanding is to obtain insight into the medical condition itself. There is a natural overlap between these fields, as an insight that has been data-mined can subsequently be used as a feature for future medical diagnosis. In this chapter, we demonstrate a method for medical image understanding by correlating two different imaging modalities to extract a relationship between the modalities. The extracted relationship can then be used to estimate important disease indicators from the more common imaging modality.

The two imaging modalities studied here are DXA and diagnostic computed tomography (dCT). The primary use of DXA is to measure bone mineral density (BMD) values for the diagnosis of osteoporosis, while dCT is a more general radiological imaging tool that is used for pre-surgical planning or general diagnosis. While DXA is the clinical gold standard used for osteoporosis detection, dCT also contains relevant densitometric information. Our motivation is to correlate DXA images with dCT images, such that a BMD value can be estimated from a dCT image. Opportunistic osteoporosis screening using routine CT images allows the physician to receive an early notification of potential bone loss and the opportunity to prescribe measures for early treatment or management.

2.1 Related Work

Osteoporosis is a skeletal disease characterized by low bone mass and microarchitectural deterioration of bone tissue with a consequent increase in bone fragility and susceptibility to fracture. The progression of osteoporosis is often gradual with few obvious symptoms before bone fracture [4, 5]. Therefore, osteoporosis has to be detected and treated early to avoid fragility fractures.

The main methods of diagnosing osteoporosis are the use of bone mineral density values measured by DXA and quantitative computed tomography (QCT). QCT can be distinguished from dCT in that it is a dedicated CT technique to determine BMD. QCT also requires the use of calibration, whereas dCT may be used in the absence of calibration for diagnosis or pre-surgical planning. While dCT is performed more frequently due to the generality of its application, bone assessments cannot currently be made based on dCT scans as the absence of calibration phantoms means that dCT-derived BMD values are less reliable than QCT-derived BMD values. dCT is also often performed with the use of an intravenous contrast agent, which further affects BMD measurements.

It has previously been shown that there is some correlation between uncalibrated CT images and BMD [6, 7]. There are several ways to exploit this densitometric information. QCT can be calibrated without a reference phantom by making comparisons with internal references such as the paraspinal muscle and subcutaneous fat [8]. Link et al. [9] conducted a study using cadaver spine samples and patient studies to replicate the calibration in absence of calibration phantoms, and then used the calibration data to obtain BMD estimates from contrast-enhanced QCT. A different line of investigation is to study the correlation between the CT images and bone mechanical properties of interest [10], such as bone density, elastic modulus [11], and bone strength [12]. Other studies have also determined by experiment conversion factors for estimating the volumetric BMD from non-dedicated contrast-enhanced standard MDCT images [13].

In recent years several papers have noted the possibility of screening for bone diseases from diagnostic or routine CT scans. Habashy et al. [14] investigated the estimation of bone mineral density in children based on dCT images and suggested that phantom-less QCT of dCT provides additional BMD information. The opportunistic screening of osteoporosis while performing CT colonography has been investigated by Pickhardt et al. [15], where the phantom-less QCT technique and a simple trabecular region-of-interest attenuation method was applied to dCT images performed for colonography and benchmarked against DXA reference. Several studies [16, 17] investigated the efficacy of BMD estimation techniques that do not require calibration phantoms; as expected, the precision of phantom-less techniques was lower compared to phantom-based QCT densitometry, but nonetheless promising for assessing fracture risk. It was also found [18] that the inclusion of calibration phantoms in dCT did not significantly affect the patient radiation dose, and hence bone loss screening may be conducted with little additional risk or cost.

Another popular approach was to use machine learning techniques to diagnose fractures [19] and osteoporotic diseases [20, 21] based on QCT images. These methods are capable of achieving good detection rates, but typically involve the use of black boxes, which makes it difficult to evaluate their reliability and generality. More extensive clinical validation is necessary, but artificial intelligence-based methods can be helpful in providing one indicator of bone disease.

While several papers have suggested the use of volumetric BMD as measured by QCT, areal bone mineral density (aBMD) from DXA remains the clinical standard for diagnosing osteoporotic diseases as it provides several advantages [22]. Biomechanical studies have shown that mechanical strength and DXA-derived BMD are strongly correlated [23], while prospective cohort studies have indicated a strong relationship between fracture risk and BMD measured by DXA [24]. Most importantly, the World Health Organization (WHO) criteria for the diagnosis of osteoporosis and for input into the fracture risk algorithm (FRAX) are based on reference data obtained by DXA [25]. As the body of work based on DXA-derived aBMD ($\mathrm{aBMD_{DXA}}$) remains more well-established than that based on volumetric BMD, it may be more feasible to determine a DXA-equivalent aBMD score from diagnostic CTs. This estimated $\mathrm{aBMD_{CT}}$ value may be directly interpreted by a physician according to existing diagnosis guidelines based on DXA.

2.2 Areal Bone Mineral Density Estimation from Diagnostic CT Images

2.2.1 Background

DXA uses two X-rays of different energies to capture a posteoanterior image of the patient's spine [26]. The absorption of each beam by bone allows the amount of bone mineral, known as the bone mineral content (BMC), in each vertebrae to be determined. This BMC is subsequently normalized by the projected vertebra's area to obtain the aBMD_{DXA}. On the other hand, the result of a dCT scan is a 3D image of the patient. We proposed to use the 3D volume from dCT to compute a similar posteoanterior projection of the spine, and compute an estimated aBMD_{CT}. Subsequently, regression techniques are used to map aBMD_{CT} to the actual aBMD_{DXA}.

2.2.2 Overview

Fig. 2.1 shows the algorithm for distinguishing osteopenic bone from normal bone. The screening algorithm consists of three major steps. The first step extracts the desired regions of interest (vertebral bodies) and performs simple Hounsfield units (HU) correction on the extracted vertebral bodies. The second step estimates aBMD_{CT} from the CT images of

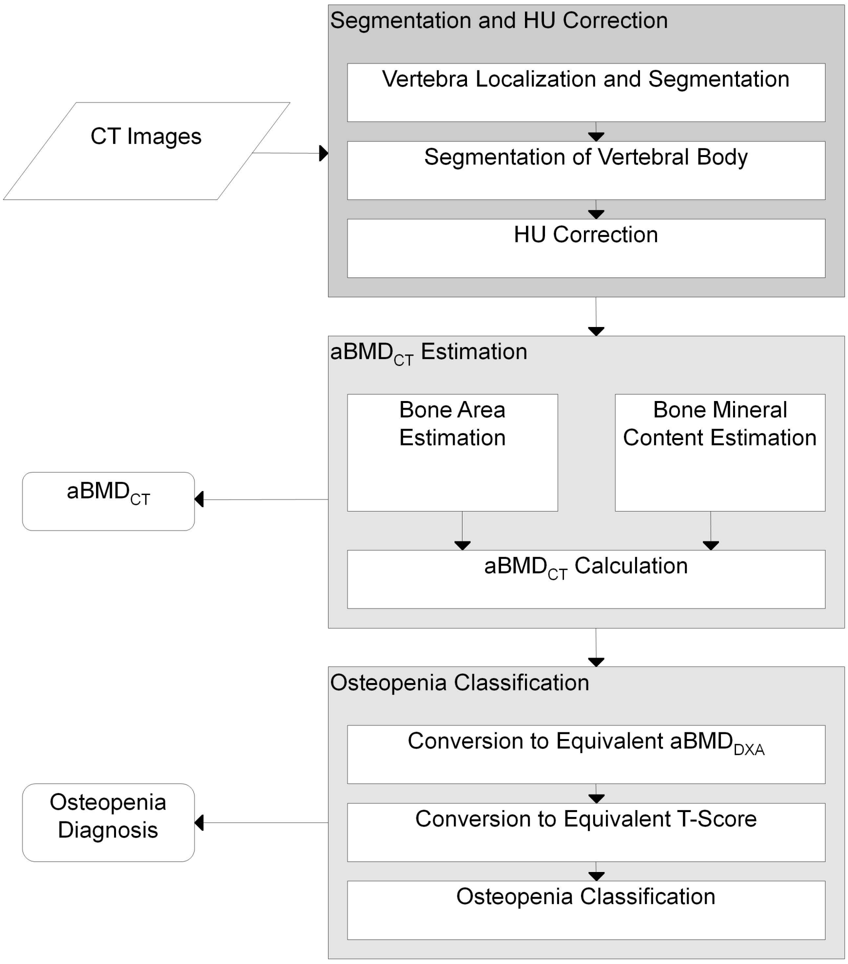

Segmentation and HU Correction
Vertebra Localization and Segmentation
Segmentation of Vertebral Body
HU Correction
CT Images
aBMD$_{CT}$ Estimation
Bone Area Estimation
Bone Mineral Content Estimation
aBMD$_{CT}$ Calculation
aBMD$_{CT}$
Osteopenia Classification
Conversion to Equivalent aBMD$_{DXA}$
Conversion to Equivalent T-Score
Osteopenia Classification
Osteopenia Diagnosis

Figure 2.1: Overview of the three-stage aBMD prediction and osteopenia screening system, performing preprocessing, aBMD prediction, and osteopenia classification tasks respectively.

the vertebral bodies by determining the area and bone mineral content of the vertebral body. The final step converts the aBMD$_{\mathrm{CT}}$ estimate to its aBMD$_{\mathrm{DXA}}$ equivalent and performs an osteopenia diagnosis using the T-score. The entire process is automated and requires no additional user input.

2.2.3 Vertebral Body Segmentation and HU Correction

This module automatically segments the vertebral body from the routine CT image and applies a HU correction on the segmented vertebral body to control for imaging performed under different beam calibration conditions. There are three sequential steps, of which two are segmentation steps and the final one being a HU correction procedure. The first segmentation step localizes the approximate position of the vertebra and performs a graph cut to obtain the entire vertebra. The second segmentation step takes the segmented vertebra and determines an appropriate cut to isolate the vertebral body from the vertebral processes. Finally, we use the HU of the adjacent paraspinal muscle to perform a correction to the HU of the segmented vertebral body.

Vertebral Localization and Segmentation

The localization of the main vertebra section is performed by an iterative window shifting technique which is inspired by mean shift clustering. First, a fixed threshold based on the likely HU for bone is used to obtain an initial segmentation of the bone regions. The centroid of the bone regions is then taken as an initial guess C_1 for the centroid of the vertebra. A local window centered about C_1 and twice the size of a typical vertebra is applied to the images, and the centroid of the bone regions contained within the local window is used as the second estimate C_2 for the vertebra centroid. The local window is subsequently re-centered to C_2 and used to produce another guess C_3 at the centroid. This iterative process continues until the centroid position converges to a static value C_{end}. The algorithm is summarized below:

1. A fixed threshold of HU > 400 is used to perform an initial segmentation of bone.

2. The centroid of the bone areas is computed as C_1.

3. A local window of twice the size of a vertebra is placed on the volume, centered about C_1.

4. The centroid of the bone areas contained within the local window is computed as C_2.

5. Repeat steps 3-4 using the latest centroid guess, until convergence to a centroid value of C_{end}.

The localization procedure captures a local window centered about the vertebra at C_{end}. The initial thresholding used to obtain the initial bone classification is not sufficiently accurate to distinguish between bone tissues for correlation and prediction, particularly for estimating the aBMD.

A graph cut algorithm [27, 28] is used instead to perform a more refined segmentation of the vertebra from the local window. Graph cut is an optimization technique commonly used in computer vision to divide an image into object and background regions. An image is represented as a graph, and the graph cut algorithm obtains a minimum set of link cuts such that the entire graph is divided into two disjoint sets of background or object nodes. The result of the graph cut is a clean segmentation of the vertebra from the surrounding tissues.

Vertebral Body Segmentation

The spinal processes (Fig. A.1) are not relevant for bone strength as the main determinant of bone strength is the vertebral body. The segmentation of the vertebral body is therefore an important step in the algorithm. To ensure repeatability of the vertebral body segmentation, the spinal canal is used as an anatomical landmark for the segmentation as it can be easily detected with high reliability. The center of the spinal canal is taken as one control point for determining the cutoff point for the vertebral body segmentation, while the centroid of the vertebral region lying above the spinal canal centroid is taken as the second control point. A line is extended to connect the two control points and profile analysis used to determine the position where there is an abrupt change in HU; this position is the boundary between the spinal canal and the vertebral body. A line perpendicular to the line connecting the two control points is used as the cutoff line to separate the vertebral body from the pedicles and the spinal processes. Finally, the upper region is taken as the vertebral body, and the lower region is taken as the spinal process. The vertebral body segmentation algorithm is summarized as:

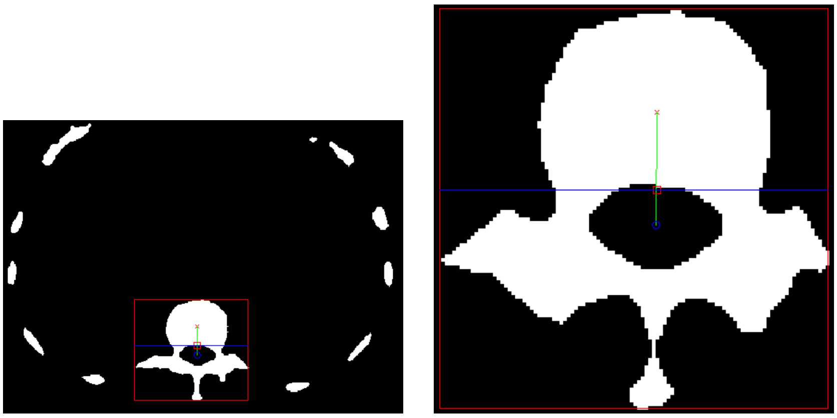
1. The spinal canal is located as a void in the vertebra and the centroid of the spinal canal, C_{sc}, is computed.

2. The centroid of the bone region lying above C_{sc} is used as a guess for the centroid of the vertebral body, C_{vb}.

3. A line $\mathrm{L}_{\text{sc-vb}}$ is extended to connect C_{sc} and C_{vb}. The gray-level profile on this line is analyzed to find a point $\mathrm{P}_{\mathrm{cutoff}}$ where there is a sudden change in HU.

4. A second line $\mathrm{L}_{\mathrm{cutoff}}$ passing through $\mathrm{P}_{\mathrm{cutoff}}$ is constructed perpendicular to $\mathrm{L}_{\text{sc-vb}}$. $\mathrm{L}_{\mathrm{cutoff}}$ is the cutoff line for the vertebral body segmentation.

5. All bone regions lying above $\mathrm{L}_{\mathrm{cutoff}}$ are labeled as vertebral body, while all bone regions lying below $\mathrm{L}_{\mathrm{cutoff}}$ are labeled as spinal processes.



(a) (b)

Figure 2.2: Two examples of vertebral body segmentation, where a) also includes the detected rib bones for context. In each image, the red outer boundary is the extracted ROI for the vertebra, the red "x" is the guess for the vertebral body centroid, the blue "o" is the centroid of the spinal canal. The green line is the line connecting the two centroids, and the red square and the blue lines are the detected cutoff point and cutoff line respectively.

The control points and segmentation lines generated using this segmentation algorithm are given in Fig. 2.2.

Intensity Correction

The HU of the CT image may differ based on the properties of the beam used to perform the CT scan. The energy spectrum of the X-ray beam affects the subsequent beam hardening when the X-ray passes through internal tissue. The algorithm proposed here must adapt to different imaging scenarios where the routine CT is obtained for diagnostic imaging purposes. A HU correction is therefore performed to reduce the variance in

HU resulting from different imaging parameters. Similar to the phantomless calibration method [8], the paraspinal muscles are used as an internal reference. We assume that the paraspinal muscles have ideal HU characteristics that do not vary significantly amongst patients, and thus the differences between the observed and ideal HU for the paraspinal muscles must largely be due to the differences in imaging parameters. Aligning the observed and ideal HU for the paraspinal muscles can therefore also correct the HU for the vertebrae.

The paraspinal muscles are first located by extending a local window horizontally about the spinal processes segmented in the previous step. The soft tissues contained within the window are assumed to consist of fat and muscle, each of which has HUs following independent Gaussian distributions. Expectation maximization is used to recover the model parameters that best explains the observed fat and muscle distribution [29]. The Gaussian mixture model is used to estimate the mode of the muscle tissue, which is used to compute the linear correction offset. The algorithm for the HUs intensity correction is:

1. A local window of twice the width of the vertebral body is extended about the spinal processes. All non-bone non-air voxels are labeled as soft tissue.

2. A Gaussian mixture model is adopted to model the soft tissues as fat

and muscle tissues [8]. Expectation maximization is used to estimate the means, standard deviations, and fractions of the fat and muscle tissues.

3. The mean of the muscle tissues, μ_{muscle}, is compared against the standard value for muscle, +40 [30]. A correction offset

$$\text{HU}_{\text{offset}} = +40 - \mu_{\text{muscle}} \tag{2.2.1}$$

is then added to each voxel of the segmented vertebral body.

2.2.4 Generation of aBMD$_{\text{CT}}$ from Routine CT

In earlier studies [11], a strong correlation was found between the HUs of a voxel and the bone mineral density ρ of that voxel. This relationship was described as:

$$\rho = 1.112 \times \text{HU} + 47 \text{ kg/m}^3. \tag{2.2.2}$$

As the volume of an individual voxel can be computed from the inter-slice spacing and the voxel spacing, this means that the bone mineral content of each voxel, and therefore the vertebral bone, can be estimated from the CT scan. For a given inter-slice spacing of S_y and a voxel spacing of S_x, the bone mineral content BMC$_{\text{CT}}$ can be estimated from the CT images as

$$\text{BMC}_{\text{CT}} = \sum \rho \times S_y \times S_x^2. \tag{2.2.3}$$

Furthermore, the area of the vertebral bone can be found by segmenting the vertebra and taking the projection area on the posteroanterior plane. The area of the bone, A_{bone} is equal to a multiple of the sum of bone pixels on the projection $\mathrm{A}_{\mathrm{pixel}}$:

$$A_{\mathrm{bone}} = A_{\mathrm{pixel}} \times S_y \times S_x. \tag{2.2.4}$$

Therefore, by dividing the estimated bone mineral content of the vertebra by the estimated area of the vertebra, a CT equivalent of the DXA aBMD can be obtained. The aBMD from CT, $\mathrm{aBMD}_{\mathrm{CT}}$ is calculated by

$$\mathrm{aBMD}_{\mathrm{CT}} = \frac{\mathrm{BMC}_{\mathrm{CT}}}{A_{\mathrm{bone}}}. \tag{2.2.5}$$

This $\mathrm{aBMD}_{\mathrm{CT}}$ may be used to gauge the bone condition and to perform a coarse diagnosis of bone diseases such as osteoporosis or osteopenia.

2.3 Robust Regression

2.3.1 Regression of $\mathrm{aBMD}_{\mathrm{DXA}}$ from $\mathrm{aBMD}_{\mathrm{CT}}$

$\mathrm{aBMD}_{\mathrm{CT}}$ is a coarse estimator of $\mathrm{aBMD}_{\mathrm{DXA}}$. It cannot be directly used to replace $\mathrm{aBMD}_{\mathrm{DXA}}$ because the bone areas and bone mineral contents used to calculate aBMD are obtained by the different radiological methods of DXA and CT. Some calibration is necessary to perform a conversion from $\mathrm{aBMD}_{\mathrm{CT}}$ to an $\mathrm{aBMD}_{\mathrm{DXA}}$ value. We assume that the $\mathrm{aBMD}_{\mathrm{CT}}$ and

aBMD_{DXA} values are related via a linear transformation of the form

$$\text{aBMD}_{\text{CT}} = k_1 \times \text{aBMD}_{\text{DXA}} + k_2, \tag{2.3.1}$$

where k_1 and k_2 are the scaling and offset constants respectively. This assumption of linearity is supported by experimental data provided in the results section. The values of the constants can be directly obtained by linear least squares regression, but the results will be adversely affected by the presence of large outliers due to infrequent but large errors in the estimation of vertebral area and bone mineral content. RAndom SAmple Consensus (RANSAC) [31] is used instead to obtain a robust estimation of the linear transformation parameters. The RANSAC procedure randomly selects pairs of points to construct linear models, and the available data is fitted to the tentative model. Points lying far away are treated as outliers and the model is only considered as a potential candidate if there are fewer than a preset number of outliers. For a valid candidate, the inlier points are collectively used to generate a regression fit. This process is continued for several iterations to yield a number of potential candidate models, which are evaluated on the basis of the standard deviation of the inlier points from the regression fit. The model with the minimum standard deviation is adopted as the best fitting model.

The RANSAC procedure is described as follows:

1. Two data points are randomly chosen to generate a linear model.

2. All data points with normalized errors of less than 0.30 are considered hypothetical inliers.

3. If more than $P_{threshold}$ (0.90) of all points are hypothetical inliers, a new linear model is estimated from all the hypothetical inliers. The sum of absolute errors of the hypothetical inliers from the new linear model is calculated and recorded along with the model parameters. Otherwise, the linear model is discarded.

4. Steps 1-3 are repeated for 1000 times.

5. The valid linear model with the lowest sum of absolute errors is used as the final regression model.

RANSAC is capable of forming outlier-free models by rejecting large random or systematic errors. Here, RANSAC is used to assist in the detection and rejection of large outliers. These outliers will subsequently be examined to determine the systematic cause, if any, that justifies their rejection.

2.3.2 Classification of Osteopenia from $\mathrm{aBMD_{CT}}$

In DXA, $\mathrm{aBMD_{DXA}}$ can be directly converted to the T-score by standardizing with respect to the $\mathrm{aBMD_{DXA}}$ of the reference population. The same standardization can be performed by first converting $\mathrm{aBMD_{CT}}$ into the estimated aBMD_{DXA} using the discovered correlations, and subsequently using the reference population to obtain the estimated T-score. The estimated T-score is used to diagnose osteoporosis and osteopenia in the same manner as conventional DXA T-scores, where bones with T-score of less than -1.0 are classified as osteopenic.

The classification rule can be modified to obtain other operating points. For example, the threshold can be increased to have a higher osteopenia detection rate at the cost of increased number of false positives. This trade-off is summarized in the receiver operating characteristic (ROC) graph, which plots the true positive rate against the false positive rate.

2.4 Results and Discussion

2.4.1 Data Sets

The data sets used in our experiments consist of paired CT scans and DXA measurements drawn from 44 male participants between 60 and 90 years of age (66.7 ± 7.47 years). The study selected patients with no preexisting medical conditions, and compression fractures and other degenerative pathologies were also excluded after radiologist review. This source data set was broken into 155 pairs of CT volumes and DXA measurements, with each pair capturing one of the vertebrae in the lumbar spine (L_1-

L_4). Approximately one-third (50) of the samples were osteopenic (46) or osteoporotic (4), while the remaining samples (105) had normal bone mineral density.

Abdominopelvic visceral adipose tissue (VAT) was determined using a 64-slice multi-detector CT scanner (Somatom Definiton, Siemens AG, Erlangen, Germany). Axial CT scan was performed with the subjects supine, from the dome of the diaphragm down to the bottom of the pelvis, using a 35 x 35 cm field of view. Non-contrast enhanced scans using routine scan parameters of 120 kVp, 210 mAs, slice collimation 0.6 mm, slice width 5.0 mm, pitch factor 1.4, and increment 5.0 mm were acquired. The thin-slice raw data was reconstructed into 1 mm sections with zero-gap intervals. No intravenous contrast agent was used in any of the CT scans.

2.4.2 Evidence of Correlation between aBMD$_{\mathbf{DXA}}$ and HU

aBMD$_{\mathrm{DXA}}$ was correlated with the mean HUs calculated from the top, middle, and bottom slices of the volume, and using all vertebral slices in the volume respectively. The squared correlation coefficients (r^2) are shown in Table 2.1, with the correlation coefficient (r) contained in brackets.

Table 2.1: Correlation coefficients using different slice sampling schemes.

	Top, Middle, Bottom Slices	Entire Volume
Mean without RANSAC	0.286 (0.535)	0.478 (0.691)
Mean with RANSAC	0.465 (0.682)	0.647 (0.804)

The raw correlation results for the method computed on the top, middle, and bottom slices without outlier rejection are poorer than the figure ($r^2 = 0.44$) reported in [7]; however, with RANSAC enabled, the correlation coefficients agree. The four samples rejected by RANSAC were found to be poorly segmented or to have osteophytes, and hence their removal was justified. Table 2.1 shows that when the entire volume is used in its computation, the mean feature correlates more strongly with the aBMD$_{\text{DXA}}$ value. This improvement in the degree of correlation occurs regardless of whether outlier rejection is used. The result suggests that it is always better to include the entire volume rather than relying on a partial selection of axial sections from the bone volume; this may be because noise and partial volume effects are reduced through averaging from several slices.

2.4.3 Estimating aBMD$_{\text{DXA}}$ from aBMD$_{\text{CT}}$

Fig. 2.3 shows the Bland-Altman plot. A systematic bias of -0.0817 g/cm^2 for aBMD$_{\text{CT}}$ was detected, while the standard deviation (SD) was 0.0908 g/cm^2. The systematic bias in the aBMD$_{\text{CT}}$ measurement can be corrected by a linear fitting model. Fig. 2.4 plots the true $\mathrm{aBMD}_{\mathrm{DXA}}$ value against the computed $\mathrm{aBMD}_{\mathrm{CT}}$. The experimental relationship between $\mathrm{aBMD}_{\mathrm{DXA}}$ and $\mathrm{aBMD}_{\mathrm{CT}}$ was found to be:

$$\mathrm{aBMD}_{\mathrm{DXA}} = 0.866 \times \mathrm{aBMD}_{\mathrm{CT}} + 0.194\ \mathrm{g/cm}^2. \qquad (2.4.1)$$

The correlation coefficient r^2 was 0.726 (r = 0.852). The root mean square error was 0.0884 $\mathrm{g/cm}^2$, which corresponds to a coefficient of variation of 8.77%. These results show that there is a strong correlation between the $\mathrm{aBMD}_{\mathrm{DXA}}$ value from the $\mathrm{aBMD}_{\mathrm{CT}}$ value, and that it is possible to predict the $\mathrm{aBMD}_{\mathrm{DXA}}$ value based on the $\mathrm{aBMD}_{\mathrm{CT}}$ value.

