Breast Cancer Risk Assessment using Mammographic Image Texture Analysis

A thesis submitted for the degree of
Doctor of Philosophy

By

Xi-Zhao Li

School of Computer Science, Engineering and Mathematics
Faculty of Science and Engineering
Flinders University

November, 2014

Contents

List of Figures

List of Tables

List of Abbreviations

BSSA —— BreastScreen South Australia

CC —— craniocaudal (top) view of the breast

DDSM —— Digital Database of Screening Mammograms

ER —— Estrogen-Receptor

FNF —— false negative fraction

FPF —— false positive fraction

MLO —— mediolateral oblique (side) view of the breast

MR8 —— Maximum Response 8 filter bank

ROC —— the recever operating characteristic

ROIs —— regions of interest

SCC —— six-category classification

SD —— Standard deviation

TNF —— true negative fraction

TPF —— true positive fraction

Summary

Breast cancer is one of the most common cancers among women and early detection plays an important role in reducing the mortality and morbidity due to breast cancer. Importantly, early breast cancer detection is facilitated by accurate breast cancer risk assessment. This thesis aims to develop computer methods for analyzing tissue texture in screening mammograms in order to assess the risk of breast cancer.

According to the literature, the breast density is a strong indicator of breast cancer risk and is independent of non-mammographic risk factors (age, race, family history, etc.). In addition, texture from screening mammograms is also considered to play an important role in predicting breast cancer risk. However, the contribution of texture alone to breast cancer risk is unclear and the role of texture for assessing breast cancer risk over time is also unknown. The focus of this thesis is on studying the role of texture, independent of density, in breast cancer risk assessment.

In this thesis, the emphasis is on characterizing texture through the use of textons. Textons can be described as ubiquitous local texture patterns. The distribution of conventional textons (referred to as first-order textons in this thesis) has been shown to characterize texture in visual images and has been successful in tasks such as separating regions corresponding to grass from regions representing trees or animals. An important contribution of this thesis is the introduction of higher-