2.4.4 Impact of Different Bone Tissues on DXA Correlation

The results of the experiment are detailed in Table 2.2. Using only the cortical or trabecular bone regions as the region of interest resulted in lower correlation to $\mathrm{aBMD}_{\mathrm{DXA}}$ compared to the case of using both bone regions. This conclusion is reasonable since our technique estimates $\mathrm{aBMD}_{\mathrm{DXA}}$, and DXA does not differentiate between cortical and trabecular bone. Another reason for not distinguishing between the two bone tissue types is that it is very difficult to separate trabecular and cortical bone with high confidence due to the partial volume effect at the scan resolutions typical for dCT. Imaging trabecular bone requires high-resolution QCT

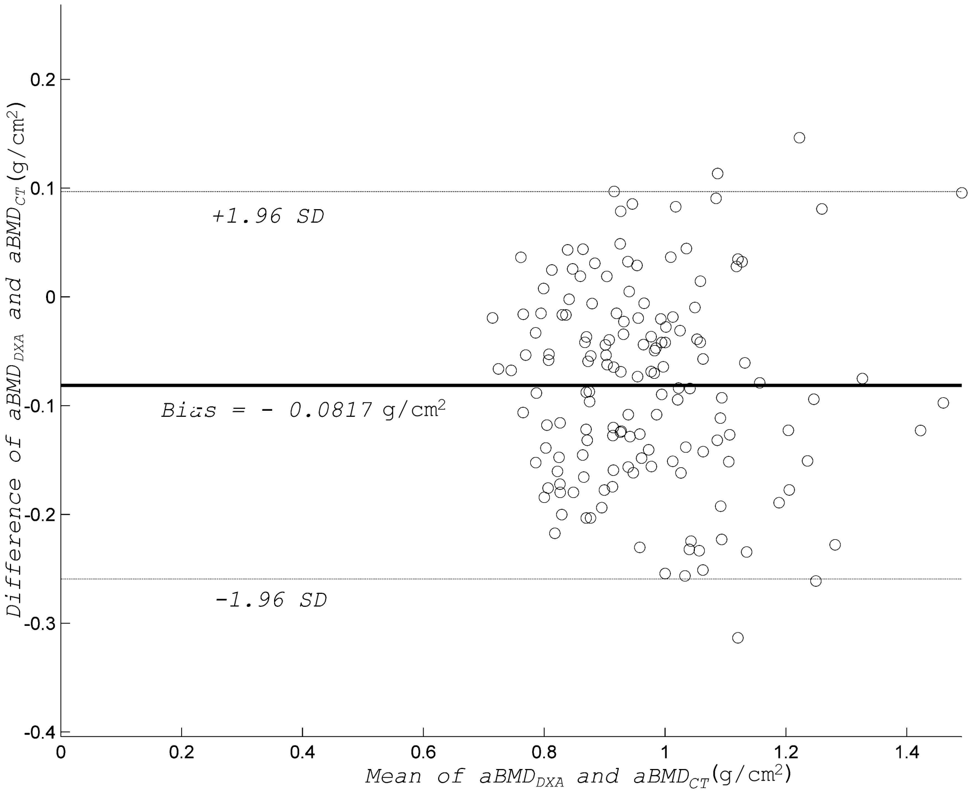


Figure 2.3: Bland-Altman plot of aBMD$_{\mathrm{DXA}}$ and aBMD$_{\mathrm{CT}}$. aBMD$_{\mathrm{CT}}$ systematically underestimates aBMD$_{\mathrm{CT}}$.

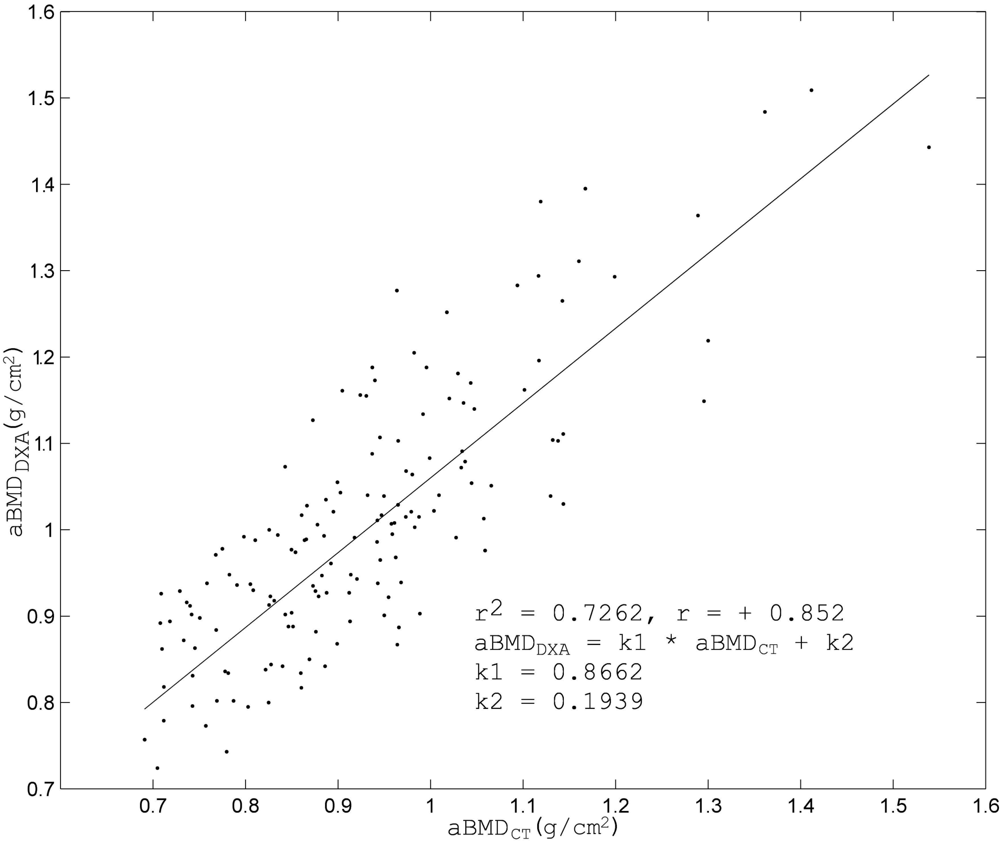


Figure 2.4: Regression plot of $\mathrm{aBMD_{DXA}}$ vs $\mathrm{aBMD_{CT}}$. 4 of 155 samples were rejected by RANSAC.

Table 2.2: Correlation of $\mathrm{aBMD_{DXA}}$ by computing $\mathrm{aBMD_{CT}}$ from different bone tissues.

Bone Tissue Used	r^2 (r)
Cortical bone	0.479 (0.692)
Trabecular bone	0.655 (0.809)
Both cortical and trabecular bone	0.726 (0.852)

and increased radiation exposure.

2.4.5 Osteopenia Classification based on T-score

At the cutoff point of T-score = -1.0, the aBMD classifier achieves an overall accuracy of 80.1% with a true and false positive rate of 73.9% and 17.1% respectively. The aBMD classifier attains an area under curve of 0.894 on the receiver operating characteristic curve, as shown in Fig. 2.5. The classifier performance is comparable in screening for osteopenia using advanced machine learning techniques [21], where AUC scores of 0.896 and 0.885 were reported.

2.4.6 vBMD and aBMD for Prediction and Classification

Table 2.3 shows the results of the comparision. While directly regressing $\mathrm{aBMD_{DXA}}$ from dCT is feasible and produces acceptable results, it is inferior to our original method using of an $\mathrm{aBMD_{DXA}}$ intermediate. The

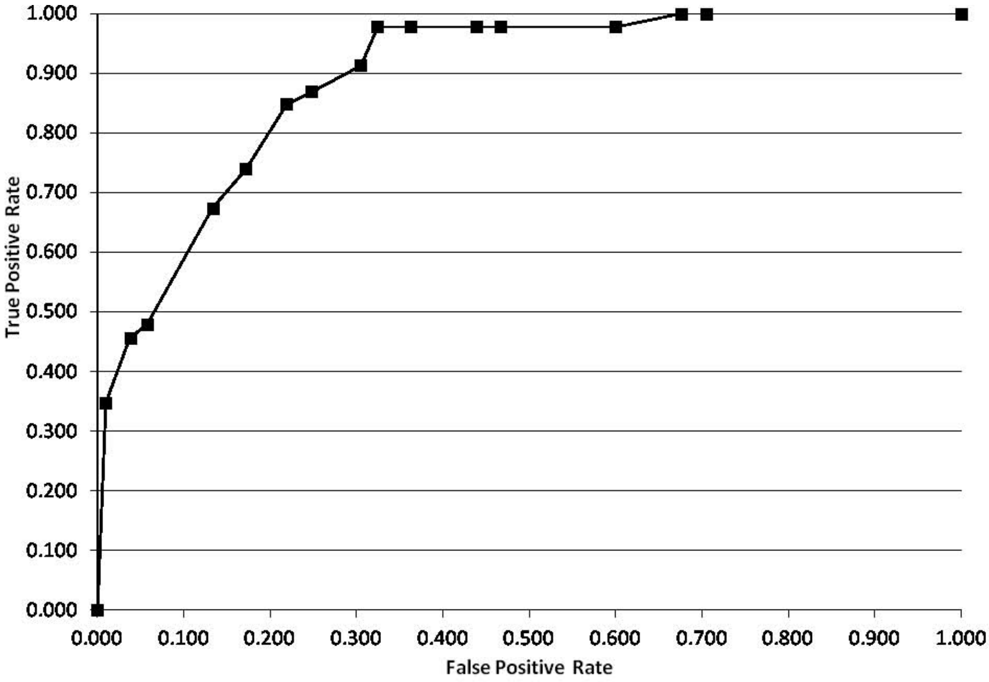


Figure 2.5: Receiver operator characteristic curve for a linear classifier using aBMD_{CT}.

Table 2.3: Comparison between vBMD and aBMD.

Method	r^2 (r)	RMSE (g/cm^2)	AUC
vBMD_{CT}	0.808 (0.653)	0.104	0.871
aBMD_{CT}	0.852 (0.726)	0.0884	0.894
Difference	-5.16% (-10.1%)	17.6%	-2.57%

vBMD_{CT} method has 17.6% greater aBMD_{DXA} estimation error than the aBMD_{CT} method, and has a corresponding decrease in classification performance by 2.57%. This difference mirrors the systematic discrepancy between the volumetric and areal measurements from QCT and DXA. It is more appropriate to compute aBMD_{CT} for predicting aBMD_{DXA}, because there are inherent differences between volumetric and areal measurements.

2.4.7 Discussion

Our results demonstrate that DXA-equivalent aBMD can be estimated from the aBMD_{CT} value derived from dCT. We have also shown that the derived aBMD value can be applied to accurately diagnose bone diseases such as osteopenia. These promising results suggest that the method of estimating aBMD from dCT is feasible.

We are optimistic that the aBMD estimation method would perform better in practice than the results reported here. The osteopenia diagnosis system using aBMD_{CT} was evaluated on its ability to distinguish between osteopenic and normal patients, whereas in clinical practice the aim of the screening application would be to separate osteoporotic and normal patients. Osteoporotic patients have an even more significant bone mineral loss than osteopenic patients, and osteoporosis should be easier to detect than osteopenia. Therefore, since osteoporotic cases were underrepresented in our experiments, we expect the real-world classification performance of the screening system to be improved. At the same time, we caution that we have not measured the precision of our technique. Since aBMD_{CT} approximates aBMD_{DXA}, it can at best only attain a precision equal to DXA (and more likely, worse than DXA). It should not be used to monitor bone density changes across time.

Several factors may affect the reliability of the aBMD estimation and osteopenia detection system. First, contrast agents were not used in any of our diagnostic CT scans and hence the effect of contrast agents cannot be studied and controlled for. Second, the aBMD$_{\mathrm{CT}}$ measure depends directly on the bone area detected, and is thus vulnerable to the mis-segmentation of bone and the resulting mis-estimation of bone area. However, these two issues may be sufficiently addressed by corrective measures. For intravenous contrast agents, correction algorithms for contrast agents have been reported in [14]. Mis-segmentation may be reduced by using automated algorithms for segmentation of bone, which can reduce the inter-operator and intra-operator variation associated with manual bone segmentation. Using more advanced segmentation algorithms in lieu of simple automated or semi-automated segmentation methods also increases the reliability of the bone segmentation.

There are some limitations to this study resulting from the patient population used. First, the study population consisted entirely of males, which may reduce the applicability of the described methods to women. This limitation may be especially important since women are generally considered to be at higher risk of osteoporosis than men. Second, the age of the participants covered only the range of 60 to 90 years, which may reduce the reliability of aBMD estimation and the osteopenia diagnosis system with a younger population. However, we note that several related studies suffer from similar age sampling limitations, the majority of which do not include healthy adults in their sample population as the risk (and therefore utility of screening for) bone loss is small. However, the methods described here should be used with caution when applied to a pediatric population, where DXA is known to be less reliable [14].

A further consideration is that the WHO definition for osteoporosis and osteopenia using BMD T-scores is meant to be applied to postmenopausal women. In clinical practice, due to the lack of a BMD-based definition for osteoporosis/osteopenia in males, it is not uncommon to apply the WHO definition to males as well [32]. This is the approach we have taken in this thesis, but there is evidence to suggest that applying the WHO standard to male populations underestimates osteoporosis risk and that a revised definition involving a higher T-score threshold is more suitable for men [32]. The classification threshold used in our algorithm can be easily modified in light of any new findings on the best T-score threshold for male osteoporosis diagnosis.

Our method of estimating the bone mineral content of a lumber vertebra relies on the empirical HU to bone mineral density conversion formula by Rho et al. [11]. Using the conversion formula assumes that the CT imaging parameters and patient setup conditions are sufficiently similar. Beam hardening effects are also dependent on the imaging conditions and may further influence the formula's reliability [33]. In practice, we have found that the conversion formula achieves satisfactory performance and is sufficient for this preliminary work. The regression equation from $\mathrm{aBMD_{CT}}$ to $\mathrm{aBMD_{DXA}}$ implicitly accounts for these variant factors, given the same device and imaging setups. Though each imaging device has different beam properties, each scanner only has to be cross-calibrated once with respect to a DXA reference to obtain the proper regression constants specific to the machine. In future work, the HU to bone mineral density conversion model can be further improved to explicitly model the different setup and beam properties for better results.

$\mathrm{aBMD_{CT}}$ has an error of 8.8% when used to estimate $\mathrm{aBMD_{DXA}}$. Since $\mathrm{aBMD_{DXA}}$ has an inherent error of 5.3% [34], this additional error increases the uncertainty associated with diagnosis. Therefore, $\mathrm{aBMD_{CT}}$ is best used as an opportunistic screening measure, and detected cases of osteopenia should be confirmed using DXA.

We propose that $\mathrm{aBMD_{CT}}$ be used as supplementary indicator of bone mineral loss, in addition to the existing phantom-less QCT method. This equips clinicians with two sets of measurements, the volumetric BMD and the areal BMD for diagnostic use. There is also the possibility of incorporating other diagnostic measures into an integrated diagnostic suite for bone disease diagnosis; machine learning algorithms can subsequently be applied to the basket of diagnostic variables to mine the disease relations. An improved prediction model for diagnosis can also be built based on the set of additional features [35]. Further investigations (described in the next chapter) indicate that the aBMD$_{\mathrm{DXA}}$ estimation can be improved by selectively including multimodal features obtained from blood and hormone measurements.

2.5 Summary

In this chapter, we have introduced a new method of screening for low bone mass which can be applied with little additional cost to existing dCT setups. By modeling the DXA test for BMD and applying the model to dCT images, we obtained a aBMD$_{\mathrm{CT}}$ value that is analogous to aBMD$_{\mathrm{DXA}}$. The aBMD$_{\mathrm{CT}}$ was then correlated with aBMD$_{\mathrm{DXA}}$ using a robust regression technique to obtain a aBMD$_{\mathrm{CT}}$-to-aBMD$_{\mathrm{DXA}}$ mapping. There was a high correlation between aBMD computed from dCT and the true DXA-derived aBMD value. The results suggest that DXA-equivalent aBMD can be reliably estimated based on aBMD computed from dCT, and that aBMD$_{\mathrm{DXA}}$ can be used in an opportunistic screening system for osteopenia; the technique thus offers the potential to have significant preventative value in the early treatment and management of osteoporosis.

CHAPTER 3

Ensembles for Classification in Osteopenia Screening

In the previous chapter, we have described a scheme for modeling an aBMD_{CT} from CT images by applying a robust regression algorithm to correlate DXA and CT images. However, while aBMD_{CT} incorporates densitometric information, it does not consider the structural and morphological properties of bone when performing an osteopenia diagnosis. Instead of relying on expert knowledge to make a disease diagnosis, machine learning can be applied to develop a black-box model of osteporosis and osteopenia. The advantage of machine learning is that it can incorporate features and modalities that are difficult to quantify, such as the structural and morphological properties of bone.

This chapter presents a genetic algorithm (GA) to evolve a weighted decision ensemble for the diagnosis of osteopenia. The weighted decision ensemble uses a novel combiner function that is able to exploit classifiers that are discriminative towards specific classes. In addition, a GA scheme is used to optimize the weights of the decision ensemble such that the final ensemble has the greatest accuracy and class separation. These contributions allow for a more accurate diagnosis of osteopenia from CT scans of lumbar vertebrae.

3.1 Related Work

The robustness and accuracy of classification can be improved by combining multiple feature sets or classification methods into ensembles. Ensembles can be broadly divided into feature and decision level ensembles [36]. In decision level ensembles, the outputs of many different classifiers are aggregated to generate a final output. The ensemble methods implemented in this thesis are examples of decision level ensembles.

The most common method of creating a classifier ensemble is to combine the outputs of the individual classifiers by some form of weighted voting scheme. The most popular weighting scheme is the majority vote, also known as plurality vote [37]. In majority voting each classifier in the ensemble is equally important, and the ensemble decision is the mode of all the individual classifier decisions. A simple extension of majority voting is weighted majority voting, where each classifier may contribute a different number of votes [38]. Heuristics are often used to choose the weightings. In accuracy-based weighting, the ensemble weights are assigned based on the accuracy performance of the underlying classifiers, thereby allowing more accurate classifiers to have a larger influence on the ensemble outcome than less accurate classifiers [39]. A related idea is variance-based weighting, where the ensemble weights are inversely proportional to the variance of the underlying classifiers [40]. Therefore, variance-based weighting gives more emphasis to classifiers with a higher prediction confidence (low variance). A similar approach, variance-optimized bagging (vogging), optimizes a linear combination of the base-classifiers such that the ensemble variance is minimized while keeping accuracy above a predetermined value [41]. Statistical methods and information theory have also been applied to weight assignment schemes. Bayes rule can be used to assign weights in a weighted voting scheme, where the weight is the posterior probability of a classifier given the training set [40, 42]. In entropy weighting, each classifier is assigned a weight that is inversely proportional to the entropy of its classification vector, where classifiers that are particularly discriminative of specific classes have lower entropy and hence higher weightings [43, 44]. Another scheme uses the Dempster-Shafer theory of evidence to combine binary weighted decision trees [45]. Finally, density-based weighting is used in combination with feature-level ensembles and the weights are assigned to each classifier based on the sampling probabilities [40, 43].

A weakness with weighted voting schemes is that the base classifiers are assumed to be equally specific to all classes. Thus weighted voting ensembles may be unable to take advantage of classifiers that have high precision for specific classes. The proposed weighted decision ensemble addresses this weakness by allowing for classifiers to contribute different weights depending on the classifier output. Another weakness with weighted voting schemes is the assumption that the component classifiers are largely independent. Weighted voting schemes tend to employ all available classifiers and to emphasize classifiers that are individually good. However, the base classifiers may be highly correlated and better classification performance may be achieved using a large subset rather than a full subset of classifiers [46]. Furthermore, an ensemble of individually poor classifiers that collectively have complementary information may outperform an ensemble made of individually good classifiers that are mutually dependent. Hence, optimization schemes that search for good weightings are able to yield better ensembles. GAs are known as general purpose optimizers and are suitable for this task.

Recently, Mehmood et al. used GA to optimize a weighted majority ensemble consisting of five types of classifiers for solving gender recognition problems [38]. Majid et al. used genetic programming to evolve a composite of support vector machine (SVM) classifiers, each with different kernel functions [47]. Zhou et al. introduced GASEN (Genetic Algorithm based Selective ENsemble) [46] and GASEN-b(it) [48] which use GA to select participating neural networks for the ensemble. Aside from classifier selection, GAs have also been used to choose good feature sets for ensembles. Miller et al. used a GA to choose a binary subset of features [49] while Pei et al. used a GA to choose a weighted subset of features [50]. Lastly, both feature and decision level fusion occur in the GAECM method, where a GA was used to simultaneously select the feature sets and to optimize the classifier vote weightings in a classifier ensemble [51].

3.2 Ensemble Classification

We hypothesize that advanced machine learning techniques could be used to obtain a good classification of normal and pathological bone from normal CT images, and introduce a new technique for the classification of osteopenia from routine CT images. The Evolved Weighted Voting Ensemble (EWVE) and Evolved Weighted Decision Ensemble (EWDE) are proposed to achieve better classification performance on the osteopenia detection problem over multiple operating points. EWVE comprises three separate components (Fig. 3.1): the EWVE module, the base classifiers in the ensemble, and the features describing the input CT data. The osteopenia diagnosis system works by taking in a routine CT volume as the input. The feature extraction module transforms the 3-D CT volume into a set of features that are relevant for osteopenia diagnosis. The extracted

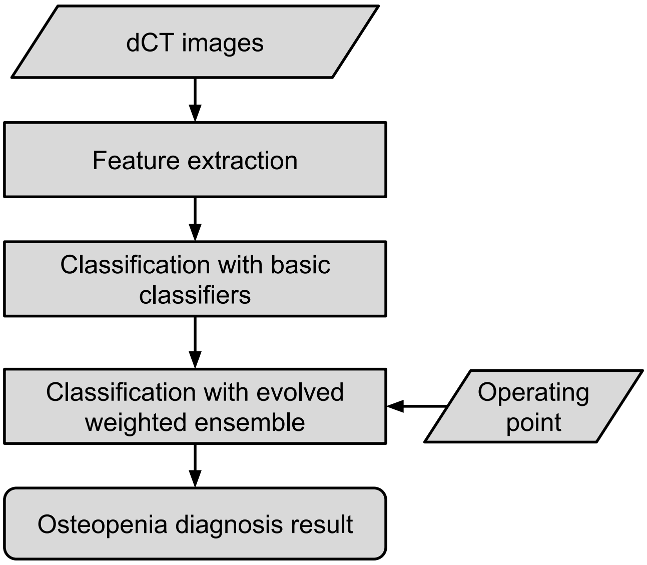

dCT images
Feature extraction
Classification with basic classifiers
Classification with evolved weighted ensemble
Operating point
Osteopenia diagnosis result

Figure 3.1: Flowchart of osteopenia screening algorithm.

features are then presented to a set of 18 basic classifiers that have been previously trained. The classifiers each return a binary result indicating whether osteopenia is detected; this is concatenated into an 18-bit pattern which is then passed to the EWVE module. The EWVE module, like the basic classifiers, has also been previously trained. The EWVE module determines the final diagnosis result by weighting the decisions of the 18 basic classifiers. The specificity of the screening system can be modified according to the clinician's preference by adjusting the operating point of the EWVE module.

3.2.1 Ensemble of Classifiers

In this work, two classifiers are applied to ensemble the classifier outputs. In addition, a genetic algorithm based ensemble method is also employed to ensemble the classifier outputs. By training classifiers on the ensemble outputs, each basic classifier acts as a binary feature transform. The four ensemble variants are described below:

The **Nearest Neighbor Ensemble** (NNE) is constructed by applying the nearest neighbor classifier onto the outputs of the 18 base classifiers. The binary classification outputs from the 18 base classifiers are concatenated to form an 18-bit pattern. The class label of an unseen test sample is assigned by first applying the 18 basic classifiers to generate the 18-bit test pattern. The Euclidean distance between the test pattern and the reference 18-bit patterns is used to find the nearest matching pattern, whose label is then applied to the test sample.

The **Random Forest Ensemble** (RFE) is constructed by applying the random forest algorithm onto the outputs of the 18 base classifiers. Random forest is a state-of-the-art classifier system that ensembles many bootstrapped decision trees. As with the nearest neighbor ensemble, the random forest ensemble is trained on the outputs of the 18 base classifiers instead of the feature data. An unseen test sample is first converted to an 18-bit pattern which is subsequently classified with a random forest classifier previously trained on the set of 18-bit patterns.

The **Evolved Weighted Voting Ensemble** (EWVE) is an extension of the most majority vote, also known as plurality vote [32]. In majority voting each classifier in the ensemble is equally important, and the ensemble decision is the mode of all the individual classifier decisions. The EWVE uses weighted majority voting, where each classifier may contribute a different number of votes [42]. The ensemble decision can be determined by a weighted vote, where the weights assigned to each classifier are assigned according to a GA. The ensemble decision is represented mathematically as:

$$\text{class}(x) = \text{argmax}(\sum_k s_k w_k g(y_k(x), c_i) - b), \tag{3.2.1}$$

where s_k is a binary switch indicating whether the corresponding classifier participates in the voting, w_k is the weight assigned to the $k-th$ classifier's decision, b is a biasing threshold to set the operating point of the system, and $g(y, c)$ is an indicator function representing the kth classifier's output, defined as:

$$g(y_k(x), c) = \begin{cases} 1, & \text{if } y = c. \\ \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.2}$$

The **Evolved Weighted Decision Ensemble** (EWDE) is proposed to further generalize the ensembler function. In the new combiner function, each classifier has a weight for each decision that it might generate.

For a binary classification problem each classifier has two weights. If a classifier returns a result of 1, it contributes the first weight towards the first decision, and nothing towards the second decision. If a classifier returns the result 2, nothing is contributed towards the first decision while the second weight is contributed towards the second decision. The ensemble decision is the decision with the greatest total sum. This combiner function can be expressed as follows:

$$class(x) = \underset{c_i \in dom(y)}{\arg\max}(\sum_k s_k h(y_k(x), c_i, k)), \tag{3.2.3}$$

where $h(y, c, k)$ is an indicator function with a weight $w_{k,c}$ for each combination of base classifier and classes:

$$h(y, c, k) = \begin{cases} w_{k,c} & y = c \\ 0 & y \neq c \end{cases}. \tag{3.2.4}$$

This combiner function allows classifiers to be more specific and discriminating towards particular classes and is helpful if the base classifiers have high precision. The weighted decision ensemble can also be easily modified for multi-class problems by appropriately changing the number of weights and decision units to match the number of target classes.

3.2.2 GA Ensemble Optimization

The weights of the Evolved Weighted Voting Ensemble are generated by using a genetic algorithm optimization technique. Choosing the weights for ensemble systems can be regarded as an optimization task. GAs are known as general purpose optimizers and are suitable for this task. However, traditional GA optimization methods have not addressed how classifier performance can be maintained or optimized across multiple operating points. Our GA approach uses a new evaluation measure that better addresses this problem. The evaluation function,

$$\text{fitness} = \text{accuracy} + k \times \text{geometricMean}, \tag{3.2.5}$$

comprises of two terms, accuracy and the geometric mean, scaled by the weighting parameter $k = 2$. Accuracy itself is not a good measure of classifier performance because it does not discriminate between true positives and true negatives. A high accuracy can be obtained simply by assigning all samples to the majority class, which is not useful for medical applications since this means that diseases and symptoms are not detected. The geometric mean is a better measure for problems with class imbalance as it takes a high value only when detection rates for both classes are high. The geometric mean is computed from the true positive rate (TPR) and the true negative rate (TNR) by

$$\text{geometricMean} = \sqrt{\text{TPR} \times \text{TNR}}. \tag{3.2.6}$$

The ensemble weights are optimized using an approach partly inspired by Xu and He [66]. Each solution is represented as an 18 gene chromosome. For an EWVE, each gene comprises of a binary bit S and a real

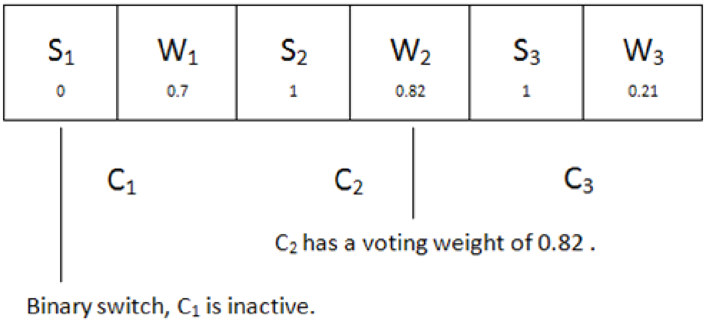

S1
0
W1
0.7
S2
1
W2
0.82
S3
1
W3
0.21
C1
C2
C3
C2 has a voting weight of 0.82 .
Binary switch, C1 is inactive.

Figure 3.2: A chromosome of an EWVE with 3 component classifiers.

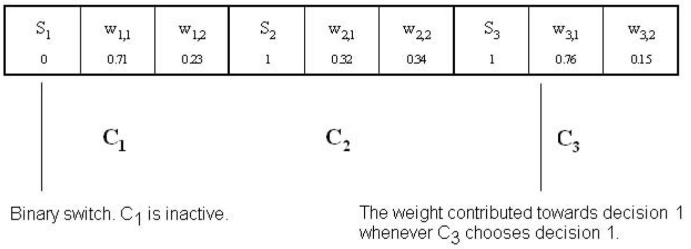

S1
0
w1,1
0.71
w1,2
0.23
S2
1
w2,1
0.32
w2,2
0.34
S3
1
w3,1
0.76
w3,2
0.15
C1
C2
C3
Binary switch. C1 is inactive.
The weight contributed towards decision 1 whenever C3 chooses decision 1.

Figure 3.3: A chromosome of an EWDE with 3 classifiers.

number W, representing respectively whether the corresponding classifier is used and the weight assigned to that classifier. A prototype chromosome with 3 component classifiers is shown in Fig. 3.2. For an EWDE, each gene comprises of a binary bit and 2 real numbers, where the binary bit indicates whether the corresponding classifier is used and the two real numbers represent the weights assigned for each class decision for that classifier, as shown in Fig. 3.3.

A pool of 250 chromosomes was used. For each generation, the top 50 chromosomes are retained and crossover ($p_c = 0.70$) populates each chromosome in the child pool by selecting random pairs of parents from the top

50 chromosomes and randomly replacing genes with a random weighted average of the corresponding parents' genes. Next, mutation ($p_m = 0.03$) modifies the solutions by randomly replacement with a random real value. Elitism preserves the top 5 individuals of the parent generation. Evolution ends when the best performance has not improved over the last 50 generations. The GA optimization scheme is illustrated in Fig. 3.4.

3.2.3 Basic Classifiers

Six different basic classifiers, each employed on each set of features, were used in our experiments. Each basic classifier was then separately trained on each of the three feature sets to yield eighteen different classifiers. The basic classifiers are:

1. K-nearest neighbor classifier
2. Naive Bayesian classifier, assuming a Gaussian data distribution
3. Naive Bayesian classifier, using kernel density estimation
4. Bayesian discriminant function, where each class shares the same covariance matrix
5. Support vector machine, with experimentally chosen polynomial kernel function $K(x_1, x_2) = (1 + x_1^T x_2)^2$ and soft margin penalty of 1.1

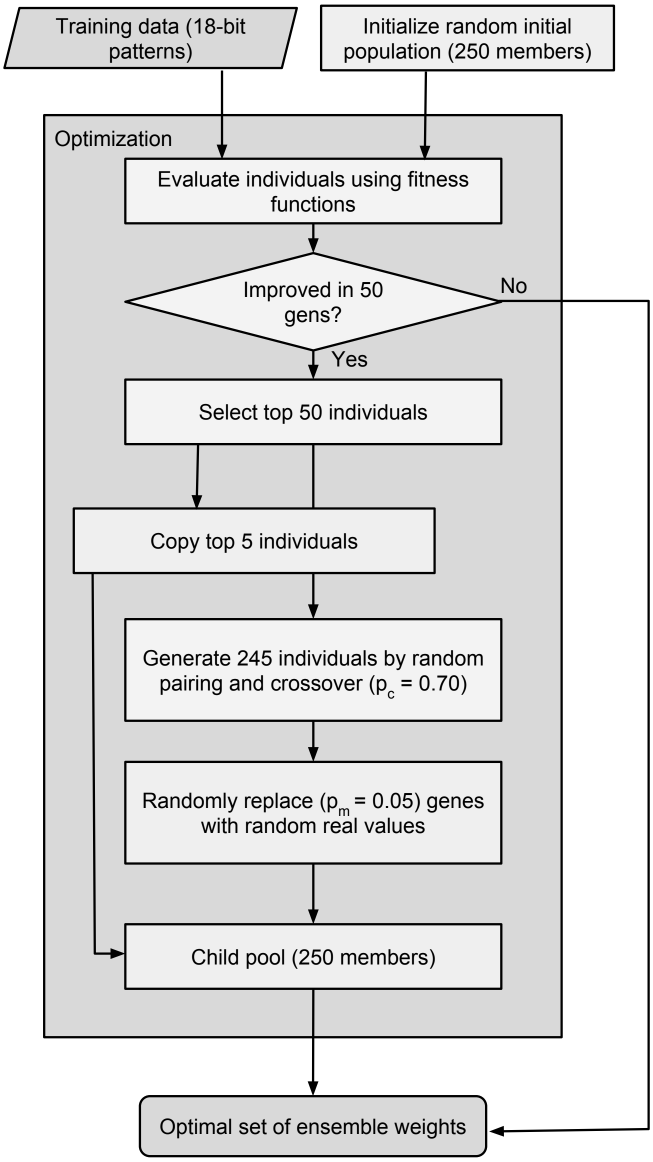

Training data (18-bit patterns)
Initialize random initial population (250 members)
Optimization
Evaluate individuals using fitness functions
Improved in 50 gens?
No
Yes
Select top 50 individuals
Copy top 5 individuals
Generate 245 individuals by random pairing and crossover (p_c = 0.70)
Randomly replace (p_m = 0.05) genes with random real values
Child pool (250 members)
Optimal set of ensemble weights

Figure 3.4: Flowchart of GA optimization.

6. Decision tree classifier, using Gini's diversity index as splitting criteria and cost complexity pruning

3.2.4 Feature Sets

The features employed are designed to be relevant to the bone mineral density and bone morphology. The distribution of gray levels is important because the CT gray levels are directly related to the bone mineral density (denser bone appears as brighter voxels in CT scans). Morphological features are also relevant in characterizing the mechanical properties of bone. These related features are used to distinguish between different classes of bone.

The first set of features is the *grey level histogram features.* The gray level distribution is quantized into a few histogram bins, with each bin containing the percentage of object voxels that fall into the bin range and each bin being mapped to a single feature dimension. The relevant range of bone gray levels is divided into 9 bins, each covering an approximately equal range of gray values. This set of features attempts to represent the gray level distributions of the CT volumes.

The second set of features is the *mean of threshold features.* The gray level distribution is divided into a number of overlapping threshold ranges. Each threshold range selects a different set of object voxels specific to that range of gray values. For each range of gray values, the mean gray level of the object voxels within that range is computed and represented as a feature. This set of features attempts to represent the statistics within the various gray level ranges in the CT volumes.

The third set of features is the *morphological features*. The average cross-section of the bone slices is first obtained by overlaying slices and thresholding. From the average cross-section, the area, height, and width of the cross-section are estimated while the perimeter of the cross-section and the minor-axis length are computed after applying morphological operations. Other morphological quantities were also explored but found to be non-relevant using feature selection algorithms [43, 47, 56] and hence excluded. These morphological features attempt to represent the size and shape of the bones being examined.

3.3 Results and Discussion

3.3.1 Experiment Methodology

The data sets employed in our experiments are drawn from a larger set consisting of CT volumes and matching DXA images and scores described in Section 2.4.1. Two smaller data sets are drawn non-exclusively from this source set, and each drawn data set is broken into two separate classes.

The first data set will be referred to as the TS-A (T-score A) data set, while the second data set will be referred to as the TS-B (T-score B) data set.

For the TS-A data set, the samples are drawn from the original data set based on the DXA-derived T-scores of the individual lumbar vertebrae. T-scores are related to the risk of bone fracture, and osteoporosis (T–score <-2.5) and osteopenia ($-2.5 <$ T–score < -1.0) are defined based on the T-score. Thus, a T-score of less than -1 standard deviation (SD) above reference was designated the lower threshold and a T-score of greater than 0 SD was designated the upper threshold. The samples are labeled as two classes correspondingly, an at-risk class (for samples with T-score below -1 SD) and a not-at-risk class (for samples above 0 SD). There are 103 samples in the TS-A data set, with 50 at-risk samples and 53 not-at-risk samples.

For the TS-B data set, the samples are also drawn from the original data set based on the T-scores of the individual lumbar vertebrae. The samples were divided into classes with T-scores of less than -1 SD and greater than -1 SD, which correspond to an at-risk class (T–score < -1.0) and a not-at-risk class (T–score > -1.0) respectively. The TS-B data set is more difficult than the TS-A data set because the separation between the two classes is smaller, and it is hard to distinguish between the boundary cases. There are 155 samples in the TS-B data set, with 50 at-risk samples and 105 not-at-risk samples.

Leave-one-out cross-validation (LOOCV) is employed to obtain the classification accuracy for each combination of features and classifiers [52]. LOOCV is performed by repeatedly training the classifier system on all-but-one of the available samples, then testing the trained classifier on the unseen sample. LOOCV ensures that each classifier is trained on the maximal number of training samples while using all available data for testing.

For the evaluation of the classifier ensembles, since GA is used to evolve the ensembles, it is too costly to employ LOOCV. Instead, 10-fold cross-validation is performed ten times for each ensemble. For each run the data set is broken into ten mutually exclusive subsets of equal size. The ensemble is tested on each subset while being trained on the union of all other subsets. To reduce the computational cost, the base classifiers are each run only once in a LOOCV fashion, and the classifier results stored in memory. Therefore, the combiner functions of the evolved ensembles operate only on the precomputed classifier results and it is not necessary to run new instances of the base classifiers.

Lastly, the effectiveness of the separation term in the GA evaluation functions was also investigated by conducting another set of trials with the separation term disabled (setting $k_2 = 0$). For all experiments hypothesis testing (t-test) was used to compare the evolved ensembles and the individual classifiers, as well as between the proposed weighted decision ensemble and other existing ensemble systems. All t-tests conducted were one-tailed at the 5% level of significance.

3.3.2 Results

Tables 3.1 and 3.2 show the classification accuracies of each individual classifier method on the various feature sets, and also the classification accuracies of the ensemble systems. For the TS-A data set, a high classification accuracy (>85%) was obtained for the best individual classifiers. However, all of the evolved ensemble classifiers significantly outperformed even the best individual classifiers, improving the classification accuracy by between 1.5% to 2.5%. For the TS-B data set, due to the increased difficulty of the data set, the best individual classifiers were only able to obtain a good classification accuracy (>75%). Only ensembles with more complex decision functions, such as the evolved weighted vote and evolved weighted decision ensembles, were able to significantly improve on the classification accuracy. On the TS-B data set the proposed evolved weighted decision ensemble gave the best classification result (83.48%) and was statistically better than all individual and ensemble classifiers.

Table 3.1: Classification accuracy on TS-A dataset

Classification Method	Feature Set	Accuracy
k-NN	Hist	86.41%
	MoR	**87.38%**
	Morpho	78.64%
Naïve Bayes (Gaussian)	Hist	80.58%
	MoR	**83.50%**
	Morpho	81.55%
Naïve Bayes (KDE)	Hist	**89.32%**
	MoR	85.43%
	Morpho	80.58%
Bayesian Discriminant Function	Hist	78.64%
	MoR	**80.58%**
	Morpho	78.64%
SVM	Hist	83.50%
	MoR	**87.38%**
	Morpho	83.50%
Decision Tree Classifier	Hist	**87.38%**
	MoR	85.44%
	Morpho	85.44%
Majority Vote of All Classifiers		88.35%
Evolved Majority Vote Ensemble (EMV)		**91.84%**
Evolved Weighted Vote Ensemble (EWV)		91.07%
Evolved Weighted Decision Ensemble (EWD)		91.65%

The results show that the proposed evolved weighted decision ensemble significantly improves on the performance of the best individual classifiers for both data sets. The proposed ensemble is significantly better than all individual and ensemble classifiers on the TS-B data set. This validates the effectiveness of the new combiner function in the weighted decision ensemble. Thus, the experimental results clearly demonstrate that the evolved weighted decision ensemble is the most suitable ensemble and classification method among all the methods studied here.

Table 3.3 shows the effect of the separation term on the classification accuracies of the evolved ensembles. Enabling the separation term generally improves the ensemble accuracy by between 0.5% to 2%. This improvement is statistically significant for the evolved weighted decision ensemble, and for the evolved weighted vote ensemble on the TS-B data set. This result demonstrates that including a separation term in the GA evaluation function is helpful as it allows the GA to discriminate between chromosomes that have the same accuracy but different class separations.

3.3.3 Discussion

All evolved classifier ensembles have statistically similar classification performances on the less difficult TS-A dataset, while the weighted decision ensemble significantly outperforms the other ensembles by about 2% on the more difficult TS-B dataset. The main difference between these ensembles is the combiner function, where the proposed method uses the most complex weighting scheme. This result suggests that the combiner model used in the evolved weighted decision ensemble is more general and

Table 3.2: Classification accuracy on TS-B dataset

Classification Method	Feature Set	Accuracy
k-NN	Hist	**81.29%**
	MoR	72.26%
	Morpho	77.42%
Naïve Bayes (Gaussian)	Hist	71.61%
	MoR	72.26%
	Morpho	**73.55%**
Naïve Bayes (KDE)	Hist	76.77%
	MoR	74.19%
	Morpho	**78.06%**
Bayesian Discriminant Function	Hist	**78.06%**
	MoR	72.90%
	Morpho	76.13%
SVM	Hist	**71.61%**
	MoR	69.68%
	Morpho	60.65%
Decision Tree Classifier	Hist	**77.42%**
	MoR	75.48%
	Morpho	74.84%
Majority Vote of All Classifiers		76.12%
Evolved Majority Vote Ensemble (EMV)		81.23%
Evolved Weighted Vote Ensemble (EWV)		81.68%
Evolved Weighted Decision Ensemble (EWD)		**83.48%**

thus potentially more powerful, but improvements in accuracy are only visible in more difficult problems where simpler models are insufficient. Furthermore, it may also imply that the evolved weighted decision ensemble requires comparatively more training samples to be fully trained. This

Table 3.3: Accuracy with and without separation term

Ensemble	W/O Separation Term		W/ Separation Term	
	TS-A	TS-B	TS-A	TS-B
EMV	91.36%	80.58%	91.84%	81.23%
EWV	91.46%	79.74%	91.07%	81.68%
EWD	90.78%	81.35%	91.65%	83.48%

hypothesis agrees with the experimental results on the separation term, where the separation term was found to have a statistically significant effect only for the more complex data sets and ensemble models.

To further investigate the merits of the evolved weighted decision ensemble, we also studied the ensemble weights of the final evolved weighted decision ensembles. Two observations were made based on the ensemble weights of the most highly evolved individuals. First, some active classifiers (classifiers selected for the decision ensemble) had class weights that were about equal while other active classifiers had class weights that were polarized. This observation demonstrates that the weighted decision ensembles incorporate both classifiers that are non-specific (classifiers with approximately equal class weights) and highly specific (polarized class weights). The second observation was that some of the active classifiers in the ensemble were those with poorer classification performance, which agrees with the theory that individually poor classifiers can provide a lot of complementary information in an ensemble.

The most costly step in the proposed algorithm is the training of the ensembles using GA, while the actual classification time is minimal as the classification step is computationally simple. For the proposed application of medical diagnosis, the computational cost of the training algorithm is usually not a major concern because the training of classifiers is performed prior to deployment of the diagnosis system. However, GA has a computational complexity of $O(mnp)$, where m is the chromosome length, n is the number of generations, and p is the size of the population pool. As a large number of evaluations have to be performed, GA is a costly method to optimize the classification algorithms and may not be suitable for problems with large data. This motivates the work in Chapter 4, which proposes less computationally expensive algorithms.

3.4 Summary

In this chapter, a decision ensemble was introduced to combine the outputs of multiple basic classifiers previously trained on a set of derived grey level and morphological features. The evolved weighted decision ensemble assigns a different voting weight to each classifier and output class, thus allowing the ensemble to incorporate classifiers that have high class specificity. A GA then optimizes the weights of the decision ensemble. The evolved weighted decision ensemble attained an accuracy of 91.65%

and 83.5% over the two data sets used and was significantly better than the best individual classifiers. The evolved weighted decision ensemble was also statistically better than other ensemble systems over the difficult data set. The results demonstrate that it is possible to identify patients at risk of low bone mass from routine CT scans with good accuracy by using advanced machine learning algorithms to model the disease condition.

CHAPTER 4

Ensembles for Regression in Osteopenia Screening

In clinical studies, besides the main modalities being studied, other medical measurements are often taken. For a radiological study, it is common to also take blood and hormone measurements for control purposes. These multimodal data are often left unstudied as they are not the focus of the investigation. However, there may be hidden relationships between the disease symptoms and these multimodal data. Although it is likely that any hidden relationships are weaker than the primary modality, there is potential for the primary relationship to be improved by exploiting the hidden information contained in multimodal data. In this chapter, we use blood, hormone, and physical measurements to improve the aBMD estimated from dCT. It is not feasible to solve the problem by directly applying multivariate regression, as the additional multimodal features are less informative. The increased ratio of features to training cases also introduces the problem of high relative dimensionality, which may lead to overfitting.

In this chapter, we study how ensemble regression methods can be applied to solve a regression problem on a multimodal medical dataset with high relative dimensionality. Based on insights obtained by using several feature selection and data transformation techniques with linear regression, an ensemble regression method using filtering is proposed. The filtering-based ensemble technique chooses a set of regressors from several candidate regressors such that the component regressors are diverse and uncorrelated. The proposed method generates the best results on the multimodal medical data and can be used to mine informative features.

4.1 Related Work

In clinical practice, DXA is a dedicated imaging modality that generates an aBMD score by which osteoporosis and osteopenia can be diagnosed [25]. To facilitate opportunistic bone screening, recent studies [9, 53] have tried to estimate an DXA-equivalent aBMD score using other imaging modalities that are commonly used in surgical planning or diagnosis. dCT is a promising modality for opportunistic screening as it is performed frequently and contains densitometric information correlated to BMD [6, 7]. However, while it is feasible to use dCT scans to estimate DXA-equivalent aBMD, several factors inherent to dCT imaging, such as beam hardening [33], can adversely affect the reliability of the estimation results. Radiological modalities may also require machine-wise calibration to account for differences in beam and source properties. One way to increase the robustness of aBMD estimation is to incorporate additional features to the prediction model [35]. These additional features can be diagnostic factors [54] that are unrelated and independent of dCT, or describe other aspects, such as the topological, morphological, and mechanical properties [55], of the dCT information. In this work, we generate two additional sets of features to improve the aBMD estimation. The first set of additional features describe the HU distributions and morphological features of the bone, and is drawn from dCT data. For the second set of features, we exploit the physical, blood, and hormone data that was also recorded during the clinical experiments. This second set of features provides a multimodal dataset that is independent of dCT, and may be helpful in increasing the robustness of the regression.

Machine learning is a popular approach for computer-aided diagnosis, and was previously used to diagnose fractures [19] and osteoporotic diseases [20, 21] based on QCT images. These methods are capable of achieving good detection rates, but typically involve the use of black boxes, which makes it difficult to evaluate their reliability and generality without more extensive clinical validation. Also, most classification algorithms return only an outcome value, or a bias value at best, which makes it difficult to estimate the severity of the diagnosed condition. Therefore, in this work, rather than focusing on the classification outcome of osteoporosis, we are interested in the aBMD value, from which the risk of osteoporosis is known based on previous studies [25].

One recurring problem with constructing diagnosis systems for medical applications is the lack of training data [56], which occurs because of the cost of acquiring patient data and the low prevalence rates of diseases [57, 58]. This lack of training data results in an undersampling of the problem space which tends to lead to poor classification performance [59, 60]. The problem is further compounded by the imbalanced nature of the class samples; typically the number of positive class instances (diseased cases) is much less than the number of negative class instances (normal cases) [61, 62]. Lastly, clinical data may have missing or incomplete features. These problems impair the performance of machine learning methods, but some ensemble techniques have been found to be robust to high dimensionality [63], high class imbalance [64], or missing features [65]. Ensemble methods are also known to improve the accuracy over single learners, and have been previously studied for use in medical diagnosis [66]. Ensemble methods work by combining the contributions of several weak component learners, which reduces the variance of errors.

4.2 Ensemble Regression

Ensemble methods can be applied to regression problems to obtain better robustness and accuracy. In this section, we describe the bootstrap aggregating method before introducing a feature-wise modification which is more helpful for datasets with high relative dimensionality. Building upon the bootstrap aggregating approach, the use of metalearners for improving ensemble performance is discussed. We review two basic metalearner ensembling schemes before presenting our correlation-based filtering technique for metalearner ensembling. The new technique is designed to form ensembles that are both diverse and robust.

Let the DXA-derived aBMD values be denoted as the target variable matrix Y. The data matrix X is then obtained by feature-wise concatenation of the dCT-derived aBMD values, the dCT-derived HU features, and the additional multimodal features from blood and physical measurements. The regression problem is defined as regressing the target Y based on the data X such that unknown future samples can be predicted.

4.2.1 Bootstrap Aggregating.

Bootstrap aggregating [67], also known as bagging, may be capable of overcoming the high dimensionality of the data relative to the number of training samples. Bagging can improve classification/regression accuracy and stability, and any learning model may be used with bagging. In this work, several linear regression models are bagged to form a regression ensemble.

In bagging, the ensemble is composed of several component classifiers, each of which is trained on a different subset of the training data, and the ensemble decision is obtained by taking an average of the individual ensemble regressors. The subsets are randomly drawn with resampling from the training set, and the subsets are traditionally drawn in a case-wise fashion. In **case-wise bagging**, each ensemble component is trained on a different resampled training set. The resampled training sets are formed by randomly drawing training cases with resampling. To reduce large instabilities in the regression and to better constrain the regression, the resampled training sets are resampled to contain more cases than there are features. For an input training set consisting of n data and target pairs $\{x_i, y_i\}$, where $i = 1 : n$, the case-wise bagging algorithm for a k-component ensemble with a case over-sampling factor of s_c is described in Algorithm 1.

While case-wise bagging is frequently used, we propose the use of feature-wise bagging as an alternative approach for bagging. **Feature-wise bagging** trains each ensemble component on a different subset of

Algorithm 1 Case-wise Bagging

```
procedure CB_Train($X, Y, k, s_c$)
    for $j = 1 : k$ do
        $S_j \leftarrow \{\emptyset\}$
        while $numel(S_j) < (s_c \times n)$ do
            $randNo \leftarrow rand(1 : n)$
            $S_{temp} \leftarrow \{x_{randNo}, y_{randNo}\}$
            $S_j \leftarrow \{S_j, S_{temp}\}$
        end while
        $R_j(x) = LR(S_j)$                    ▷ Linear regression of $S_j$
    end for
    return $R$
end procedure

procedure CB_Test(R, $x_{test}$)
    return $\sum_j R_j(x_{test})/k$
end procedure
```

training features, and the subset of training features are formed by randomly selecting features for inclusion. This reduces the dimensionality of the data relative to the number of available training cases, and is better suited for datasets with a high relative dimensionality. For an input training set consisting of n d-dimensional data and target pairs $\{x_i, y_i\}$, where $i = 1 : n$ and $x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,d}\}$, the feature-wise bagging algorithm for a k-component ensemble with a feature sampling factor of s_f is given by:

Algorithm 2 Feature-wise Bagging

```
procedure FB_Train($X, Y, k, s_f$)
    for $j = 1 : k$ do
        $dim \leftarrow \{1, 2, ..., d\}$
        while $numel(dim) > (s_f \times d)$ do
            $randNo = rand(1 : numel(dim))$
            $dim(randNo) \leftarrow \{\emptyset\}$
        end while
        $S_j \leftarrow \{x_{dim}, y\}$
        $R_j(x) = LR(S_j)$                    ▷ Linear regression of $S_j$
    end for
    return $R$
end procedure

procedure FB_Test(R, $x_{test}$)
    return $\sum_j R_j(x_{test})/k$
end procedure
```

Bootstrap aggregating may dampen large instabilities in the regression when the resampled training sets are resampled to contain more cases than there are features, or when the resampled subsets contain small data dimensionality relative to the number of available training cases.

4.2.2 Metalearner Ensembles

Instead of taking the average of the component regressors, the ensemble components can also be combined by using a metalearner. A metalearner is typically a machine-learner that is capable of learning the properties of the model and assigning the appropriate weightings to the component regressors. The metalearner uses the outputs of the component regressors on the training data as the inputs, and takes the target aBMDs as the

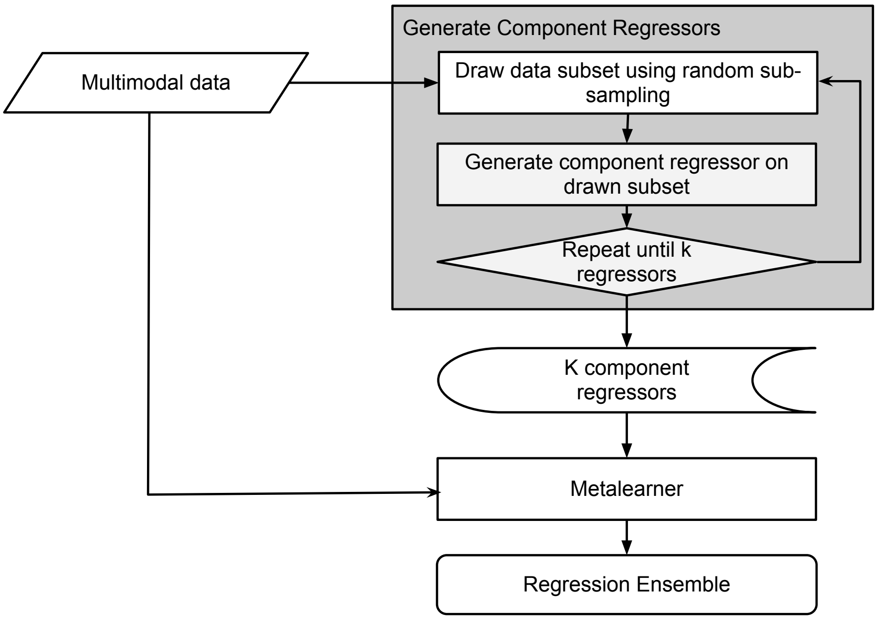

Generate Component Regressors
Multimodal data
Draw data subset using random sub-sampling
Generate component regressor on drawn subset
Repeat until k regressors
K component regressors
Metalearner
Regression Ensemble

Figure 4.1: Overview of the generation of a metalearner regression ensemble.

desired outputs. The metalearner then learns the model most capable of matching the regressor outputs to the target. Metalearners may be considered as a separate classification or regression problem on the regressor outputs. The metalearner training process is given in Fig. 4.1. A few metalearner candidates are explored here:

Regression weighted metalearner. The regressor outputs are mapped to the target output by a regularized regression. Each regressor is assigned a weight, and the ensemble decision is the weighted sum of the regressor outputs. The weighting is assigned to a higher level regressor based on the errors committed by each component error on the training set. For a set of k candidate component regressors $\{R_1(x), R_2(x), ..., R_k(x)\}$, the ensembling algorithm is given in Algorithm 3.

Algorithm 3 Regression Weighted Metalearner

```
procedure RWM_Train($X, Y, R$)
    for $i = 1 : k$ do
        $p_i \leftarrow R_i(X)$
    end for
    $P \leftarrow \{p_1, p_2, ..., p_k\}$
    $W \leftarrow LR(P, Y)$                    ▷ Linear regression of P on Y
    return $W$
end procedure

procedure RWM_Test($R, W, x_{test}$)
    $R(x_{test}) \leftarrow \{R_1(x_{test}, R_2(x_{test}), ..., R_k(x_{test})\}$
    return $WR(x_{test})$
end procedure
```

Multi-layer perceptron metalearner. A perceptron network is trained on the regressor outputs to match the aBMD$_{\text{DXA}}$. For a set of k candidate component regressors $\{R_1(x), R_2(x), ..., R_k(x)\}$, the ensembling algorithm is given by Algorithm 4.

Correlation-based filtering. We propose a simple technique for constructing diverse and robust regression ensembles. The outputs of each component regressor are compared to outputs of each other component regressor, and the sum of the correlation coefficients is computed. The component regressors that generate the least correlated outputs are selected to ensure diversity in the ensemble. The ensemble result is the mean of the component regressor outputs. For a set of k candidate component

Algorithm 4 Multi-layer Perceptron Metalearner

```
procedure MLPM_Train($X, Y, R$)
    for $i = 1 : k$ do
        $p_i \leftarrow R_i(X)$
    end for
    $P \leftarrow \{p_1, p_2, ..., p_k\}$
    $Nnet(P) \leftarrow Nnet_{train}(P, Y)$        ▷ train neural network for P on Y
    return $Nnet$
end procedure

procedure MLPM_Test($R, NNet(P), x_{test}$)
    $R(x_{test}) \leftarrow \{R_1(x_{test}, R_2(x_{test}), ..., R_k(x_{test})\}$
    return $Nnet(R(x_{test}))$
end procedure
```

regressors $\{R_1(x), R_2(x), ..., R_k(x)\}$ where l components are chosen, the ensembling algorithm is given by Algorithm 5.

Algorithm 5 Correlation-based Filtering

```
procedure CF_Train($X, Y, R, l$)
    for $i = 1 : k$ do
        $p_i \leftarrow R_i(X)$
    end for
    $P \leftarrow \{p_1, p_2, ..., p_k\}$
    $C \leftarrow Filter(X, Y, P, l)$    ▷ C contains the chosen regressor indices
    $R_C(x) \leftarrow \{R_{C(1)}(x), R_{C(2)}(x), ..., R_{C(l)}(x)\}$
    return $R_c(x)$
end procedure

procedure CF_Test($R_C(x), x_{test}$)
    return $R_C(x_{test})$
end procedure
```

The correlation-based filtering method can be performed with several different filtering schemes. We propose three different strategies for filtering component regressors. These strategies are aimed at building ensembles with high diversity.

Strategies for Filtering Component Regressors

Standard deviation ranking assigns a score to each component regressor based on the standard deviation of each regressor's training set prediction error. The standard deviation of prediction error determines the stability and consistency of a component regressor, and good regressors have low error standard deviation. For the training set targets Y and the component regressor predictions P, the l regressors are selected using the following algorithm:

Algorithm 6 Standard Deviation Ranking

```
procedure $SD\_Filter(X, Y, P, l)$
    for $i = 1 : k$ do
        $s_i \leftarrow stdev(Y - P_i)$    ▷ Standard deviation of prediction errors
    end for
    $s \leftarrow \{s_1, s_2, ..., s_k\}$
    $C = argmin(s, l)$    ▷ Indices of $l$ lowest items in $s$
    return $C$
end procedure
```

Stepwise partial correlation iteratively selects component regressors based on the partial correlation factor of each regressor's predictions. Partial correlation measures the correlation between two variables after controlling for a given set of variables, and is more useful if several variables are inter-related. Let $\rho_{AB \cdot C}$ denote the partial correlation coefficient between variables A and B while controlling for variable C, and let the training set targets be Y and the component regressor predictions be P.

The stepwise partial correlation is described in Algorithm 7.

Algorithm 7 Stepwise Partial Correlation

```
**procedure** $PC\_Filter(X, Y, P, l)$
    $NC \leftarrow \{1, 2, ..., k\}$                         ▷ Not chosen ensemble elements
    $C \leftarrow \{\emptyset\}$                                  ▷ Chosen ensemble elements
    $temp \leftarrow argmax(corr(Y, P))$      ▷ Index of regressor whose predictions are most correlated to $Y$
    $C \leftarrow temp$
    $NC \leftarrow temp \notin NC$                          ▷ Delete $temp$ from $NC$
    **for** $i = 1 : l$ **do**
        **for** $j = 1 : numel(NC)$ **do**
            Update $\rho_{P_{NC(j)}Y \cdot P_C}$
        **end for**
        $temp \leftarrow argmax(\rho_{P_{NC}Y \cdot P_C})$
        $C \leftarrow \{C, temp\}$
        $NC \leftarrow temp \notin NC$                      ▷ Delete $temp$ from $NC$
    **end for**
    **return** $C$
**end procedure**
```

This method of constructing the ensemble selects the most informative regressors while controling for the effect of previously selected regressors.

Stepwise least correlation iteratively selects the component regressor whose predictions are the least correlated with all the remaining regressor predictions, in order to build ensembles comprising of diverse regressors. For the component regressor predictions $P = \{p_1, p_2, ..., p_k\}$, the filtering algorithm is described by Algorithm 8.

Algorithm 8 Stepwise Least Correlation

```
procedure $LC\_Filter(X, Y, P, l)$
    $C \leftarrow \{\emptyset\}$
    for $u = 1 : k$ do
        for $v = 2 : k$ do
            $r_{u,v} = corr(P_u, P_v)$
        end for
        $s_u = \sum_v r_{u,v}$
    end for
    for $i = 1 : l$ do
        $temp \leftarrow argmin(s)$
        $C \leftarrow \{C, temp\}$
        for $u = 1 : k$ do
            $s_u \leftarrow s_u - r_{u,temp}$                    ▷ Update scores
        end for
        $s_{temp} \leftarrow max(s)$                          ▷ Set dummy value
    end for
    return $C$
end procedure
```

4.2.3 Time Complexity Analysis

The time complexity of the proposed metalearner ensembles and filtering-based strategies can be expressed in terms on the number of component regressors L, the number of training samples N, and the number of data features C. The time complexity of simple regression is dominated by matrix multiplication, $O(C^2N)$. The time complexity of the proposed metalearner ensembles can be derived as:

1. $O(LC^2N + N^2L)$ for the regression weighted metalearner

2. $O(LC^2N+L)$ for correlation-based filtering using standard deviation

3. $O(LC^2N + L^3)$ for correlation-based filtering using stepwise partial correlation

4. $O(LC^2N + L^2)$ for correlation-based filtering using stepwise least correlation

For the multi-layer perceptron metalearner, the time complexity is a non-linear function of the neural network configuration (number of nodes, network structure, activation function); generally, the multi-layer perceptron is slower than the metalearner algorithms above.

4.3 Experiments

4.3.1 Data

Our experiment data set consists of paired CT scans and DXA measurements. Patients with preexisting medical conditions were excluded from the study, while compression fractures and other degenerative pathologies were also excluded after a radiologist's review. The data set was divided into 155 pairs of CT volumes and DXA measurements, with each pair containing one of the vertebrae in the lumbar spine (L_1-L_4). Approximately two-thirds (100) of the samples had normal bone mineral density, while the remaining samples were osteopenic (46) or osteoporotic (4). The detector and scanner parameters for the study were previously described in

Section 2.4.1.

Three distinct feature sets were obtained, of which two were derived from information contained within the CT scans. The final feature set consists of physical and blood tests that were taken during the study as controls.

1. **aBMD$_{\mathrm{CT}}$**. The aBMD$_{\mathrm{CT}}$ is a good approximation of the actual aBMD$_{\mathrm{DXA}}$ value. aBMD$_{\mathrm{CT}}$ is estimated based on CT scans of the lumbar spine using the method presented in Chapter 2.

2. **CT image features**. These features are derived from the CT scans, and include histogram features and morphological features. The features are described in Section 3.2.4.

3. **Physical and blood measurements**. This feature set consists of physical and blood tests that were taken during the study as controls. Although physical and blood measurements are not expected to have strong predictive capability on the bone state, the additional information provided may be helpful in improving regression results.

For the regression task, the target output variable was the DXA measurements. For classification, aBMD$_{\mathrm{DXA}}$ was converted into age-calibrated T-score values and categorized into normal, and osteopenic and osteoporotic bone. The positive class was the osteopenic and osteoporotic cases.

10-fold cross-validation was performed 20 times each. Regression performance was measured by the root-mean-square error (RMSE). The regressed aBMD values were then used to diagnose osteopenia. Area under the receiver operating characteristic curve (AUROC) was used as the evaluation metric for classification.

4.3.2 Experiments

The following five experiments were conducted:

1. **Full linear regression based on aBMD, CT features, and multimodal data.** The three sets of data features were concatenated in various combinations, and linear least squares was used to regress the data to aBMD_{DXA}. The prediction and classification performance was measured. This experiment is performed to determine if combining the multimodal feature sets improves the regression and classification performance, and to see which combinations are most promising.

2. **Feature selection and data transformation on combined multimodal data.** The three sets of data features are combined into a single large multimodal data set for subsequent experiments. The combined dataset was subjected to several linear regression schemes (described in Appendix C). Feature selection and data transforma-

tion schemes were also tested here. This experiment studies whether feature selection or data transformation strategies are sufficient to improve regression performance.

3. **Ensembles by bootstrap aggregating.** Using the combined multimodal data set, bootstrap aggregating is applied. A random forest algorithm is used for comparison. This experiment compares case-wise bagging with feature-wise bagging to determine the most appropriate ensembling approach.

4. **Ensemble metalearners.** The three metalearner algorithms described in Section 4.2.2 are trained based on the outputs of the component regressors on the training data. Regression adaboost is used as a benchmark for comparison. The RMSE and AUROC are recorded to determine the most suitable metalearner for ensemble regression.

5. **Most significant features**. The regressor ensembles are used to determine the most helpful features. The selected regressors in the ensemble are averaged to form a single regression equation. From the composite regression equation, the features corresponding to the regression components with the highest magnitudes are recorded as the most significant features. The most significant features for each

Table 4.1: Regression on different combinations of multimodal features

CT	aBMD	Physical	RMSE	AUROC
x	x	x	0.0684	0.935
x	x		0.0836	0.918
	x	x	0.0806	0.933
x		x	0.0675	0.934
x			0.0900	0.903
	x		0.0979	0.871
		x	0.1066	0.873

fold and trial are accumulated to calculate the probability of a feature being a most significant feature in a regression ensemble. This provides insight into the most relevant and important features for aBMD regression.

4.4 Results and Discussion

In this section, we present our experimental results for the linear regression and ensemble regression methods.

4.4.1 Linear Regression on Different Combinations of Multimodal Features

Multivariate linear regression (Appendix C) was applied to various combinations of the three feature sets, and the results are presented in Table

Table 4.2: Evaluation of linear regression methods

Method	RMSE	AUROC
Linear least squares	0.0684	0.935
Linear least squares (Tikhonv regularization)	0.0664	0.944
Linear least squares (discard minor components)	0.0631	0.945
Principal feature analysis	0.0617	0.931
Principal components regression	0.0648	0.937
Partial least squares regression	0.0649	0.937

4.1. Comparing the individual sets of features, the CT features provide the best regression and classification performance, while the basket of physical and blood measurements provides the least information for the $\mathrm{aBMD}_{\mathrm{DXA}}$ estimation. $\mathrm{aBMD}_{\mathrm{CT}}$ was the single best feature. The results show that including additional features for regression significantly improves the estimation of $\mathrm{aBMD}_{\mathrm{DXA}}$, even when the dimensionality of the combined multimodal data approaches the number of samples.

4.4.2 Simple Feature Selection on Combined Multimodal Data

Table 4.2 presents the results of the regression and feature selection schemes. Principal feature analysis was found to produce the best regression result, but at the same time it had degraded classification performance. Simple feature selection by discarding the features with the smallest contributions was competitive with regression methods that transform

Table 4.3: Evaluation of ensemble methods

Ensemble method	Parameters	Regressors	RMSE	AUROC
Random forest	-	50	0.0683	0.926
Random forest	-	500	0.0666	0.931
Case-wise bagging	300% samping	10	0.0700	0.935
Case-wise bagging	300% samping	25	0.0694	0.934
Case-wise bagging	300% samping	50	0.0692	0.934
Feature-wise bagging	50% features	10	0.0608	0.944
Feature-wise bagging	50% features	25	0.0601	0.945
Feature-wise bagging	50% features	50	0.0599	0.944
Case+Feature-wise bagging	50% features, 300% sampling	25	0.0605	0.940

the data without performing feature selection.

4.4.3 Ensembles by Bootstrap Aggregating

Several ensemble methods were applied to the combined multimodal dataset, and the results are shown in Table 4.3. Feature-wise bagging was found to greatly improve the regression of aBMD_{DXA}, while case-wise bagging was ineffective. Changing the number of component regressors does not improve case-wise bagging over regularized linear least squares. Using both feature-wise and case-wise bagging was better than regularized linear least squares, but the results were still inferior to using feature-wise bagging alone.

Table 4.4: Evaluation of ensemble metalearning algorithms

Metalearner	Parameters	Regressors	RMSE	AUROC
Feature-wise bagging	50% features	25	0.0599	0.944
Adaboost	50 iterations	-	0.0700	0.937
Regression weighted	-	20	0.0627	0.941
MLP	1 layer, 10 nodes	10	0.0628	0.944
MLP	1 layer, 20 nodes	10	0.0625	0.944
MLP	1 layer, 50 nodes	10	0.0606	0.947
MLP	1 layer, 10 nodes	25	0.0628	0.943
MLP	1 layer, 20 nodes	25	0.0642	0.943
Standard deviation	c=10	25	0.0595	0.948
Partial correlation	c=10	25	0.0596	0.946
Stepwise least correlation	c=10	25	0.0590	0.946

4.4.4 Ensemble Metalearners

Table 4.4 presents the results of metalearner ensembles on the combined multimodal dataset. All the metalearner regression ensembles outperformed regularized linear least squares. The best regression result was obtained by the stepwise least correlation method, where an improvement of 11.3% and 1.50% RMSE over regularized linear least squares and feature-wise bagging respectively was observed. There was only a marginal improvement in AUROC, as the classification error was already low. Adaboost using regression trees performed poorly, producing results that were worse than linear least squares.

4.4.5 Most Significant Features

The most significant features in each regression ensemble were collected and to estimate the probability of a feature being a most significant feature. Table 4.5 lists the top features identified as most significant features. The column "correlation" indicates the sign that is most often assigned to the regression weight for that significant feature, and thus can be used to determine if the feature is positively or negatively linked with the target variable.

Table 4.5: Features identified as most significant features

Feature Name	Selection Frequency	Std Dev	Correlation
% voxels in HU range [100,200]	0.721	0.119	-
% voxels in HU range [600,750]	0.511	0.257	+
% voxels in HU range [750, 1000]	0.639	0.260	+
Mean HU of voxels in HU range [50, 1350]	0.812	0.159	+
Vertebral area	0.761	0.239	-
Vertebral minor axis length	0.718	0.262	-
Mean left femoral neck fat	0.811	0.164	-
Mean left femoral neck trabecular	0.596	0.313	+
Mean left femoral neck unmineralized	0.873	0.143	+
Mean right femoral trochanter fat	0.853	0.173	-
Mean right femoral neck cortical	0.623	0.321	-
Mean right femoral neck unmineralized	0.553	0.111	+
Insulin-like growth factor 1 (IGF-1)	0.623	0.209	+
Serum type 1 N-terminal procollagen (P1NP)	0.593	0.198	-
Homeostasis model assessment insulin resistance (HOMA-IR)	0.811	0.174	+
Model spheroidal high-density lipoprotein (ms_hdl)	0.530	0.199	+
$\mathrm{aBMD}_{\mathrm{CT}}$	0.224	0.203	+

4.4.6 Discussion

The results in Sec. 4.4.1 show that there is significant redundancy between the features in the combined multimodal dataset; for example, adding the aBMD feature to the CT and physical measurements does not improve RMSE. This suggests that several features are not helpful, and removing some features may improve overall regression performance. Removing spurious features also helps to reduce the data dimensionality and prevents overfitting. In Sec. 4.4.2, discarding features with the lowest contributions is more effective than data transformation methods that do not discard any features. This suggests that the main source of noise is not noisy samples, but spurious features.

Case-wise bagging was found to be ineffective in Sec. 4.4.3, whereas feature-wise bagging improved the regression performance. Apart from the possibility that spurious features are more significant than noisy samples in this multimodal dataset, another explanation involves the diversity and stability of the component regressors. Case-wise bagging is typically performed using unstable classifiers/regressors where minor changes in the training subset result in significant changes in the classification/regression, hence the ensemble be relatively diverse. In our case, the component regressor was a linear regression, which is a highly stable regressor. As suggested in [68], feature-wise bagging is more suitable for stable classifiers/regressors, and can be used to generate diverse ensembles.

The benchmark random forest algorithm was also outperformed by regularized least squares, implying that tree-based regression methods are not suitable for the problem which is better modeled by linear methods. Similarly, adaboost using regression trees was inferior to linear least squares, reinforcing the conclusion that regression trees are unsuitable for the multivariate regression problem. Tree-based methods may be overfitting the dataset due to the high relative dimensionality.

In general, reducing the regression error also reduces the classification error. This relation can be seen by correlating the RMSE and AUROC; a correlation coefficient of r = -0.930 was found. However, this does not imply that reducing the RMSE always improves the classification performance. A few algorithms were able to achieve comparable or superior classification performance while having larger regression error. This difference could lie in the regions where the regression algorithms are optimal over. For example, it is possible to improve the regression error by training on extreme samples, but this results in very little improvement in classification error as these samples are far from the decision boundary and are unlikely to be misclassified in the first place. One possible way to overcome this issue is to build an additional regressor on the region near the decision boundary. Reducing the regression error on this restricted region should be more effective in reducing the classification error. However, doing so may result in fewer training samples for the regression and training, which may have a negative impact on performance. It is possible to simply use a weighted linear least squares procedure, where the central samples are given more weight than extreme samples. The modified regression equation for such a weighted least squares procedure, given a diagonal sample weight matrix W, is

$$K = (X^T W X)^{-1} X^T W Y. \tag{4.4.1}$$

The sample weight matrix can be assigned based on a number of ad-hoc strategies. One method is to use the regression ensemble to produce an initial prediction, and to use the second set of regression constants if the prediction falls within a certain distance of the decision boundary.

One of the limitations of supervised learning is their black box manner of operation. The method of using regression ensembles for feature filtering presents a simplified list of significant features to clinicians, thus explaining the rationale behind the ensemble decision and helps to build expert knowledge. The most significant features selected (Sec. 4.4.5) may indicate a hidden relationship between the features and osteopenia. Among the selected CT features, the percentage of voxels belonging to lower density bone (from 600 to 1000 HU) and the mean HU of the non-soft tissue regions was found to be important in determining BMD. The area and minor axis length of the lumbar vertebra was also found to be relevant features, but had a negative impact on BMD. For the hormonal measurements, insulin-like growth factor 1 (IGF-1) and serum type 1 N-terminal procollagen (P1NP) were useful. No physical measurement was found to be a significant feature, which means that the height, weight and body mass index are not useful in determining osteopenia. In medical literature, P1NP is a biochemical marker that reflects osteoblast activity and is linked to increased rates of bone turnover [69]. P1NP was negatively related to BMD, supporting the literature. There is some evidence to suggest that IGF-1 concentration is reduced in osteoporotic patients [70, 71], which agrees with our results. Interestingly, aBMD_{CT}, which is the single feature with the highest correlation to aBMD_{DXA}, was only ranked 34 (out of 143) in the list of most significant features. This could mean that significant redundancies are present between aBMD feature and other features; aBMD could be strongly correlated with other features.

4.5 Summary

We have described a filtering-based ensemble method for performing multivariate regression on multimodal medical data. Several feature-wise data subsets are randomly selected to form a set of candidate regressors. The regression predictions of each candidate regressor are then compared to the outputs of each other candidate regressor to select a set of candidate regressors that are least correlated and most diverse. The chosen regressors are combined into an ensemble regressor to generate an ensemble regression prediction. The proposed method generates the best results on the multimodal medical data, increasing the accuracy and robustness of regression. The filtering approach can also be used to identify potential relationships between features and the target variable by analyzing the frequency at which a feature is selected in the component regressors.

CHAPTER 5

Clustering for Transfer Function Design in Medical Image Visualization

In medical image visualization, image understanding can be used to extract the underlying structures contained within volumetric data so that the extracted structures can be individually displayed or highlighted. Clustering is a class of unsupervised learning techniques that is used to identify and group similar elements, and is particularly useful if the properties and distributions of the data are unknown. In this chapter, a non-parametric clustering technique is applied to extract the material boundaries within volumetric data so that each distinct boundary can be visualized.

Volume rendering is a powerful tool for displaying 3-D medical data, as it provides a spatial perspective that is absent in 2-D slice views. In volume rendering, transfer functions (TF) are often used to assign optical properties to various voxel data properties. While a good TF can reveal important structures in the data, the process is not trivial for complex volumes composed of several materials and structures. Furthermore, it is not possible to entirely automate the process of TF design as the desired visualization result is dependent on the user's visualization objectives. Clustering is useful for TF design in volume rendering, as clustering can be used to extract the underlying structures in volumetric data and to present the extracted structure for automatic or user-assisted TF design.

The method presented in this chapter applies a non-parametric clustering technique on LH space [72] to organize the voxels into several groups, each representing a material boundary in volumetric data. Each material boundary can then be assigned visual properties using an automated TF design module, and occlusions within the volume are reduced by a data-driven post-processing step that considers the spatial distributions of each boundary. Manual manipulation of the visualization results can be easily achieved by modifying the clustering parameter, or by editing the cluster boundaries in LH space. The proposed innovations significantly reduce the time and effort required to obtain good TFs for volume rendering and enable visualizations with quality approaching that of existing methods to be automatically generated.

5.1 Related Work

TFs are mapping functions that assign various optical properties such as opacity and color to voxels depending on the voxel properties. Typically, a voxel's value and gradient magnitude are used in 2-D transfer functions for visualizing structures within volumes. Kindlmann et al. [73] used the first derivative (i.e., gradient) as an attribute to generate multi-dimensional TFs. In the 2-D TF domain, which incorporates the intensity and gradient magnitude, material boundaries can be interpreted as arches. Thus, they can be selected and visualized by manipulating certain TF widgets to approximate the arches. However, these arches often overlap, which prevents proper isolation of a material from others. One possible approach to overcome this drawback is to include the second directional derivative along with the gradient direction [74, 75]. Nevertheless, these methods cannot fully solve the blur effect in the intensity-derivative histogram which is caused by noise. Lum et al. [76] used the two intensity values on both sides of the border to set up a TF with the assumption that the width of the border represented by the distance between these two sample positions varies with the amount of blur in the volume. Šereda et al. [72] proposed another method to represent boundaries by searching for low and high intensity values in both the negative and positive gradient directions of the voxels in a boundary. The representation of those low and high values in a 2-D plane is called the *LH histogram*. An important advantage of LH histograms over the 2D intensity-gradient magnitude TF is that boundaries appear as blobs rather than arches. Blobs are easier to parameterize for clustering and are less likely to overlap in complicated datasets than arches; thus LH histograms allow for boundaries to be more easily separated either manually or automatically through clustering. Another advantage is that LH histograms have greater robustness to noise, bias and partial volume effects than intensity-gradient magnitude histograms. Recently, a semi-automatic generation of LH TFs using a fast generation of LH values has been introduced by Praßni et al. [77].

Apart from finding new TF feature domains, much work has also been put into developing clustering or segmentation algorithms to separate different regions in the TF domain. Tzeng et al. [78] presented a method to create TFs based on material classes extracted from the spatial domain using the ISODATA technique. Šereda et al. [79] applied hierarchical clustering to LH space to group voxels based on their LH values. Maciejewski et al. [80] used non-parametric kernel density estimation to extract patterns from intensity-gradient-magnitude feature space and guide the generation of TFs. Wang et al. [81] modeled the intensity-gradient-magnitude transfer function space as a Gaussian mixture, and designed the TF by taking each Gaussian component as a separate structure. Cheuk et al. [82] introduced a hierarchical volume exploration scheme based on a normalized-cut segmentation of the TF domain. Finally, Wang et al. [83] adopted Morse theory to automatically decompose the feature space into a set of valley cells for TF assignment. In our work, we apply mean-shift clustering in LH space to identify the unique material boundaries for further visualization. Our method is non-parametric and robust, and allows the visualization results to be easily modified by manipulating the clustering variable.

5.2 Automatic Transfer Function Design using Mean-shift Clustering

Our method considers volumetric data that consists of multiple boundaries, each of which is represented by a cluster in the LH histogram. These clusters are automatically extracted using mean-shift clustering. Then, the visual parameters of color and opacity are assigned to the voxels in each cluster. A bounding polygon based interaction widget allows for further manual modification of the TF. Fig. 5.1 presents an overview of our method.

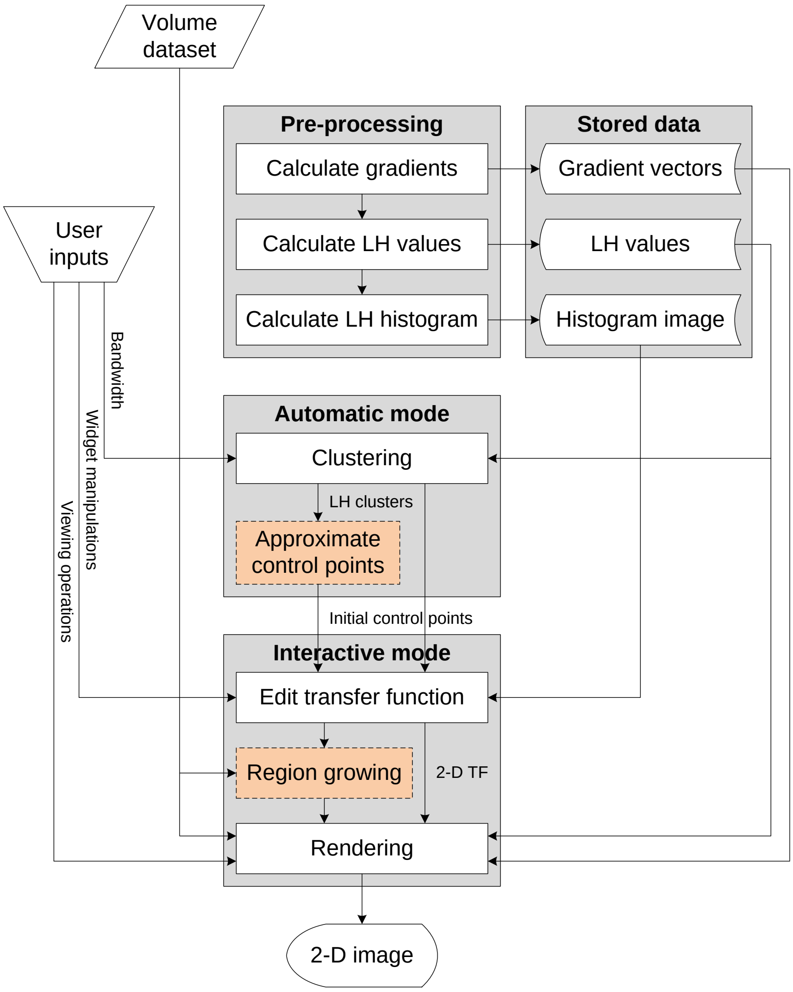


Figure 5.1: Overview of the method. Dotted rectangles represent optional processes for the semi-automatic mode.

5.2.1 Pre-processing

The gradient vector and LH values corresponding to each voxel are first computed in a pre-processing step. To calculate the voxel gradients, a second-degree polynomial function is used to approximate the local neighborhood density function [84]; the voxel gradients can then be obtained by solving for the coefficients of the polynomial function with an error minimization strategy. The advantages of this approximation method are: (1) the difference between the pixel spacing and the spacing between slices can be accounted for; (2) no computationally expensive interpolation method is needed to estimate the gradient vector of an arbitrary sampling point between voxels; and (3) this method is robust to noise since it does not interpolate the curve passing through all the given data points.

The lower (L) intensity and higher (H) intensity values of each voxel can be determined by tracking the boundary path using gradient integration along both gradient directions. Heun's method, which is a modified Euler's method, is applied to integrate the gradient field:

$$u_{i+1} = u_i + \frac{1}{2}d\left(\nabla f\left(u_i\right) + \nabla f\left(u_i + d\nabla f\left(u_i\right)\right)\right), \tag{5.2.1}$$

where u_i and u_{i+1} are positions of the current and the next sampling voxels, respectively, ∇f denotes normalized gradient vector when tracking H or L, and d is the step size of the integration. A step size of one voxel was experimentally found to be a good balance between accuracy and computation speed. The integration is halted upon reaching a local extremum or an inflexion point. To emphasize voxels on the boundary of two materials, each pair [L, H] is weighted by a factor w when being accumulated to create the LH histogram. The weight w is determined from

$$w = 1 - \frac{|d_L - d_H|}{d_L + d_H}, \tag{5.2.2}$$

where d_L and d_H are the accumulated distances along the boundary path from the current voxel to the sampling voxels corresponding to L and H, respectively. The approximation-based interpolation method reduces the effect of noise in the LH histogram, and thus improves the resulting visual quality.

An LH histogram is represented as an image of $N \times N$ pixels, where $N = 512$ as a compromise between the memory requirements and the visual quality. The histogram image is constructed by determining the correct bin for each [L, H] pair, scaling the sum of all corresponding weight factors taking the logarithm, and then mapping the resulting value to a color band, e.g. the cold-to-hot spectrum (Fig. 5.2). At the end of this preprocessing step, all the gradient vectors, the LH values, and the histogram image are stored in an intermediate data file for further processing.

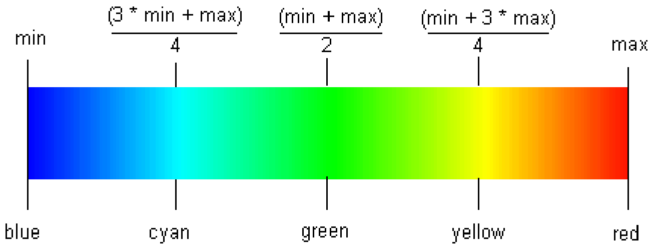


Figure 5.2: Cold-to-hot color ramp.

5.2.2 Mean Shift Clustering in LH Space

Mean shift clustering is a non-parametric feature-space analysis technique that seeks the modes of the given sample space. Compared with other clustering methods, mean shift clustering does not assume any specific structure or distribution of the data, and the number of clusters does not need to be known a priori. Mean shift clustering is more robust for general data, and hence is suitable for our application where the number and properties of structures in volumetric data is unknown. Also, mean shift clustering relies only on the bandwidth parameter B_w which correlates to the sensitivity of the clustering process, and thus is intuitive for the user to tune. From our experiments, a good B_w lies between 3%-12% of the maximum LH value, $\max_{LH} = \max(\max_L, \max_H)$. We apply mean shift clustering on the LH space to divide the LH histogram into multiple clusters. The procedure of mean shift clustering is summarized in the following algorithm:

1. Set the window bandwidth B_w.
2. For a point in the LH histogram, find all points that have LH values within the bandwidth B_w.
3. Find the mean μ_n of the set of neighboring points, with each point weighted by its voxel frequency.
4. Shift the window center to the new mean, and continue steps 2-4 until convergence. A cluster is deemed to have converged if the distance between successive means is less than ρB_w where ρ is a threshold which is preset as 0.001 in our experiments.
5. Repeat steps 2-4 for each point in the LH histogram.
6. Points that converge to the same modes (the converged cluster mean) are grouped as a single cluster, and clusters that have modes within $B_w/2$ of each other are also grouped as one cluster.

In our implementation of mean shift clustering, mean shift clustering is computed over discrete values in the LH histogram rather than over all the points in the volume. Since all voxels can only have discrete LH values, and since the LH histogram is relatively sparse, this speeds up the clustering operation and reduces the memory requirements. The resulting operation will be equivalent to an unmodified mean-shift clustering as long as each LH point is weighted by its voxel frequency (the number of occurrences of a particular LH value in the voxel volume) during the mean computation step.

5.2.3 Cluster-based Region Growing

The results of the mean shift clustering are sufficient for simple datasets, but for the more complicated datasets that are typical of medical imaging applications, additional information is needed to have a sufficient image quality. After mean shift clustering, each cluster is further passed through a region growing algorithm. The region growing algorithm is a means of incorporating spatial information to improve the visualization results.

In earlier work by Huang and Ma [85] on using region growing for volume visualization, a number of prior information and parameters, such as the initial seed points and the weighting factors for the cost function, must be provided to the region growing algorithm. In our approach, manual tuning of the region growing parameters is not necessary as the parameters and seed points will be assigned automatically based on the clusters obtained earlier during mean shift clustering. For each cluster previously extracted after mean shift clustering (and after manual user adjustment), the cluster voxels are used as the initial seed points. The standard deviation of the LH values of the cluster voxels are used to set the similarity tolerance of the region growing algorithm.

After each cluster has been passed through the region growing algorithm, the separate volumes must be merged into a single volume. If there are voxels belonging to more than one cluster, we simply merge all overlapping clusters. In future work, other criteria may be added to restrict merging to only cases where there is significant overlap between clusters. The algorithm for region growing enhancement is given below:

1. For each convex hull H_i obtained earlier, obtain the set of voxels V_i that have LH values lying within H_i.
2. For each cluster i, use the set of voxels V_i as the initial seeds for the region growing. The parameter τ_i is used as the LH tolerance.
   (a) Add each voxel in V_i to the output volume O_i.
   (b) For each voxel in O_i, add neighboring voxels to O_i only if they do not already belong to O_i and have LH values within τ_i of the seed voxel.
   (c) Repeat step 2b until no more voxels can be added to O_i.
3. For each pair of enhanced output volumes O_a and O_b, merge them if they overlap.

5.2.4 Assignment of Visual Parameters for TF Design

Our strategy to assign visual parameters to a cluster is based on the size of the region in the volume described by the cluster and the relative distance between that region and its neighbors. The *size* of the region R_i corresponding to the cluster C_i is coarsely estimated by the standard deviation σ_i of the positions of all the voxels $v_j = \left(v_x^j, v_y^j, v_z^j\right) \in R_i$

$$\sigma_i = \sqrt{\frac{1}{N_i} \sum_{v_j \in R_i} \left|v_j - \mu_i\right|^2}, \tag{5.2.3}$$

where N_i is the number of voxels in R_i, and μ_i is the mean of the positions of all voxels in R_i:

$$\mu_i = \frac{1}{N_i} \sum_{v_j \in R_i} v_j. \tag{5.2.4}$$

The *distance* between two regions R_i and R_j is defined as the Euclidean distance between the two corresponding mean values:

$$D\left(R_i, R_j\right) = \sqrt{\left(\mu_x^i - \mu_x^j\right)^2 + \left(\mu_y^i - \mu_y^j\right)^2 + \left(\mu_z^i - \mu_z^j\right)^2} \tag{5.2.5}$$

A region R_i *occludes* region R_j if

$$\begin{cases} \sigma_i > \sigma_j \\ \sigma_i > k_d D\left(R_i, R_j\right) \end{cases}, \tag{5.2.6}$$

where $k_d \geq 1$ is a pre-defined value. The opacity α_i assigned to region R_i is calculated by

$$\alpha_i = \frac{\alpha_i^*}{k_s\left(S_i + 1\right)}, \tag{5.2.7}$$

where k_s is an adjustable factor, S_i is the number of regions occluded by R_i, and α_i^* is the value corresponding to σ_i in the linear mapping of $\left[\min\limits_j \sigma_j, \max\limits_j \sigma_j\right]$ to a predefined opacity range $[\alpha_{\min}, \alpha_{\max}]$:

$$\alpha_i^* = \frac{\max\limits_j \sigma_j - \sigma_i}{\max\limits_j \sigma_j - \min\limits_j \sigma_j}\left(\alpha_{\max} - \alpha_{\min}\right) + \alpha_{\min} \tag{5.2.8}$$

Since smaller structures are more likely to be occluded than larger structures, this method of opacity assignment renders large structures more transparent than small structures. For the enhancement of voxels near boundaries, the voxel opacity α_v^i corresponding to a voxel v in the region R_i is individually modulated by the ratio of its gradient magnitude and the maximum gradient magnitude of all the voxels in the region:

$$\alpha_v^i = \alpha_i \frac{\|\nabla v\|}{\max\limits_{u \in R_i} \|\nabla u\|}. \tag{5.2.9}$$

The color parameter is difficult to assign as the materials have true colors that cannot be discerned from the CT/MRI volumes; assigning appropriate colors thus requires external knowledge. In our method, the color of each region can be assigned according to the ratio between the size of the region and the maximum size of all the regions, mapped onto a cold-to-hot spectrum. This operation will map small regions to hot colors, and large regions to cooler colors. Alternatively, since the number of regions is relatively small in most cases, we can use a pre-defined color array for this mapping. In addition, the color c_v^i of the individual voxel v in the region R_i is scaled by the ratio of its intensity value f_v and the maximum intensity of all voxels in the region:

$$c_v^i = c_i \frac{f_v}{\max\limits_{u \in R_i} f_u}. \tag{5.2.10}$$

For a better rendering result, this scaling is only applied to the brightness value of the corresponding color in the HSV color space.

5.2.5 Cluster Bounding Polygons for Manual Interaction

While mean shift clustering automatically assigns labels to each voxel in the volume, no automatic method can simultaneously satisfy the requirements of all users since different users have different visualization requirements and regions of interest. Minor adjustments made by the user will improve the quality and relevance of the visualization. To facilitate easy modification of the automatically extracted clusters, the voxel cluster labels are used to generate a set of cluster-bounding polygons. The advantage of cluster polygons is that they are easy to manipulate and modify via polygon and vertex operations. Entire clusters or individual vertices can thus be edited on the LH histogram. By creating or manipulating the control points (vertices) of a control polygon, the user encapsulates a region on the histogram and can thus select, remove, change the shape of, and assign optical properties to the region. Based on the properties of all polygons, a 2-D TF is generated and transferred to the renderer to produce the final image. Optionally, the user can apply a post-processing step using region growing to enhance the visualization result.

Ideally, each cluster polygon should only contain all voxels assigned to that cluster, but this requires computing concave bounding polygons which is computationally expensive. We simplify the computation by assuming that the bounding polygons are convex polygons, which can be computed in $\Omega(n \log(n))$ time by fast convex hull algorithms such as Andrew's monotone chain algorithm [86]. To resolve overlaps between bounding polygons, collision detection is performed for each pair of polygons. For each overlap, there are two intersections. A dividing line is drawn between the two intersections and each partitioned area is assigned to the cluster it is nearest to. This disambiguation scheme is illustrated in Fig. 5.3.

Finally, the regions along the main diagonal of the LH histogram belong to voxels lying within the same material, i.e. material not lying on the material interfaces [72]. These clusters are unimportant for visualization and can be discarded or rendered with a low opacity value. After

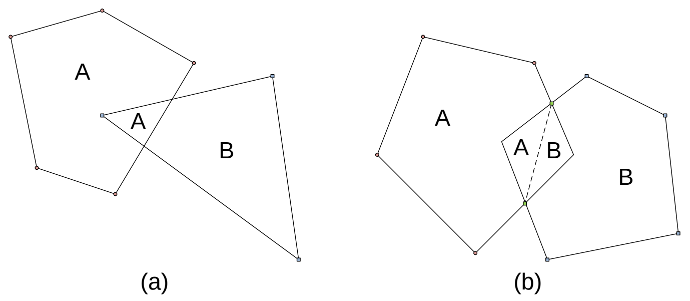


Figure 5.3: Two examples of the overlap disambiguation scheme.

the cluster bounding polygons are generated, a check is rendered to detect and discard such clusters. All polygons with at least one vertex within a diagonal window of the main diagonal of the LH histogram are treated as clusters of non-boundary material. The diagonal window is experimentally defined to have a width of 2% of the range of LH values. The procedure for computing the cluster bounding polygons is demonstrated in Fig. 5.4 and summarized in the following algorithm:

1. For each cluster C_i obtained from the mean shift algorithm, obtain the set of points P_i and compute a convex hull H_i containing all the points in P_i.

2. Construct a convex polygon H_{diag} using the following 6 coordinates: $[0, 0]$, $[0, 0.01 \times \text{max}_H]$, $[0.99 \times \text{max}_L, \text{max}_H]$, $[\text{max}_L, \text{max}_H]$, $[\text{max}_L, 0.99 \times \text{max}_H]$, $[0.01 \times \text{max}_L, \text{max}_H]$, where max_L and max_H are the maximum values in the LH histogram. For each convex hull H_i, if any vertex in H_i lies in H_{diag}, the cluster C_i is treated as a non-boundary cluster

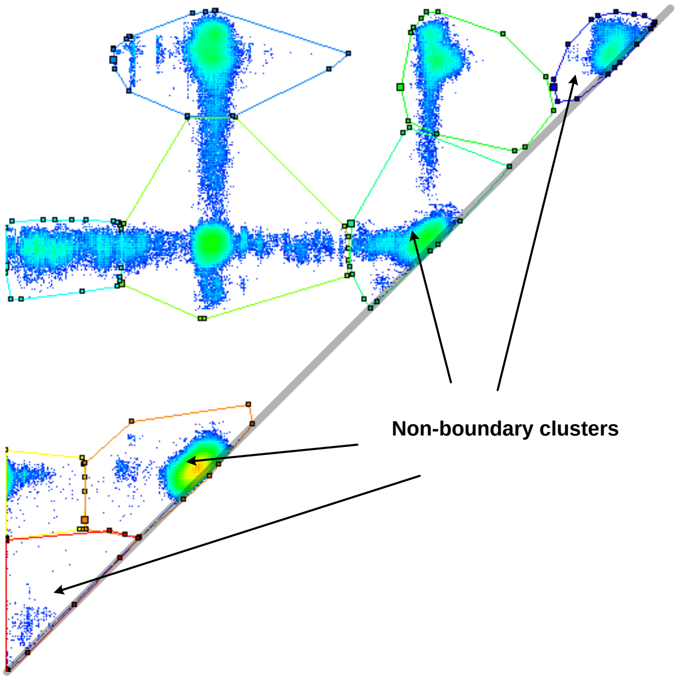


Figure 5.4: Demonstration of non-boundary cluster removal on the Tooth dataset.

and is removed or rendered with low opacity.

3. For each pair of remaining convex hulls H_a and H_b, compute the intersection, if any, between each combination of hull segments. If there are intersections denote them as I_a and I_b. Add both I_a and I_b to both hulls H_a and H_b, and remove all hull points interior to the line segment created by H_a and H_b.

5.3 Results and Discussion

Four 16-bit CT volumes were used in our experiments: the Tooth (256 × 256 × 161), Feet (256 × 256 × 125), Head (128 × 256 × 156), and Pig

(256 × 256 × 128) datasets. The computing platform was a 2.66 GHz Intel i5-750 system equipped with 4 GB RAM and a NVIDIA Quadro FX 3800 graphics card. The times required to compute the pre-processed data were 151s, 213s, 200s, and 205s, respectively. Using a GPU-based renderer employing ray marching through a 3-D texture and setting the display window resolution to 512 × 512 pixels, a real-time frame rate was achieved for all the four datasets. In our experiments, the value of k_d was set to 1.

Fig. 5.5 shows the result of applying our method in automatic mode to the Tooth data set. The bandwidth B_w was chosen as 7% of max_{LH}, and the total time for clustering was 157ms. The clustering result generated by the mean shift clustering algorithm (Fig. 5.5(b)) closely resembles the optimal manual LH clustering from a previous work [72]. Hence, the mean shift algorithm is capable of quickly generating clusters of similar quality to semi-automatic methods. The clustering speed is also sufficiently fast to allow the user to interact with the bandwidth parameter and receive the updated visualization results on the fly. When operating in the automatic mode, non-boundary clusters (clusters along the main diagonal of the LH histogram) are rendered with a low opacity. Occluding regions are also assigned lower opacity values to ensure that smaller and interior structures are visible. These steps ensure that separate regions within the volume are visible and distinct in the resulting visualization (Fig. 5.5(c)).

Fig. 5.6 shows the same volume rendered in the semi-automatic mode. The semi-automatic mode allows user to modify the TFs generated previously in the automatic mode. This TF modification is performed on the polygonal approximations of the clusters. In Fig. 5.6(a) the opacity of the cylinder was set to 0 and the pulp-dentine boundaries were set to the same color. This pulp-dentine boundary was separated into two disjoint clusters on the LH histogram because of the thin object effect [72]. After manually adjusting the color and opacities for the two clusters, the rendering result was improved (Fig. 5.6(b)). However, some discontinuities in the pulp boundary still existed. These discontinuities cannot be resolved by clustering on the LH space, or similar methods that rely solely on the LH histogram for classification. Our algorithm includes a region growing step to address these issues. Fig. 5.6(c) shows the result after region growing was performed. The discontinuity in the pulp has been filled by the region growing algorithm to yield a single continuous boundary.

For the Feet dataset (Fig. 5.7), the automatic mode with B_w as 7% of $\max_{LH}$ was employed to generate the initial clusters for the LH histogram (Fig. 5.7(a)) and initial rendering (Fig. 5.7(b)). The clustering operation took 3578ms to complete. The opacities of the skin and base plate were edited in the semi-automatic mode to obtain the final visualization

(Fig. 5.7(c)), which clearly showed the bones within the feet.

For the Head dataset (Fig. 5.8), we used our algorithm in the automatic mode and varied k_s to examine its effect on the visualization. B_w was set to 6% of $\max_{LH}$ and the clustering operation was completed in 4594ms. In Figs. 5.8(b) and 5.8(c), the TF assignment algorithm was run with k_s set to $k_s = 0.1$ and $k_s = 0.3$, respectively. The results confirm that by increasing k_s, occluded internal regions can be selectively revealed. This demonstrates that our algorithm is capable of automatically assigning colors to distinct regions within volumes, and also capable of automatically assigning the opacities of each region such that all regions are visible and not occluded.

For the Pig dataset (Fig. 5.9) which we acquired from a surgical planning experiment, finding a suitable TF is difficult due to the complexity and number of structures within the volume. It is difficult for the user to properly select any clusters from the LH histogram. Mean shift clustering ($B_w = 4\% \max_{LH}$) alleviates this problem by producing an initial set of clusters (Fig. 5.9(a)) which can be quickly modified to achieve the desired visualization. Due to the complexity of the volume, clustering took more time to complete (7500ms). The results from the automatic mode (Fig. 5.9(b), 5.9(c)) show that the regions of the volume that could be important for surgical planning, such as the bones, blood vessels, and surgical

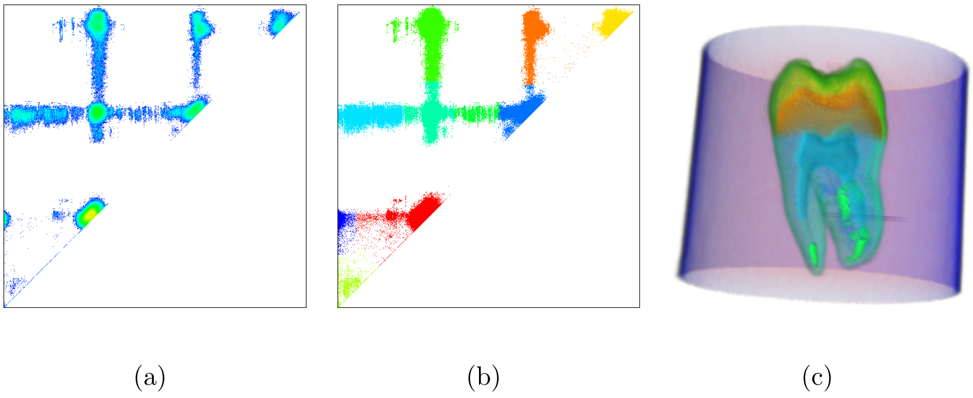
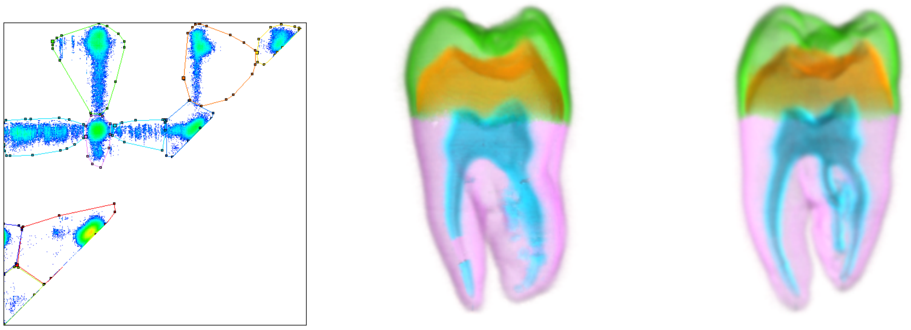


Figure 5.5: Automatic TF design for rendering the Tooth dataset: (a) The LH histogram; (b) Generated clusters; (c) Rendered image from clusters.

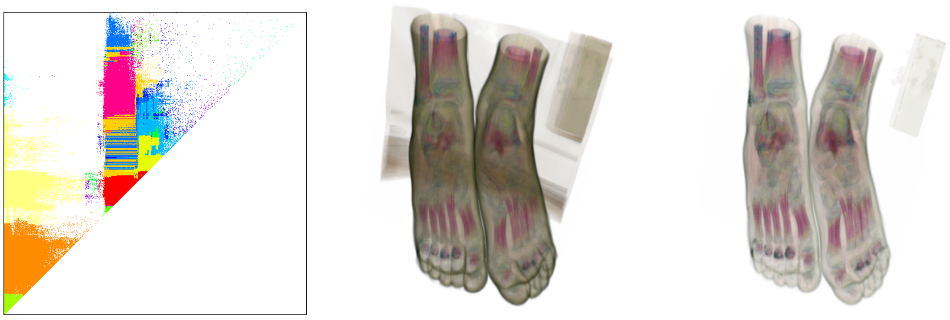
markers, are clearly visible. Hence, our automatic method is suitable for medical visualization, particularly for surgical planning tasks, where good visualization with clear indication of the regions of interest is important.

The visualization results show that our automatic method is capable of assigning the visual properties of color and opacity to obtain good renderings. Comparing with Šereda's method that uses hierarchical clustering [79], our method does not need to generate initial clusters which may strongly affect the rendering results. Furthermore, the user is not required to adjust the cluster colors or opacities as these are determined automatically by our algorithm.



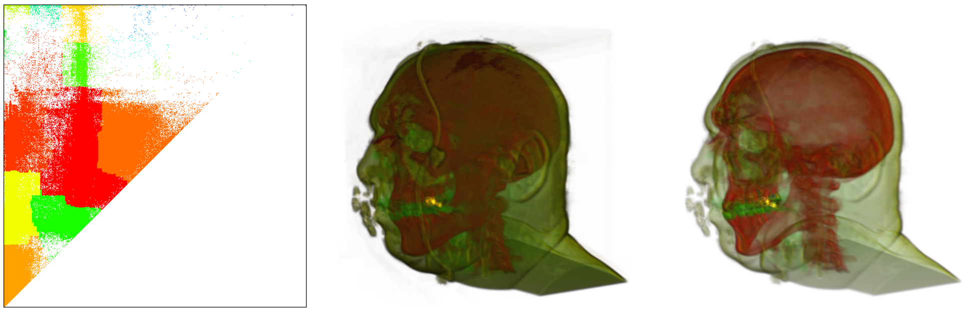
(a) (b) (c)

Figure 5.6: Semi-automatic TF design for rendering the Tooth dataset: (a) New TF based on approximated polygons; (b) Rendered image; (c) Rendered image using region growing.



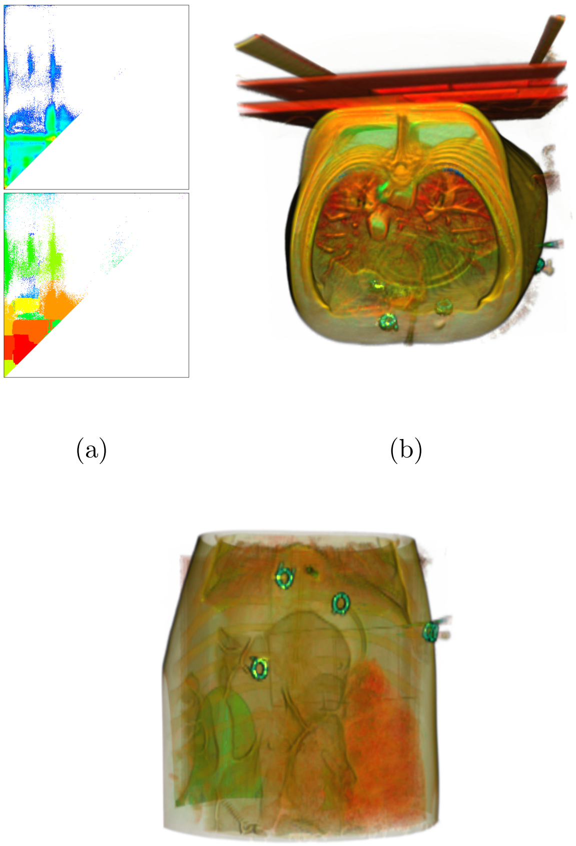
(a) (b) (c)

Figure 5.7: Volume rendering of the Feet dataset: (a) Clusters; (b) Rendered image with $k_s = 0.3$; (c) Rendered image with $k_s = 0.3$ then decrease the opacity of the skin and set zero-opacity for the back plate.



(a) (b) (c)

Figure 5.8: Volume rendering of the VisMaleHead dataset: (a) Clusters; (b) Rendered image with $k_s = 0.1$; (c) Rendered image with $k_s = 0.3$.



(a) (b)

(c)

Figure 5.9: Volume rendering of the Pig dataset: (a) LH histogram (upper) and clusters (lower); (b) and (c) Rendered images using automatic mode.

5.4 Summary

We have developed a system for the automatic generation of TF for medical volume visualization. Mean shift clustering identifies clusters in the LH domain that correspond to material boundaries, and also generates the seed information for a region growing algorithm to improve clusters by incorporating spatial constraints. An automatic TF design module then assigns color and opacity to each cluster based on the relative sizes and distances between clusters. The proposed system automatically generates good visualizations while preserving a high degree of freedom for the user to adjust the rendering results. The visualizations generated by the proposed automated method are comparable to existing state-of-the-art approaches.

CHAPTER 6

One-class Classifiers for Biometric Recognition in a Surgical Data Access Application

There is much interest in touch-free computer interfaces for remote computer interaction. Remote computer interaction may be motivated by several different reasons, such as interaction from a distance (for making presentations), a need for an unencumbered 'desk-free' interface (for games), or sterility requirements (for surgical applications). In particular, sterility requirements for surgical settings motivate research into touch-free computer input and interaction. Gesture-based approaches are popular for remotely inputting one of several pre-determined commands, or to translate gestures into more traditional mouse plus cursor commands to interact with existing computer interfaces.

For surgical augmented reality with multiple users, context-selection offers the possibility of interaction that is more efficient. Instead of offering only a single mode of interaction, several work-contexts can be defined and applied to each identified unique user . This functionality can be exploited in several novel ways for augmented reality in a surgical environment. First, different gesture-profiles can be loaded for each user, allowing the same gesture to carry out different actions when performed by different users. This is useful for reducing the number of gestures operators have to memorize, and can help to limit the gesture set to the simplest and most consistently-recognized set of gestures. Another way of employing user-specific context selection is to project different interfaces or data for each user. For example, surgical assistants may be assigned different roles for an operation, such as manipulation of the ablation system or maintaining of patient homeostasis, and the projected AR system can intelligently switch interfaces depending on the current user. In this way, context-selection can greatly improve the efficiency of human-robot interaction in the surgical setting.

This chapter describes a method for multi-user biometric recognition in a gesture-based surgical data access system. A Kinect sensor is used to capture depth images of a user's palm, and biometric features are then extracted from the palm depth images. Based on the palm-based biometrics, users are identified and the specific work environments specific to each user are loaded, allowing users to quickly access data and interfaces unique to their work scope. For the biometric recognition task, we propose a one-class classifier based on the nearest neighbor distance (NN-d). A one-class classifier system is applied to correctly recognize and classify palms of previously registered users, while rejecting unknown and unregistered users. The results demonstrate that one-class classifier systems are useful for learning the properties of unknown distributions, and can be used for simple biometric recognition in the gesture-based surgical data access system.

6.1 Related Work

Palm-based biometric recognition is typically performed using scanners or CCD cameras as the input sensor; such recognition devices require physical contact with the sensing device for the palm images to be acquired, and are not suitable for a non-touch surgical setting. Non-contact biometric verification is more hygienic, and has the potential to be used in settings other than surgery. Ong et al. [87] introduced a webcam-based system for touch-less palm print recognition from low-resolution hand images. Their method applies hand tracking to extract a square palm print ROI; the local binary pattern texture descriptor is then used to describe the distinctive texture information contained in the palm region, and the resulting features applied to train a probabilistic neural network. Ong's method can be considered to adopt the statistical approach to biometric verification, where palm print features are not explicitly designed using expert knowledge (i.e., to detect and track explicit ridges and structures), but instead extracted automatically using machine learning techniques. Other methods that follow the statistical approach may use principal components analysis [88], Fisher's linear discriminant [89], or independent components analysis [90] to perform subspace analysis to extract a descriptive representation of the distinctive palm features. The advantage of statistical over structural approaches is that structural features may be unreliable and structural matching computationally expensive, while statistical features are robust if given sufficient training samples and have low computation cost for classification.

The main weakness of appearance-based biometric verification approaches is that in a surgical setting, appearance features are non-usable. Surgical gloves and stains acquired during the operation obscure hand textural features. Surgical environments have strong lighting which may result in large lighting variations for image capture, further reducing recognition accuracy [91]; under harsh lighting and geometric conditions, the extracted hand edges may be unstable and do not capture internal structure [92]. Lastly, the distance between the hand and sensor impedes accurate capture of hand texture details [87], or requires costly high resolution cameras [93]. These disadvantages are not present in a geometry-based hand recognition system [91, 94]. Geometric features possess the important qualities of being time-invariable, difficult to counterfeit, and are unique to the individual, and hence are good biometric features [94]. Furthermore, the usage of gloves does not seriously impair geometric features.

Instead of using RGB cameras for acquiring geometric features, we choose Microsoft Kinect as the sensing device. The Kinect is commodity hardware that has a built-in depth camera capable of capturing depth images; this depth information can be used to control for changes in the distance between the hand and the imaging sensor. A depth-based image plane realignment also allows some variation in hand orientation and pose. These advantages allow a depth-based geometric approach to be applied in a more general setting with fewer constraints on the hand position and pose.

Biometric verification consists of two main tasks; the first is to recognize a registered user, and the second is to reject unregistered users. Conventional classification algorithms learn a decision boundary between a target class and other classes and excel in dealing with the first problem, but are unable to reject samples from unknown classes that are absent during training [95]. Therefore, conventional classification algorithms are unsuitable for biometric verification as they are not designed to detect novel outliers.

The detection of unknown users can be considered a novelty detection task [96]. Approaches to novelty detection typically involve modeling the data distribution from known samples and using a distance or similarity measure to detect abnormalities. Broadly, novelty detection methods may belong to two major classes, parametric or non-parametric. Parametric approaches model the data based on assumed statistical distributions or properties and apply the constructed model to determine the probability of a sample being an unknown outlier, while non-parametric approaches make no assumptions on the data distribution. Amongst the simplest of parametric approaches is to model the data as a Gaussian distribution and to reject outliers by the number of standard deviations away from the class mean [97], or to use box-plot summaries to identify atypical samples [98]. More advanced methods apply more complex data modeling techniques such as Gaussian mixture modeling [99]. Unfortunately, parametric methods require a priori knowledge and may not be suitable for real-world problems with unknown data distribution, where the data may have multiple discontinuous modalities that are not Gaussian [100]. Non-parametric methods may use Parzen density estimation [101, 102] to obtain a non-parametric density estimate, or apply K-NN technique to estimate the width of the local density.

Related to novelty detection is one-class classification. Unlike conventional classification, one-class classifiers assume that only training instances of the object class are available; therefore, one-class classifiers focus on constructing a parsimonious class model with a minimal chance of accepting outliers [103]. Like novelty detection, density estimation methods such as Gaussian mixture modeling and Parzen density estimation are often used to model the distribution of the class samples. Besides density estimation, Vapnik [104] has argued for the more direct solution of constructing a data boundary without explicitly modeling the data density. Boundary-based methods include the K-centers method, and the nearest neighbor distances (NN-d) method. K-centers involves the fitting of several hyper-spheres of equal radii to the training data such that the maximum distance of all minimum distances between the hypersphere centers and training samples is minimized; the hyperspheres thus enclose the data density and can be used as a decision boundary for outlier rejection [105]. Instead of fixing a distance radius, the NN-d method adaptively determines a local radius about each sample point by comparing the distance of a test point to its nearest neighbor in the training data with the distance from the nearest neighbor to *its* nearest neighbor [95]. Thus, NN-d produces a tight boundary in densely sampled regions where the confidence of classification is higher, and a looser boundary in sparsely sampled regions where there is less confidence of the true boundary. In our work, we propose a modified NN-d method to improve the trade-off between outlier rejection and sample classification. In addition, we propose the use of NN-d in a two-stage method to allow any classifier to be used in a biometric recognition capacity.

6.2 A System for Biometric Recognition

The biometric verification interface consists of two components, a feature extraction module that produces a set of feature descriptors from the depth map of a palm, and a classification module to recognize the presented palm. In this section, we describe the image preprocessing and feature extraction elements of the biometric recognition system.

6.2.1 Finger Segmentation from Palm Depth Images

The dimensions of the palm and fingers provides a physical invariant that can be applied for biometric recognition tasks, but it is generally not possible to reconstruct the physical dimensions of the palm and fingers from only a 2D projection such as a color image. However, the additional depth information from the Kinect depth sensor allows for the local scale to be estimated. From Fig. 6.1, a line of length L at depth D appears to be of the same length as a line of length $2L$ at depth $2D$. Similarly, a pixel of depth D represents an area of only a quarter that of a pixel at

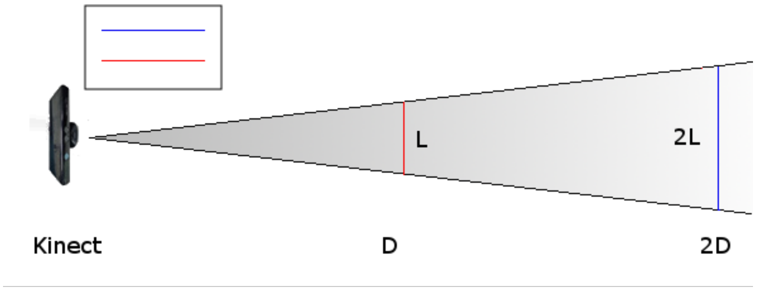


Figure 6.1: Scale variation with depth.

depth $2D$.

A palm is represented by a set of feature descriptors for each finger in the palm. The fingers are first segmented from the palm using a variant of the valley-peaks extraction [87, 106] algorithm in order to define a polygonal ROI about each finger, as shown in Fig. 6.2. The valley-peaks finger segmentation algorithm determines the finger tips and finger webs, which lie at the maximum and minimum distances to the palm center, and constructs a bounding polygon about each finger. The finger segmentation algorithm (also shown in Fig. 6.3) is as follows:

1. Apply thresholding to obtain the set of edge pixels, $P_{\text{edge}} \in p_{\text{edge}}$.

2. Compute the central point, $p_{\text{central}} = \text{argmax}(\min(|p_{\text{edge}} - p_{\text{central}}|))$, which maximizes the distance from itself to the closest edge pixel.

3. Compute the distance $d_{i,j}$ from each edge pixel p_i to every other edge pixel p_j. Also compute the distance $d_{i,c}$ from each edge pixel p_i to the central point p_{central}.

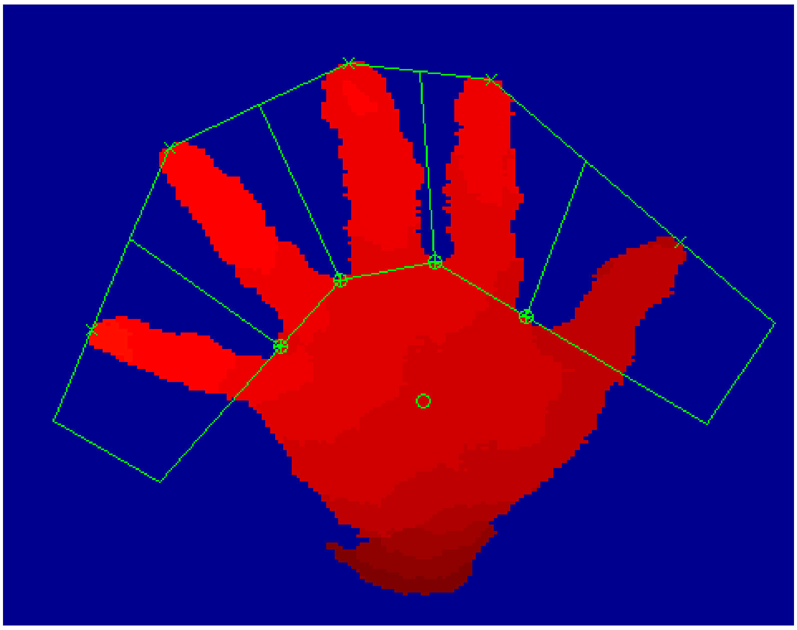


Figure 6.2: Segmented polygonal ROIs for fingers, where '+' indicates the peak points, '⊕' indicates the valley points, and 'o' represents the central point.

4. Using a neighborhood radius r, determine the valley points $P_{\text{valley}} \in p_{\text{valley}}$, where a point p_i is a valley point if $d_{i,c} < d_{j,c}$ for $j \in d_{i,j} < r$.

5. Using a neighborhood radius r, determine the peak points $P_{\text{peak}} \in p_{\text{peak}}$, where a point p_i is a peak point if $d_{i,c} > d_{j,c}$ for $j \in d_{i,j} < r$.

6. Sort both P_{valley} and P_{peak} according to the angle from p_{central}.

7. A bounding polygon for each finger is constructed from a quintuple comprising of successive pairs of valley points, a peak point, and the midpoints between successive pairs of peak points.

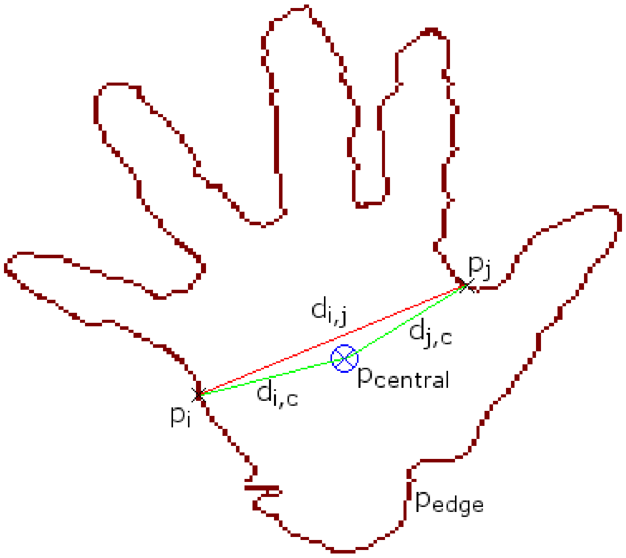


Figure 6.3: Finger segmentation algorithm, where p_{edge} is indicated by the red edges, and the points p_{central}, p_i, and p_j are indicated accordingly.

6.2.2 Palm Feature Descriptors

Each finger is extracted using the polygonal ROI, and a set of 18 descriptors $F_1 - F_{18}$ are computed for each finger from its depth image; these finger descriptors are concatenated to form a 90-dimensional descriptor for each palm. The descriptors $F_{10} - F_{18}$ measure the dimensions (area and lengths) of each finger and phalanx segment (Fig. 6.4), while descriptors $F_1 - F_9$ represent the same quantities multiplied by a scaling factor S computed on the finger or phalanx segment. The scaling factor S for a region is the mean of the depth D of each pixel in the region:

$$S = \bar{D}. \tag{6.2.1}$$

The inclusion of the scaling factor controls for the effect of apparent size variations resulting from objects at different depths.

The scaled descriptors $F_1 - F_9$ are given below:

1. The scale-adjusted area of the finger, $F_1 = \Sigma(D^2)$.
2. The scale-adjusted major axis length, $F_2 = L_{\mathrm{maj}} S_{\mathrm{finger}}$.
3. The scale-adjusted minor axis length, $F_3 = L_{\mathrm{min}} S_{\mathrm{finger}}$.
4. The scale-adjusted average width of each third of the finger, $F_4 = W_{\mathrm{s1,ave}} S_{\mathrm{s1}}$, $F_5 = W_{\mathrm{s2,ave}} S_{\mathrm{s2}}$, $F_6 = W_{\mathrm{s3,ave}} S_{\mathrm{s3}}$.
5. The scale-adjusted maximum width of each third of the finger, $F_7 = W_{\mathrm{s1,max}} S_{\mathrm{s1}}$, $F_8 = W_{\mathrm{s2},max} S_{\mathrm{s2}}$, $F_9 = W_{\mathrm{s3,max}} S_{\mathrm{s3}}$.

Lastly, each feature is standardized to zero-mean and unit-variance.

6.3 Nearest Neighbor Distances for Biometric Recognition

Biometric recognition uses classifiers to match presented palms to their preregistered feature representations stored in the database. Based on the user identity, the appropriate data and interface settings and preferences unique to that user can be loaded. In the case of an unregistered user, for

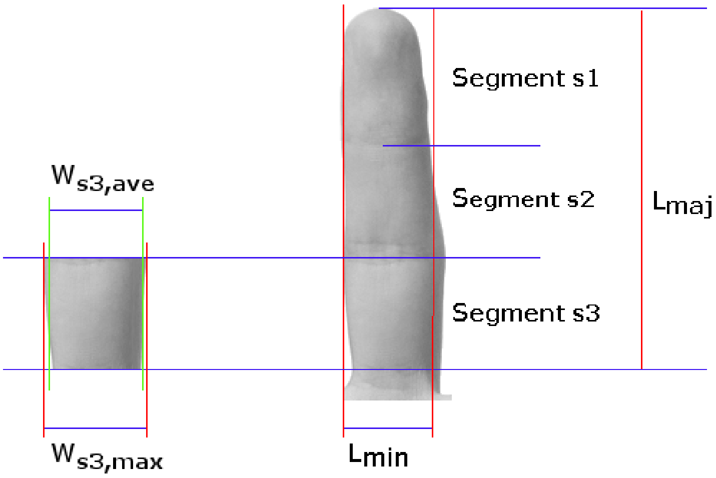


Figure 6.4: Finger and phalange lengths used in feature descriptors. The average and maximum widths of the third finger segment ($W_{\text{s3,ave}}$ and $W_{\text{s3,max}}$ respectively) are indicated.

example a guest surgeon or another member of the surgical staff, a default interface with fewer access privileges can instead be loaded.

For biometric recognition with only known users, any classifier can be applied to the palm features described earlier. However, if unknown users are present, novelty detection schemes are needed to identify these guest users. In this section, we describe three innovations for novelty detection using NN-d.

NN-d is a boundary method which estimates the class boundary for each individual class based on the local density [95] , and is suitable for outlier rejection. In NN-d, the distance $d_{\text{NN}_1}(x)$ from a sample x to its nearest neighbor $\text{NN}_1(x)$ in the training set is compared with the distance from the nearest neighbor $\text{NN}_1(x)$ to its nearest neighbor $\text{NN}_1(\text{NN}_1(x))$.

The NN-d decision rule for determining if a new sample x belongs to a class i is

$$\begin{cases} \text{Accept if} & d_{\mathrm{NN}_{1,i}}(x) <= d_{\mathrm{NN}_{1,i}}(d_{\mathrm{NN}_{1,i}}(x)) \\ \text{Reject if} & d_{\mathrm{NN}_{1,i}}(x) > d_{\mathrm{NN}_{1,i}}(d_{\mathrm{NN}_{1,i}}(x)) \end{cases} . \tag{6.3.1}$$

Intuitively, if x is as close or closer to a class sample than other items of the same class, it is likely to be a class inlier. Also, if the local density about a training point is dense, then the estimated boundary about the point is tight, and outliers are less likely to be accepted; conversely, a sparse local density results in a loose estimated boundary with a higher probability of accepting out-of-class samples.

6.3.1 Large Margin Nearest Neighbor Distances

In k-NN and in NN-d, the distance metric used is typically not optimized for classification. Large margin methods compute a space reprojection that attempts to maximize the separation between different classes by minimizing the number of impostors (nearest neighbors that belong to different classes) for all data samples in the training set [107]. Under the large margin reprojection, the distance between classes is increased, thus reducing the classification error. A large margin reprojection also reduces the impact of spurious feature dimensions. Large margin nearest neighbors is computed using semi-definite programming, which can be computationally expensive.

Large margin reprojection can be applied to NN-d as a preprocessing method. We use the large margin method described by Weinberger et al. [108] to compute a projection matrix M on the training data. The distance between points x_i and x_j under large margins reprojection is thus

$$d(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j). \tag{6.3.2}$$

6.3.2 Class Specific Radius Optimization

One weakness of NN-d is that parts of the feature space within the target distribution may be incorrectly rejected [95]. Consider a data set drawn from a uniform distribution; under a LOOCV evaluation scheme, only samples which are mutual nearest neighbors can be correctly identified as in-class members. In particular, for data which is poorly sampled and where sampled objects are tightly bunched together, the NN-d is likely to give poor results.

Instead of using the distance from the closest training sample to its nearest neighbor, we add a fixed distance r to that distance to increase the size of the class boundary. This distance radius is computed for each training class, and serves as a heuristic to control for the regularity of sampling present in each class. The distance radius r_i for the i-th class C_i

is computed from all training samples from C_i via

$$r_i = \max_{x \in C_i}(d_{\mathrm{NN}_{2,i}}(x) - d_{\mathrm{NN}_{1,i}}(x)), \tag{6.3.3}$$

where $d_{\mathrm{NN}_{1,i}}(x)$ and $d_{\mathrm{NN}_{2,i}}(x)$ denote the distance from a sample x to its first and second nearest neighbors in C_i. If a class is regularly sampled, then $d_{\mathrm{NN}_{1,i}}(x)$ and $d_{\mathrm{NN}_{2,i}}(x)$ should be very close, thus reducing r_i, whereas r_i is larger if the class is irregularly sampled. Therefore, r_i is a class-wise smoothing parameter to reduce the impact of sample bunching. The modified decision rule for determining if a new sample y belongs to a class i is

$$\begin{cases} \text{Accept if} & d_{\mathrm{NN}_{1,i}}(y) <= d_{\mathrm{NN}_{1,i}}(d_{\mathrm{NN}_{1,i}}(y)) + r_i \\ \text{Reject if} & d_{\mathrm{NN}_{1,i}}(y) > d_{\mathrm{NN}_{1,i}}(d_{\mathrm{NN}_{1,i}}(y)) + r_i \end{cases}. \tag{6.3.4}$$

Under the new decision rule, the decison boundary about each training sample is a combination of both an adaptive distance based on the local density and a fixed radius for smoothing.

6.3.3 A Two-stage Method for Adapting Classifiers for Outlier Rejection in Multi-class Problems

In order to use conventional classifiers for biometric verification, we propose a two-stage method using NN-d as a outlier filter. Under the two-stage model (Fig. 6.5), the training data is used to train both a NN-d classifier and a conventional classifier. When presented with new samples, NN-d performs the novelty detection task while the conventional classifier sorts all samples accepted as inliers into the trained class labels. No additional modifications are required for either NN-d or the conventional classifier. For best results, the NN-d can be a modified NN-d scheme incorporating both large margins method and the class-specific radius optimization.

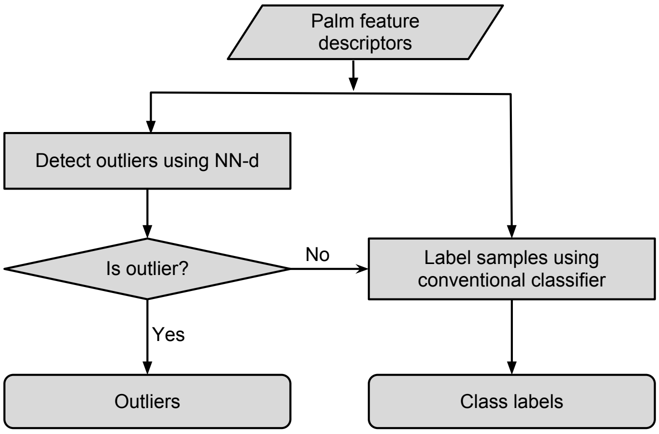

Palm feature descriptors
Detect outliers using NN-d
Is outlier?
No
Label samples using conventional classifier
Yes
Outliers
Class labels

Figure 6.5: Two-stage model for outlier rejection using conventional classifiers.

6.4 Results and Discussion

6.4.1 Experiment Methodology

The data sets employed in our experiments were collected using a data collection interface. Both palms of each user are recorded at varying distances (60 to 200 cm) with the fingers spread out at different extents. User palms were captured with frontal angle variation of ±15°. Two sets of experiments were conducted; the first set of experiments was to validate the biometric recognition system for bare and gloved palms, while the second set of experiments was to evaluate the biometric verification and novelty detection performance.

Two experiments were conducted to evaluate and determine the most appropriate classifiers for the biometric recognition system on bare and gloved palms. In the first experiment, volunteers were instructed to present their bare palms to the data capture system; this task mimics the traditional biometric palm recognition task. In the second experiment, volunteers were instructed to wear surgical gloves and present the gloved palms to the data capture system; this task mimics aseptic environments where operators are required to wear surgical gloves to preserve sterility. In total, 1602 bare palm samples were collected from eight users for the first experiment; as the left and right palms have different dimensions, this forms a total of 16 different ungloved palms labels. For the second experiment, 858 gloved palm samples were collected from six users, resulting in a total of 12 different gloved palm labels.

A further two experiments were conducted to evaluate the biometric verification and novelty detection performance. In the first evaluation task, the classifiers are trained on the training set with all class labels represented, and the trained classifier is used to assign class labels to the test set; this task mimics the case where all users are registered. In the second evaluation task, the training set excludes all cases of one class, and the trained classifier is tested on the complete testing set. Test samples from the unseen class are to be identified as outliers, and not assigned to one of the seen classes; this task mimics the more general biometric verification task, where unregistered users may gain access to the system and should not be wrongly verified. In total, 723 bare palm samples were collected from three users; as the left and right palms have different dimensions, this forms a total of six possible palms labels.

Leave-one-out cross-validation (LOOCV) is employed to obtain the classification accuracy for each combination of features and classifiers [52]. LOOCV is performed by repeatedly training the classifier system on all-but-one of the available samples, then testing the trained classifier on the unseen sample. LOOCV ensures that each classifier is trained on the maximal number of training samples while using all available data for testing. The evaluation metrics used are accuracy and macro-averaged F-measure. The macro-averaged F-measure is a generalization of the F-score for multi-class problems [109], and it reflects a classifier's precision and recall performance. The F-measure value ranges from $(0, 1)$, where a larger value corresponds to a higher classification quality. While accuracy is dominated by the classifier's performance on common classes, the macro-averaged F-measure assigns equal weight to all classes regardless of the class frequency, and thus is influenced more strongly by infrequent categories. The macro-averaged F-measure can be computed by:

$$F_i = \frac{2\pi_i\rho_i}{\pi_i + \rho_i}, F(\text{macro-averaged}) = \frac{\sum F_i}{M}, \tag{6.4.1}$$

where M is the total number of classes, and π_i and ρ_i are defined respectively as:

$$\pi_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{6.4.2}$$

$$\rho_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, \tag{6.4.3}$$

where TP_i is the number of true positives, FP_i the number of false positives, and FN_i the number of false negatives for class i.

6.4.2 Benchmarking against Conventional Classifiers

To benchmark the performance of the proposed novelty detection methods against conventional classifiers, we apply a simple outlier rejection scheme to the conventional classifiers. For a given conventional classifier trained on the training set X, the training samples x are passed into the classifier to obtain the posterior probability $p(L|x)$ for each class label L. Let the class labels with the highest and second highest posterior probabilities be denoted by L_1 and L_2 respectively. A quotient $q_x = \frac{L_1}{L_2}$ is computed for all training samples in X. Subsequently, put q_x into the group $q_{correct}$ or q_{wrong} based on whether the training sample x was correctly labeled by the classifier.

The quotient q is an indicator of the relative confidence of a class label compared to the next most probable label. If a sample is a class inlier, q should be high for some label L; if a sample does not belong to any trained label, then q would be low. Choosing a threshold q_τ for q would allow some out-of-class samples to be detected. q_τ is chosen by minimizing the cost C of misclassification on the training set, where C is the sum of the number of items in $q_{correct}$ and q_{wrong} that are smaller and larger than q_τ respectively.

6.4.3 Results: Bare Palms and Gloved Palms

Table 6.1 shows the evaluation results using different classifiers to recognize bare palms and gloved palms. Bare palms were well recognized with most classifiers, and the best results were obtained with large margin K-nearest neighbors. For gloved palm recognition, biometric recognition accuracy is comparatively degraded, but a classification accuracy of 95% is still possible.

Table 6.1: Evaluation of classification methods on bare and gloved palms

	Bare Palms		Gloved Palms	
Classifier	Accuracy	Macro-F	Accuracy	Macro-F
3-NN	0.9148	0.9086	0.8924	0.8759
Bayesian	0.8285	0.8180	0.7114	0.7180
Linear discriminant	0.9479	0.9420	0.9416	0.9334
Decision tree classifier	0.6823	0.6601	0.7011	0.6868
Random forest	0.9229	0.9171	0.8947	0.8820
SVM (SVM$^{\text{light}}$)	0.9139	0.9093	0.8811	0.8613
SVM (LIBSVM)	0.9433	0.9388	0.9098	0.9037
ELM	0.8845	0.8785	0.8345	0.6188
Random Subspace (K-NN)	0.9323	0.9280	0.9049	0.8887
Large margin NN	0.9656	0.9615	0.9569	0.9535
BOOSTMETRIC	0.9406	0.9355	0.9233	0.9152

6.4.4 Results: All Users Registered

Table 6.2 shows the evaluation results for the biometric task when all users are registered. The results for one-class classifiers are not included here, as they revert to a k-nearest neighbor classifier if outlier rejection is not used. Most classification algorithms are able to achieve an acceptable ($\geq$ 90%) classification accuracy, with the exception of decision tree classifiers. The best results were obtained using linear discriminant analysis and large margin K-nearest neighbors, suggesting that some form of space reprojection is needed to improve classification results.

Table 6.2: Evaluation of classification methods, all users known.

Classifier	Accuracy	Macro-F measure
Naive Bayes classifier	0.9281	0.9300
Linear discriminant analysis	0.9834	0.9824
Decision tree classifier	0.8465	0.8418
Random forest	0.9710	0.9718
Radial basis function SVM	0.9800	0.9808
K-nearest neighbors classifier	0.9710	0.9723
Random subspace K-nearest neighbors	0.9723	0.9735
Large margin K-nearest neighbors	0.9862	0.9860

6.4.5 Results: Some Users Unregistered

Table 6.3 shows the evaluation results for the biometric task when some users are unknown. The novelty detection performance of the classifiers can be observed from the outlier and inlier recall rates, which are computed by the total fraction of unregistered and registered users correctly detected. Meanwhile, the inlier accuracy is the fraction of accepted inliers that have been correctly assigned to the right users, and it estimates the traditional classification performance of the registered user palms.

As expected, the biometric recognition task with unregistered users is more difficult, and there is a significant decrease in the classification performance. NN-d and its variants clearly outperform most conventional classifiers; of the conventional classifiers, only random forest was able to have a comparable outlier and inlier recall rate. The use of large margin reprojection in large margin nearest neighbors and large margin NN-d improves the inlier recall rate and the overall accuracy. The class-specific radius in NN-d improves the overall classification accuracy with a small tradeoff in outlier detection. An increase in the overall accuracy over the conventional classifiers is also seen using the two-stage model for conventional classifiers, primarily due to better outlier detection.

Table 6.3: Evaluation of classification methods, some users unknown.

Classifier	Accuracy	Macro-F measure	Outlier recall	Inlier recall	Inlier accuracy
Naive Bayes classifier	0.7803	0.7814	0.0249	0.9812	0.9314
Linear discriminant analysis	0.8200	0.8137	0.2476	0.9423	0.9345
Decision tree classifier	0.7328	0.7301	0.0041	0.9957	0.8785
Random forest	0.8520	0.8466	0.7524	0.8726	0.8719
Radial basis function SVM	0.8244	0.8203	0.4867	0.8970	0.8919
K-NN classifier	0.8246	0.8297	0.0733	0.9944	0.9749
Large margin K-NN	0.8596	0.8598	0.2144	0.9956	0.9886
1-NN-d	0.8580	0.8568	0.7275	0.9000	0.8841
Large margin NN-d	0.8824	0.8809	0.3734	0.9876	0.9842
1-NN-d with class-specific radius optimization	0.8792	0.8785	0.6432	0.9452	0.9264
Large margin NN-d with class-specific radius	0.9041	0.9014	0.7510	0.9363	0.9347
Two-stage model for linear discriminant classifier	0.8997	0.8965	0.7510	0.9363	0.9294
Two-stage model for radial basis function SVM	0.8960	0.8946	0.7510	0.9363	0.9250

6.4.6 Discussion

The results demonstrate the usefulness of the innovations introduced in this thesis in improving the outlier detection rate as well as the overall classification accuracy.

For the recognition of gloved palms, the best accuracy achieved by our system was 95.7%, which represents a slight reduction in accuracy compared with the recognition of bare palms (96.5%). This degradation could be due to an imperfect surgical glove fit, resulting in small air pockets at the fingertips of the gloves, increasing the apparent length of those fingers. Nonetheless, the recognition accuracy is still high and is sufficient to demonstrate the viability of biometric recognition without appearance features.

In novelty detection, the performance of different methods can be interpreted in the context of the type I (rejected inliers) and type II (accepted outliers) errors. Finding a good trade-off between type I and type II errors is key in achieving a good overall accuracy. In general, the modified conventional classifiers have low outlier recall rates, which impacts their overall classification accuracy.

Large margin reprojection improves the classification accuracy for nearest neighbors, as seen in the increase in accuracy in Table 6.2. For the evaluation task with unregistered users, the class-labeling of inliers is also improved using the large margin variants of K-NN and NN-d. Therefore, large margin methods can successfully reproject the input space for superior classification. For biometric recognition with unregistered users, the effect of large margin methods is less straightforward. The primary effect of a large margin projection is to improve the recall on the *trained* samples, which improves both the inlier recall rates as well as the inlier classification accuracy; these improvements were observed for both K-nearest neighbors and NN-d using large margins. Compared with the unmodified K-nearest neighbors algorithm, large margin K-NN offers better outlier detection without compromising on inlier recall. However, outlier recall was degraded using NN-d classifier using large margins, although overall classification accuracy was still better. This is the result of a drastic trade-off between outlier and inlier recall.

To a smaller extent, this trade-off was also seen in the class-specific radius optimization, which had a modest improvement and degradation in inlier recall and outlier recall respectively. This result is expected, as the class-specific radius increases the decision boundary to improve the inlier acceptance rate at the cost of accepting outliers. However, both modifications of the NN-d resulted in improved overall classification performance, which means that the trade-off between accepting inliers and rejecting outliers was ultimately advantageous. Combining both the large margin method and the class-specific radius yields the best results overall, and there were improvements in outlier detection performance.

The two-stage model uses the large margin with class-specific radius NN-d as an outlier filter, and thus the recall rates are identical and differences in the overall classification rate are attributable to correct labeling of inlier samples. As the linear discriminant and SVM classifiers were marginally outperformed by large margin K-nearest neighbors in the registered users task, the overall classification accuracy obtained using the two-stage model is still slightly inferior to NN-d, but nonetheless vastly superior to the original conventional classifiers.

The results demonstrate a promising option for biometric recognition using Kinect depth images, and the biometric recognition rate approaches that of state-of-the-art approaches with more constraints on hand pose or using more sensitive imaging sensors. However, our system is calibrated for a smaller base of registered users and is more suitable for biometric recognition rather than dedicated biometric verification. For the dynamic scenario with the possibility of unknown users, the unknown user rejection rate of 75.1% offers a good chance of detecting unknown guest users in practical settings with few non-registered users.

Lastly, to further validate the innovations introduced in this paper for novelty detection, we conducted additional experiments on other datasets.

As these datasets are not related to the problem of biometric recognition, the results to these additional experiments are contained in Appendix D.

6.5 Summary

In this chapter, biometric recognition was proposed for user identification to perform context-selection in a surgical computer interface. Depth information from Kinect is used to construct scale-invariant features for the classification of users. For the detection of unregistered users, large margin NN-d is proposed to increase the class separation and the classification accuracy. In addition, a class-specific radius is proposed to modify the classifier decision boundaries to obtain a better trade-off between inlier acceptance and outlier rejection. The one-class classifier system is able to correctly recognize and classify palms of previously registered users while rejecting unknown and unregistered users, demonstrating that novelties introduced are useful for learning the properties of unknown distributions. The biometric recognition results were comparable to state-of-the-art approaches and are promising for detecting unregistered users.

CHAPTER 7

Conclusion and Future Work

Computational intelligence and machine learning continue to play increasingly important roles in medical analysis and visualization. This dissertation has introduced several novel computational intelligence approaches to address problems in medicine.

In Chapter 3, we described an ensemble-based method for diagnosing osteopenia. The weighted decision ensemble exploits classifiers that are discriminative towards specific classes by using a novel combiner function. The weights of the decision ensemble are optimized using a GA scheme, ensuring that the final ensemble has the greatest accuracy and class separation. These contributions allow for a more robust and accurate diagnosis of osteopenia from CT scans of lumbar vertebrae.

In Chapters 2 and 4, regression was used to predict a patient's BMD from dCT images. A filtering-based ensemble technique is applied to solve a regression problem on a multimodal medical dataset with high relative dimensionality. By choosing a set of regressors from several candidate regressors such that the component regressors are diverse and uncorrelated, the regression ensemble reduces the influence of spurious features and noisy data samples. Compared with simple multivariate regression, the ensemble regression approach is more powerful and robust, and yields better results on multimodal medical datasets.

In Chapter 5, mean shift clustering was applied to detect and group voxels with similar properties in the LH domain; each cluster was representative of a structure or material boundary. The extracted boundaries were subsequently improved using a region growing algorithm to smooth the boundaries. Lastly, TFs were automatically designed to reduce occlusions by considering the relative sizes and distances between clusters. Because mean shift clustering is non-parametric, clusters corresponding to material boundaries can be identified automatically with little parameter tuning. The proposed system therefore allows visualizations comparable to state-of-the art approaches to be generated while reducing the amount of manual labor required.

In Chapter 6, users were identified using biometric recognition based on depth images of the palm captured using Microsoft Kinect, and the user identities were used to customize the work-interfaces specific to each user. When no unregistered users were expected, good accuracies ($\geq 95\%$) were attained by standard classification algorithms, with the best algorithms achieving a recognition rate comparable to state-of-the-art biometric recognition algorithms. For detecting unregistered users, one-class classifiers such as NN-d obtained better outlier detection rates and overall classification accuracy. By projecting the data using a large margins method for NN-d and adding a cluster-specific radius to the decision boundary, the modified NN-d algorithm offered the best trade-off between outlier rejection and classification accuracy.

7.1 Future Work

In this section, we propose several areas where the thesis work can be expanded upon.

7.1.1 Classifier Design for Osteopenia Diagnosis

While classification is capable of diagnosing a disease condition, the black-box nature of most classifiers means that it is usually not possible to describe the rationale behind a machine diagnosis. Even rule-based classifier systems, such as decision trees, generate complex rules which are difficult for a human to understand. This impacts the confidence of the medical community in any black-box machine learning diagnosis system, and makes it difficult for any machine learning diagnosis system to be adopted. Furthermore, it is difficult to extract any useful insight into the disease condition based on the black-box. However, it is possible to process the classifier ensemble in Chapter 3 such that the ensemble decision is more comprehensible to an expert user. Fundamentally, the classifier ensemble is a combination of several basic classifiers trained on different features modes, which different weights assigned to each classifier based on its importance and relevance. We can organize the ensemble by grouping the basic classifiers according to the features they are trained on. Then, when a diagnosis is made, the net contribution of each set of feature classifiers can be calculated and presented. Thus, the ensemble decision is augmented with the feature-wise breakdown behind the decision, allowing the clinician to determine which disease symptoms are most prominent.

7.1.2 Bone Mineral Density Prediction

A problem common to regression techniques for prediction is the tendency for large errors when predicting extreme or outlier values. This problem arises because small errors in the estimation of the slope accumulate to large errors when the data point is far from the training space. While these errors may not be important for medical diagnosis, as outlier points are far away from the decision boundary and their class labels are unaffected by large absolute errors, this issue should not be neglected. One concern is that in multivariate data, the influence of large outliers in one or a few feature dimensions may result in large overall regression errors. We propose a simple modification to our ensemble regression scheme to reduce the impact of large outliers in some feature dimensions. The detection of outliers is simple, as they lie outside the typical range of values for a given feature. For a given feature dimension with an outlier, we cap that feature's contribution to a known maximum/minimum; the cap value can be determined by analyzing the range of feature values seen in the training data. This modification ensures that ensemble regression occurs entirely in the operating range that it has been trained on, and reduces the impact of outliers.

7.1.3 Automated Transfer Function Design

The clustering-based transfer function design method is general and can be applied to complex volumetric datasets from different sources. However, to obtain better visualization results, we can specialize to focus on medical datasets. By including domain knowledge, such as the typical voxel intensities of well-defined tissues like bone, instances of spurious clusters or mis-merged clusters can be reduced. Domain knowledge can be built using insights from domain experts, or by using machine learning to extract the properties of recurring anatomical structures in medical volumes.

Computational intelligence can also be applied to obtain more precise cluster segmentations, thus improve the sharpness and crispness of material and structure boundaries in medical visualizations. A voxel-wise classification on the material boundaries can be performed by comparing each boundary voxel to the labels of neighboring voxels. Compared to the region-growing heuristic used in subsection 5.2.3, the machine learning approach should yield crisper edges. However, a large corpus of labeled medical volumes and visualizations is required for a machine learning approach to be viable.

7.1.4 Biometric Recognition

To increase the reliability of palm biometric recognition, the information from multiple sequential frames can be combined to allow for the palm identity to be refined across multiple frames; this reduces the impact of sensor errors or motion-induced artifacts, but also introduces a time delay depending on the number of frames used. The simplest implementation of this idea is to classify each individual frame and to take the majority label. Another possibility is to build a palm image by registering across the sequential frames, and to perform classification on this refined palm image.

The user palm recognition system can also be extended to allow for a fast-registration mode where a new user can be quickly granted access to the system. The advantage of one-class classifiers is that the training of each class is independent of all other classes; thus, the addition of new users does not require all the classifier decision boundaries to be recomputed. However, if large margins NN-d is used, then the large margins projection becomes increasingly unsuitable as more new users are added; a new projection matrix will need to be recomputed to include the newly registered users.

In a broader context, the novelty detection algorithms could be applied to detect atypical samples in medical screening without necessarily training on a specific disease condition. This can reduce the requirement for diseased cases in medical studies, as diseased cases are typically much rarer than healthy cases.

Bibliography

[1] G. D. Magoulas and A. Prentza, “Machine learning in medical applications,” in *Machine Learning and Its Applications.* Springer, 2001, pp. 300–307.

[2] N. Savage, “Better medicine through machine learning,” *Communications of the ACM*, vol. 55, no. 1, pp. 17–19, 2012.

[3] M. Wernick, Y. Yang, J. Brankov, G. Yourganov, and S. Strother, “Machine learning in medical imaging,” *Signal Processing Magazine, IEEE*, vol. 27, no. 4, pp. 25–38, 2010.

[4] “Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis,” *American Journal of Medicine*, vol. 94, pp. 646–650, Jun 1993.

[5] “Osteoporosis prevention, diagnosis, and therapy,” *Journal of the American Medical Association*, vol. 285, pp. 785–795, Feb 2001.

[6] M. Revilla, J. L. Cardenas, E. R. Hernndez, L. F. Villa, and H. Rico, “Correlation of total-body bone mineral content determined by dual-

energy x-ray absorptiometry with bone mineral density determined by peripheral quantitative computed tomography," *Academic Radiology*, vol. 2, no. 12, pp. 1062 – 1066, 1995.

[7] J. Schreiber, P. Anderson, G. Humberto, A. Buchholz, and A. Au, "Hounsfield units for assessing bone mineral density and strength: A tool for osteoporosis management," *The Journal of Bone and Joint Surgery (American)*, vol. 93, no. 11, pp. 1057–1063, 2011.

[8] S. Boden, D. Goodenough, C. Stockham, E. Jacobs, T. Dina, and R. Allman, "Precise measurement of vertebral bone density using computed tomography without the use of an external reference phantom," *Journal of Digital Imaging*, vol. 2, pp. 31–38, 1989.

[9] T. Link, B. Koppers, T. Licht, J. Bauer, Y. Lu, and E. Rummeny, "In vitro and in vivo spiral CT to determine bone mineral density: Initial experience in patients at risk for osteoporosis," *Radiology*, vol. 231, no. 3, p. 805, 2004.

[10] S. Teoh and C. Chui, "Bone material properties and fracture analysis: Needle insertion for spinal surgery," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 1, no. 2, pp. 115 – 139, 2008.

[11] J. Rho, M. Hobatho, and R. Ashman, "Relations of mechanical properties to density and CT numbers in human bone," *Medical*

Engineering Physics, vol. 17, no. 5, pp. 347 – 355, 1995.

[12] T. Baum, J. Carballido-Gamio, M. Huber, D. Müller, R. Monetti, C. Räth, F. Eckstein, E. Lochmüller, S. Majumdar, E. Rummeny *et al.*, “Automated 3D trabecular bone structure analysis of the proximal femur - prediction of biomechanical strength by CT and DXA,” *Osteoporosis International*, pp. 1–12, 2010.

[13] J. Bauer, T. Henning, D. Mueller, Y. Lu, S. Majumdar, and T. Link, “Volumetric quantitative CT of the spine and hip derived from contrast-enhanced MDCT: conversion factors,” *American Journal of Roentgenology*, vol. 188, no. 5, p. 1294, 2007.

[14] A. H. Habashy, X. Yan, J. K. Brown, X. Xiong, and S. C. Kaste, “Estimation of bone mineral density in children from diagnostic CT images: A comparison of methods with and without an internal calibration standard,” *Bone*, vol. 48, no. 5, pp. 1087 – 1094, 2011.

[15] P. Pickhardt, L. Lee, A. del Rio, T. Lauder, R. Bruce, R. Summers, B. Pooler, and N. Binkley, “Simultaneous screening for osteoporosis at CT colonography: Bone mineral density assessment using MDCT attenuation techniques compared against the DXA reference standard,” *Journal of Bone and Mineral Research*, vol. 26, no. 9, pp. 2194–2203, 2011.

[16] H. Gudmundsdottir, B. Jonsdottir, S. Kristinsson, A. Johannesson, D. Goodenough, and G. Sigurdsson, "Vertebral bone density in Icelandic women using quantitative computed tomography without an external reference phantom," *Osteoporosis International*, vol. 3, no. 2, pp. 84–89, 1993.

[17] D. Mueller, A. Kutscherenko, H. Bartel, A. Vlassenbroek, P. Ourednicek, and J. Erckenbrecht, "Phantom-less QCT BMD system as screening tool for osteoporosis without additional radiation," *European Journal of Radiology*, vol. 79, no. 3, pp. 375–381, 2010.

[18] S. Hui, V. Weir, K. Brown, and J. Froelich, "Assessing the clinical utility of quantitative computed tomography with a routinely used diagnostic computed tomography scanner in a cancer center," *Journal of Clinical Densitometry*, vol. 14, no. 1, pp. 41–46, 2011.

[19] W. Li, J. Kornak, T. Harris, J. Keyak, C. Li, Y. Lu, X. Cheng, and T. Lang, "Identify fracture-critical regions inside the proximal femur using statistical parametric mapping," *Bone*, vol. 44, no. 4, pp. 596–602, 2009.

[20] A. Valentinitsch, J. Patsch, D. Mueller, F. Kainberger, and G. Langs, "Texture analysis in quantitative osteoporosis assessment: Characterizing microarchitecture," in *Biomedical Imaging: From*

Nano to Macro, 2010 IEEE International Symposium, 2010, pp. 1361–1364.

[21] W.-L. Tay, C.-K. Chui, S.-H. Ong, and A. C.-M. Ng, "Detection of osteopenia from routine CT images," in *The 7th Asian Conference on Computer-Aided Surgery*, Bangkok, August 26-27 2011.

[22] E. Lewiecki and N. Watts, "Assessing response to osteoporosis therapy," *Osteoporosis International*, vol. 19, pp. 1363–1368, 2008.

[23] J. Lotz, E. Cheal, and W. Hayes, "Fracture prediction for the proximal femur using finite element models: Part I - linear analysis," *Journal of Biomechanical Engineering*, vol. 113, p. 353, 1991.

[24] D. Marshall, O. Johnell, and H. Wedel, "Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures," *BMJ*, vol. 312, no. 7041, p. 1254, 1996.

[25] World Health Organization, *Assessment of fracture risk and its application to screening for postmenopausal osteoporosis.* World Health Organization, 1994, no. 843.

[26] B. Richmond, "Dxa scanning to diagnose osteoporosis: Do you know what the results mean?" *Cleveland Clinic Journal of Medicine*, vol. 70, no. 4, pp. 353–360, 2003.

[27] Y. Boykov and O. Veksler, “Graph cuts in vision and graphics: Theories and applications,” *Handbook of Mathematical Models in Computer Vision, Springer-Verlag*, pp. 79–96, 2006.

[28] Y. Li, J. Sun, C. Tang, and H. Shum, “Lazy snapping,” in *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3. ACM, 2004, pp. 303–308.

[29] D. Reynolds, “Gaussian mixture models,” *Encyclopedia of Biometric Recognition*, 2008.

[30] G. Hounsfield, “Computed medical imaging,” *Medical Physics*, vol. 7, p. 283, 1980.

[31] M. Fischler and R. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[32] K. G. Faulkner and E. Orwoll, “Implications in the use of T-scores for the diagnosis of osteoporosis in men,” *Journal of Clinical Densitometry*, vol. 5, no. 1, pp. 87 – 93, 2002.

[33] J. Zhang, C. Yan, C. Chui, and S. Ong, “Accurate measurement of bone mineral density using clinical CT imaging with single energy

beam spectral intensity correction," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 7, pp. 1382–1389, 2010.

[34] O. Svendsen, C. Hassager, V. Skødt, and C. Christiansen, "Impact of soft tissue on in vivo accuracy of bone mineral measurements in the spine, hip, and forearm: a human cadaver study," *Journal of Bone and Mineral Research*, vol. 10, no. 6, pp. 868–873, 1995.

[35] J. Carrino and L. Ohno-Machado, "Development of radiology prediction models using feature analysis," *Academic Radiology*, vol. 12, no. 4, pp. 415–421, 2005.

[36] A. D'Costa and A. Sayeed, "Data versus decision fusion for classification in sensor networks," in *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, 2003.

[37] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[38] Y. Mehmood, M. Ishtiaq, M. Tariq, and M. Arfan Jaffar, "Classifier ensemble optimization for gender classification using genetic algorithm," in *Information and Emerging Technologies (ICIET), 2010 International Conference on*, 2010, pp. 1 –5.

[39] D. W. Opitz and J. W. Shavlik, "Generating accurate and diverse members of a neural-network ensemble," in *Advances in Neural Information Processing Systems.* MIT Press, 1996, pp. 535–541.

[40] V. Tresp and M. Taniguchi, "Combining estimators using non-constant weighting functions," in *Advances in Neural Information Processing Systems 7.* MIT Press, 1995, pp. 419–426.

[41] P. Derbeko, R. El-Yaniv, and R. Meir, "Variance optimized bagging," in *In ECML 2002.* Springer-Verlag, 2002, pp. 60–71.

[42] W. L. Buntine, "A theory of learning classification rules," 1992.

[43] L. Rokach, *Pattern Classification using Ensemble Methods.* World Scientific, 2009.

[44] Z. Wu, C. hung Li, and V. Cheng, "Large margin maximum entropy machines for classifier combination," in *Wavelet Analysis and Pattern Recognition, 2008. ICWAPR '08. International Conference on*, vol. 1, 2008, pp. 378 –383.

[45] S. Shlien, "Multiple binary decision tree classifiers," *Pattern Recognition*, vol. 23, no. 7, pp. 757 – 763, 1990.

[46] Z.-H. Zhou, J. Wu, and W. Tang, “Ensembling neural networks: Many could be better than all,” *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239 – 263, 2002.

[47] A. Majid, A. Khan, and A. Mirza, “Intelligent combination of kernels information for improved classification,” in *Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on*, 2005, p. 6 pp.

[48] Z.-H. Zhou, W. Tang, Z. hua Zhou, and W. Tang, “Selective ensemble of decision trees,” in *Lecture Notes in Artificial Intelligence*. Springer, 2003, pp. 476–483.

[49] M. T. Miller, A. K. Jerebko, J. D. Malley, and R. M. Summers, “Feature selection for computer-aided polyp detection using genetic algorithms,” in *Medical Imaging 2003: Physiology and Function: Methods, Systems, and Applications, Proceedings of the SPIE*, vol. 5031.

[50] M. Pei, E. D. Goodman, W. F. P. Iii, and Y. Ding, “Genetic algorithms for classification and feature extraction,” in *Annual Meeting, Classification Society of North America*, 1995.

[51] R. Xu and L. He, “Gacem: Genetic algorithm based classifier ensemble in a multi-sensor system,” *Sensors*, vol. 8, no. 10, pp. 6203–6224,

2008.

[52] R. O. Duda, P. E.Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

[53] W.-L. Tay, C.-K. Chui, S.-H. Ong, and A. C.-M. Ng, "Osteopenia screening using areal bone mineral density estimation from diagnostic CT images," *Academic Radiology*, vol. 19, no. 10, pp. 1273–1282, 2012.

[54] D. Mantzaris, G. Anastassopoulos, L. Iliadis, K. Kazakos, and H. Papadopoulos, "A soft computing approach for osteoporosis risk factor estimation," *Artificial Intelligence Applications and Innovations*, pp. 120–127, 2010.

[55] A. Akgundogdu, R. Jennane, G. Aufort, C. Benhamou, and O. Ucan, "3d image analysis and artificial intelligence for bone disease classification," *Journal of Medical Systems*, vol. 34, no. 5, pp. 815–828, 2010.

[56] J. Serrano, M. Tomeckova, and J. Zvarova, "Machine learning methods for knowledge discovery in medical data on atherosclerosis," *European Journal for Biomedical Informatics*, vol. 2, no. 1, pp. 6–33, 2006.

[57] H.-P. Chan, B. Sahiner, and L. Hadjiiski, "Sample size and validation issues on the development of cad systems," *International Congress Series*, vol. 1268, pp. 872 – 877, 2004, cARS 2004 - Computer Assisted Radiology and Surgery. Proceedings of the 18th International Congress and Exhibition.

[58] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2-3, pp. 427 – 436, 2008, advances in Neural Networks Research: IJCNN '07, 2007 International Joint Conference on Neural Networks IJCNN '07.

[59] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, 1991.

[60] B. Brumen, M. Jurič, T. Welzer, I. Rozman, H. Jaakkola, and A. Papadopoulos, "Assessment of classification models with small amounts of data," *Informatica*, vol. 18, no. 3, pp. 343–362, 2007.

[61] M. Mazurowski, P. Habas, G. Tourassi, and J. Zurada, "Impact of low class prevalence on the performance evaluation of neural network

based classifiers: Experimental study in the context of computer-assisted medical diagnosis," in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, aug. 2007, pp. 2005 –2009.

[62] Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data mining on imbalanced data sets," in *Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on.* Ieee, 2008, pp. 1020–1024.

[63] H. Moon, H. Ahn, R. Kodell, S. Baek, C. Lin, and J. Chen, "Ensemble methods for classification of patients for personalized medicine with high-dimensional data," *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 197–207, 2007.

[64] H. Lo, C. Chang, T. Chiang, C. Hsiao, A. Huang, T. Kuo, W. Lai, M. Yang, J. Yeh, C. Yen *et al.*, "Learning to improve area-under-froc for imbalanced medical data classification using an ensemble method," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 2, pp. 43–46, 2008.

[65] L. Nanni, A. Lumini, and S. Brahnam, "A classifier ensemble approach for the missing feature problem," *Artificial Intelligence in Medicine*, 2011.

[66] B. Antal, I. Lázár, A. Hajdu, Z. Torok, A. Csutak, and T. Peto, "A multi-level ensemble-based system for detecting microaneurysms in fundus images," in *Soft Computing Applications (SOFA), 2010 4th International Workshop on*. IEEE, 2010, pp. 137–142.

[67] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[68] P. Cunningham and J. Carney, "Diversity versus quality in classification ensembles based on feature selection," *Machine Learning: ECML 2000*, pp. 109–116, 2000.

[69] P. Delmas, R. Eastell, P. Garnero, M. Seibel, and J. Stepan, "The use of biochemical markers of bone turnover in osteoporosis," *Osteoporosis International*, vol. 11, no. 18, pp. 2–17, 2000.

[70] J. Langlois, C. Rosen, M. Visser, M. Hannan, T. Harris, P. Wilson, and D. Kiel, "Association between insulin-like growth factor i and bone mineral density in older women and men: the framingham heart study," *Journal of Clinical Endocrinology & Metabolism*, vol. 83, no. 12, pp. 4257–4262, 1998.

[71] E. Kurland, C. Rosen, F. Cosman, D. McMahon, F. Chan, E. Shane, R. Lindsay, D. Dempster, and J. Bilezikian, "Insulin-like growth

factor-i in men with idiopathic osteoporosis," *Journal of Clinical Endocrinology & Metabolism*, vol. 82, no. 9, pp. 2799–2805, 1997.

[72] P. Šereda, A. V. Bartroli, I. W. O. Serlie, and F. A. Gerritsen, "Visualization of boundaries in volumetric data sets using LH histograms," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 208–218, 2006.

[73] G. Kindlmann and J. W. Durkin, "Semi-automatic generation of transfer functions for direct volume rendering," in *Proceedings of IEEE Symposium on Volume Visualization*, 1998, pp. 79–86.

[74] J. Kniss, G. Kindlmann, and C. Hansen, "Interactive volume rendering using multi-dimensional transfer functions and direct manipulation widgets," in *Proceedings of IEEE Symposium on Volume Visualization*, 2001, pp. 255–262.

[75] J. Kniss, G. Kindlmann, and C. Hansen, "Multi-dimensional transfer functions for interactive volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 3, pp. 270–285, 2002.

[76] E. B. Lum and K.-L. Ma, "Lighting transfer functions using gradient aligned sampling," in *Proceedings of IEEE Visualization*, 2004, pp. 289–296.

[77] J.-S. Praßni, T. Ropinski, and K. H. Hinrichs, “Efficient boundary detection and transfer function generation in direct volume rendering,” in *Proceedings of the 14th International Fall Workshop on Vision, Modeling, and Visualization (VMV09)*, 2009, pp. 285–294.

[78] F.-Y. Tzeng and K.-L. Ma, “A cluster-space visual interface for arbitrary dimensional classification of volume data,” in *Proceedings of IEEE/Eurographics Symposium on Visualization*, 2004, pp. 17–24.

[79] P. Šereda, A. Vilanova, and F. A. Gerritsen, “Automating transfer function design for volume rendering using hierarchical clustering of material boundaries,” in *Proceedings of IEEE/Eurographics Symposium on Visualization*, 2006, pp. 243–250.

[80] R. Maciejewski, W. Chen, I. Woo, and D. S. Ebert, “Structuring feature space - a non-parametric method for volumetric transfer function generation,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1473–1480, 2009.

[81] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi, “Efficient volume exploration using the gaussian mixture model,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 11, pp. 1560–1573, 2011.

[82] C. Y. Ip, A. Varshney, and J. JaJa, “Hierarchical exploration of volumes using multilevel segmentation of the intensity-gradient histograms,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2355 –2363, dec. 2012.

[83] Y. Wang, J. Zhang, D. J. Lehmann, H. Theisel, and X. Chi, “Automating transfer function design with valley cell-based clustering of 2d density plots,” in *Computer Graphics Forum*, vol. 31, no. 3pt4. Wiley Online Library, 2012, pp. 1295–1304.

[84] D. Hong, G. Ning, T. Zhao, M. Zhang, and X. Zheng, “Method of normal estimation based on approximation for visualization,” *Journal of Electronic Imaging*, vol. 12, no. 3, pp. 470–477, 2003.

[85] R. Huang and K.-L. Ma, “RGVis: Region growing based techniques for volume visualization,” in *Proceedings of Pacific Conference on Computer Graphics and Applications*, 2003, pp. 355–363.

[86] A. Andrew, “Another efficient algorithm for convex hulls in two dimensions,” *Information Processing Letters*, vol. 9, no. 5, pp. 216–219, 1979.

[87] G. K. Ong Michael, T. Connie, and A. B. Jin Teoh, “Touch-less palm print biometrics: Novel design and implementation,” *Image and Vision Computing*, vol. 26, no. 12, pp. 1551–1560, 2008.

[88] G. Lu, D. Zhang, and K. Wang, “Palmprint recognition using eigenpalms features,” *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1463–1467, 2003.

[89] X. Wu, D. Zhang, and K. Wang, “Fisherpalms based palmprint recognition,” *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2829–2838, 2003.

[90] T. Connie, A. T. B. Jin, M. G. K. Ong, and D. N. C. Ling, “An automated palmprint recognition system,” *Image and Vision computing*, vol. 23, no. 5, pp. 501–515, 2005.

[91] A. Kumar, D. C. Wong, H. C. Shen, and A. K. Jain, “Personal verification using palmprint and hand geometry biometric,” in *Audio-and Video-Based Biometric Person Authentication.* Springer, 2003, pp. 668–678.

[92] C. Schwarz and N. d. V. Lobo, “Segment-based hand pose estimation,” in *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on.* IEEE, 2005, pp. 42–49.

[93] R. Sanchez-Reillo, C. Sanchez-Avila, and A. Gonzalez-Marcos, “Biometric identification through hand geometry measurements,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1168–1171, 2000.

[94] R. M. Luque-Baena, D. Elizondo, E. LóPez-Rubio, E. J. Palomo, and T. Watson, “Assessment of geometric features for individual identification and verification in biometric hand systems,” *Expert Systems with Applications*, 2012.

[95] D. M. Tax, “One-class classification,” Ph.D. dissertation, Technical University of Delft, 2001.

[96] M. Markou and S. Singh, “Novelty detection: a reviewpart 1: statistical approaches,” *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[97] G. Manson, S. G. Pierce, K. Worden, T. Monnier, P. Guy, and K. Atherton, “Long-term stability of normal condition data for novelty detection,” in *SPIE’s 7th Annual International Symposium on Smart Structures and Materials.* International Society for Optics and Photonics, 2000, pp. 323–334.

[98] J. Laurikkala, M. Juhola, E. Kentala, N. Lavrac, S. Miksch, and B. Kavsek, “Informal identification of outliers in medical data,” in *Proceedings of the 5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, 2000, pp. 20–24.

[99] T. Odin and D. Addison, “Novelty detection using neural network technology,” in *Proceedings of the COMADEN Conference*, 2000.

[100] L. Tarassenko, D. A. Clifton, P. R. Bannister, S. King, and D. King, "Novelty detection," *Encyclopedia of Structural Health Monitoring*, 2009.

[101] C. M. Bishop, "Novelty detection and neural network validation," in *Vision, Image and Signal Processing, IEEE Proceedings-*, vol. 141, no. 4. IET, 1994, pp. 217–222.

[102] D.-Y. Yeung and C. Chow, "Parzen-window network intrusion detectors," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4. IEEE, 2002, pp. 385–388.

[103] O. Mazhelis, "One-class classifiers: a review and analysis of suitability in the context of mobile-masquerader detection." *South African Computer Journal*, vol. 36, pp. 29–48, 2006.

[104] V. N. Vapnik, "Statistical learning theory," 1998.

[105] A. Ypma and R. P. Duin, "Support objects for domain approximation," in *Proceedings of International Conference on Artificial Neural Networks*, 1998.

[106] Z. Mo and U. Neumann, "Real-time hand pose recognition using low-resolution depth images," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1499–1505.

[107] J. Blitzer, K. Q. Weinberger, and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” in *Advances in Neural Information Processing Systems*, 2005, pp. 1473–1480.

[108] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[109] Y. Yang and X. Liu, “A re-examination of text categorization methods,” in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1999, pp. 42–49.

[110] I. Cohen, Q. Xiang, X. Sean Zhou, Z. Thomas, and T. Huang, “Feature selection using principal feature analysis,” *ICIP’02*, 2002.

[111] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, “Human detection using partial least squares analysis,” in *Computer Vision, 2009 IEEE 12th International Conference on.* IEEE, 2009, pp. 24–31.

[112] K. Bache and M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[113] E. Alpaydin and F. Alimoglu, “Pen-based recognition of handwritten digits data set,” 2013. [Online].

Available: http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits

[114] C. Brodley, “Image segmentation data set,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Image+Segmentation

[115] A. Srinivasan, “Statlog (landsat satellite) data set,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)

APPENDIX A

Vertebral Anatomy

Fig. A.1 presents a 2-D view of a typical lumbar vertebra. The vertebral body is the main weight-bearing structure of the vertebra. The vertebral body can be segmented based on nearby anatomical landmarks, such as the spinal canal which houses the spinal cord.

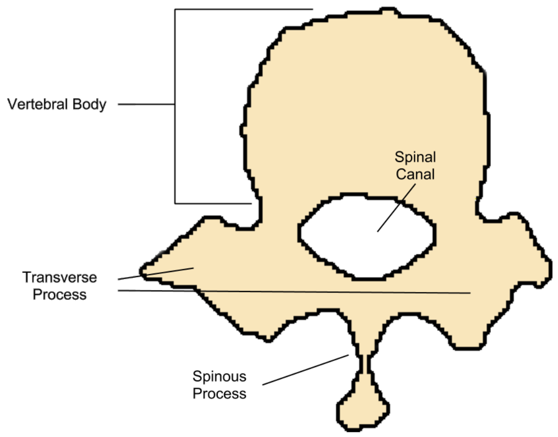


Figure A.1: A lumbar vertebra, with spinal processes and vertebral body labeled.

APPENDIX B

Dual-energy X-ray Absorptiometry

Dual-energy X-ray absorptiometry (DXA, also known as DEXA) is the most prevalent technology for bone density measurement, and is primarily used in the diagnosis and following of osteoporosis in the spine and hip. DXA provides a measurement of the bone mineral density (BMD) which provides an indicator to the bone strength and fracture risk. DXA operates by radiating two X-ray beams with different energy levels at skeletal sites; because the two X-ray beams possess different energies, they are attenuated at different rates by bone [26]. After subtracting the contribution of soft tissue absorption, the mineral content contained within each bone can be determined based on the absorption rates of each beam by bone. This BMC is subsequently normalized by the projected bone's area to obtain the aBMD.

As DXA is the most widely-studied bone measurement technology, it is used in the WHO's definition for osteoporosis [25]. The aBMD measurement from DXA is compared to a reference population to generate a score for diagnosis. For osteoporosis screening, the T-score is used, where the reference population is a healthy 30-year-old white female (WHO recommendation) or a healthy 30-year-old of the same ethnicity and sex (US standard). Three conditions are defined based on the T-score, where a T-score of -1.0 represents an aBMD that is one standard deviation below the mean for the reference population:

1. **Normal**, with normal risk of fracture, defined as a T-score of -1.0 or higher.

2. **Osteopenia**, with low bone mass and considered a precursor to osteoporosis, defined as a T-score of between -1.0 and -2.5.

3. **Osteoporosis**, with increased risk of fracture, defined as a T-score of -2.5 or lower.

The T-score definition of osteoporosis is typically applied for osteoporosis screening in post-menopausal women and men of over age 50. For other patient groups where osteoporosis is normally infrequent, such as premenopausal women, men below 50, and children, the Z-score is applied instead to screen for severe osteoporosis. The Z-score is calculated against a matched reference population of the same age, sex, and ethnicity. A low Z-score (-1.5) can be an indicator of metabolic bone disease and justify for further evaluation for osteoporosis [26]. However, because different reference populations have different fracture risks, the Z-score may provide a misleading picture of the actual fracture risk.

APPENDIX C

Linear Regression Methods

In this appendix, we discuss the theory and methods for linear regression, as well as feature selection and data transformation techniques that can improve regression performance on large datasets.

The objective is to relate the multimodal data matrix X with the aBMD from CT, Y, through a set of linear constants k. It is also desired to perform feature selection on the data, such that only $f \times s$ features are used. This feature selection method is known as the *filter* method, where the subset of chosen features is selected as a pre-processing step independent of the chosen classifiers.

C.1 Linear Least Squares Regression

The simplest way to relate the target variable Y and the data matrix X is to use linear least squares. A set of constants k is assumed to relate the two variables, with some residual error e.

$$Y = Xk + e. \tag{C.1.1}$$

To recover the least squares solution, the sum of squared errors is minimized using the pseudoinverse,

$$k = (X^T X)^{-1} X^T Y. \tag{C.1.2}$$

To reduce the impact of noise on the regression constants, Tikhonov regularization is used. A regularization term λ, with an experimentally-determined value of 0.5, is included in the pseudoinverse,

$$k = (X^T X + \lambda I)^{-1} X^T Y. \tag{C.1.3}$$

The linear regression solution typically involves all features of X, not all of which are useful for determining Y. Some features of X can be discarded to increase the robustness of the regression on unknown data. The components of k with the smallest magnitudes contribute the least to the regression, and discarding them does not have a large impact on the final result. Let X_{fs} be the data matrix X where all columns corresponding to the features with the fs smallest absolute components in k are set to zero. Then, to compensate for the removed features, a new set of constants k_{llsfs} is computed

$$k_{\text{llsfs}} = (X_{fs}^T X_{fs} + \lambda I)^{-1} X_{fs}^T Y. \tag{C.1.4}$$

C.2 Principal Components Regression

In principal components regression (PCR), principal components analysis is first used to obtain a set of principal components, P, for the data. The principal components describe the maximum variation possible that describes the original data matrix X. If the singular value decomposition of X is

$$X = W\Sigma V^T, \tag{C.2.1}$$

where the $m \times m$ matrix W is the matrix of eigenvectors of the covariance matrix XX^T, the matrix Σ is an $m \times n$ rectangular diagonal matrix with nonnegative real numbers on the diagonal, and the $n \times n$ matrix V is the matrix of eigenvectors of X^TX, then the PCA transformation of X is given by

$$X_{\text{PCA}} = V\Sigma^T. \tag{C.2.2}$$

In PCR, the principal components with the largest eigenvalues are used to form a regression to the target variable. Assuming that the matrix formed by retaining the fs columns in X_{PCA} corresponding to the largest eigenvalues in W is X_{PCR}, then the linear regression components k_{PCR} are

$$k_{\text{PCR}} = W_{fs}(X_{PCR}^T X_{PCR} + \lambda I)^{-1} X_{PCR}^T Y. \tag{C.2.3}$$

C.3 Principal Feature Analysis

Principal feature analysis (PFA) [110] is an algorithm based on PCR. However, the principal components created by PCR span over the entire set of features in the original data, hence it is not a feature selection technique. PFA imposes a feature selection condition during the construction of the principal components to restrict the number of features used.

For PFA, the principal components V and eigenvalues of the X are first computed. Construct the vectors W by taking the rows of V; therefore W should contain as many vectors as there are dimensions in X. $|W|$ is clustered using k-means with q clusters, where q is chosen depending on the amount of data variability to be retained. For each cluster, the vector W closest to the cluster mean is computed, and the corresponding feature is chosen as a principal feature. There are therefore q principal features. Lastly, linear regression is performed on the set of principal features.

C.4 Partial Least Squares Regression

Partial least squares regression (PLS) is a method that has recently been used for computer vision[111]. PLS attempts to decompose X into a set of latent variables that are highly correlated with Y, and to then regress Y based on the latent variables.

$$X = TP^T + E_1. \tag{C.4.1}$$

$$Y = Uq^T + e_2. \tag{C.4.2}$$

T and U are the matrices containing the extracted latent vectors, while P and q represent the loadings. E_1 and e_2 are the residual errors. T and U are then constructed using iterative PLS algorithm, where

$$[\mathrm{cov}(t_i, u_i)]^2 = \max_{|w_i|=1} [\mathrm{cov}(Xw_i, y)]^2. \tag{C.4.3}$$

Since the latent vectors are orthogonal and uncorrelated, there are at most $rank(X)$ latent vectors.

APPENDIX D

Additional Experiments for NN-d Validation

To validate the modified NN-d algorithms, we perform additional experiments on different test datasets from the UCI Machine Learning Repository [112]. The datasets chosen had a similar number of classes to our earlier experiments.

Table D.1: Test datasets used

Name	Samples	Classes	Features
pendigits [113]	10992	10	16
segmentation [114]	2100	7	19
Statlog [115]	6435	7	36

The results show that the proposed NN-d classifiers also improve the overall accuracy on other datasets. In particular, the two-stage model achieves the best results for the Statlog dataset (Table D.4), demonstrating that a combination of a NN-d outlier filter and a conventional classification algorithm can outperform either component by itself. These results agree with our earlier findings on the biometric recognition problem.

Table D.2: Evaluation results on pendigits dataset

Classifier	Accuracy	Macro-F measure	Outlier recall	Inlier recall	Inlier accuracy
Naive Bayes classifier	0.7856	0.7210	0.1078	0.9684	0.8609
Linear discriminant analysis	0.8001	0.7317	0.2595	0.9384	0.8602
Decision tree classifier	0.8660	0.7878	0.0179	0.9956	0.9602
Random forest	0.9171	0.8301	0.2958	0.9906	0.9862
Radial basis function SVM	0.7605	0.6866	0.7125	0.7667	0.7658
K-NN classifier	0.9027	0.8160	0.0795	0.9993	0.9941
Large margin K-NN	0.9037	0.8173	0.0830	0.9993	.9949
1-NN-d	0.9392	0.8473	0.4597	0.9967	0.9924
Large margin NN-d	0.9482	0.8552	0.8000	0.9673	0.9647
1-NN-d with class-specific radius optimization	0.9439	0.8514	0.5985	0.9860	0.9822
Large margin NN-d with class-specific radius	0.9447	0.8520	0.6123	0.9846	0.9816
Two-stage model for linear discriminant classifier	0.8503	0.7785	0.6123	0.9846	0.8767
Two-stage model for radial basis function SVM	0.9442	0.8521	0.6123	0.9850	0.9811

Table D.3: Evaluation results on segmentation dataset

Classifier	Accuracy	Macro-F measure	Outlier recall	Inlier recall	Inlier accuracy
Naive Bayes classifier	0.7873	0.4708	0.0283	0.9747	0.9391
Decision tree classifier	0.8190	0.4930	0.0442	0.9972	0.9740
Random forest	0.8583	0.5133	0.2225	0.9803	0.9855
Radial basis function SVM	0.5505	0.6296	0.5442	0.5608	0.5518
K-NN classifier	0.8139	0.4999	0.0208	0.9919	0.9725
Large margin K-NN	0.8269	0.4995	0.0108	0.9954	0.9901
1-NN-d	0.8400	0.5091	0.4042	0.9414	0.9272
Large margin NN-d	0.8588	0.5122	0.7150	0.9014	0.8876
1-NN-d with class-specific radius optimization	0.8386	0.5109	0.2725	0.9707	0.9518
Large margin NN-d with class-specific radius	0.8657	0.5181	0.3933	0.9717	0.9602
Two-stage model for radial basis function SVM	0.8548	0.5170	0.3933	0.9717	0.9471

Table D.4: Evaluation results on Statlog dataset

Classifier	Accuracy	Macro-F measure	Outlier recall	Inlier recall	Inlier accuracy
Naive Bayes classifier	0.7616	0.3573	0.0837	0.9531	0.8972
Linear discriminant analysis	0.7994	0.3731	0.0731	0.9643	0.9447
Decision tree classifier	0.7984	0.3765	0.0380	0.9938	0.9505
Random forest	0.8466	0.3879	0.1296	0.9835	0.9900
Radial basis function SVM	0.7549	0.3573	0.4127	0.8186	0.8233
K-NN classifier	0.8175	0.3824	0.0072	0.9910	0.9795
Large margin K-NN	0.8199	0.3834	0.0044	0.9920	0.9830
1-NN-d	0.8313	0.3759	0.6055	0.9022	0.8764
Large margin NN-d	0.7844	0.3572	0.7562	0.8181	0.7900
1-NN-d with class-specific radius optimization	0.8362	0.3859	0.3047	0.9646	0.9425
Large margin NN-d with class-specific radius	0.8649	0.3934	0.4837	0.9699	0.9411
Two-stage model for linear discriminant classifier	0.8564	0.3897	0.4837	0.9699	0.9309
Two-stage model for radial basis function SVM	0.8810	0.3974	0.4837	0.9698	0.9605

List of Publications

Journal Papers

1. Binh P. Nguyen, Wei-Liang Tay, Chee-Kong Chui, and Sim-Heng Ong, "A clustering-based system to automate transfer function design for medical image visualization," *The Visual Computer*, vol. 28, issue 2 (2012), pages 181-191, 2012.

2. Wei-Liang Tay, Chee-Kong Chui, Sim-Heng Ong, and Alvin Choong-Meng Ng, "Osteoporosis screening using areal bone mineral density from diagnostic CT," *Academic Radiology*, vol. 19, issue 10 (2012), pages 1273-1282, 2012.

3. Wei-Liang Tay, Chee-Kong Chui, Sim-Heng Ong, and Alvin Choong-Meng Ng, "Ensemble-based regression analysis of multimodal medical data for osteopenia diagnosis," *Expert Systems with Applications*, vol. 40, issue 2 (2013), pages 811-819, 2013.

4. Lile Cai, Wei-Liang Tay, Binh P. Nguyen, Chee-Kong Chui, and Sim-Heng Ong, "Automatic transfer function design for medical vi-

sualization using visibility distributions and projective color mapping," *Computerized Medical Imaging and Graphics*, vol. 37, issue 7 (2013), pages 450-458, 2013.

5. Rong Wen, Wei-Liang Tay, Binh P Nguyen, Chin-Boon Chng, and Chee-Kong Chui, "Hand gesture guided robot-assisted surgery based on a direct augmented reality interface," *Computer Methods and Programs in Biomedicine*, in press.

Conference Papers

1. Binh P. Nguyen, Wei-Liang Tay, Chee-Kong Chui, and Sim-Heng Ong, "Automatic transfer function design for volumetric data visualization using clustering on LH space," *Proceedings of Computer Graphics International*, pages 1-10, 2011.

2. Wei-Liang Tay, Chee-Kong Chui, Sim-Heng Ong, and Alvin Choong-Meng Ng, "Detection of osteopenia from routine CT Images," Presented at *7th Asian Conference on Computer-Aided Surgery*, 2011.

3. Wei-Liang Tay, Chee-Kong Chui, and Sim-Heng Ong, "Single camera-based remote pointing and recognition for monitor-based augmented reality surgical systems," *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Ap-*

plications in Industry, pages 35-38, 2012.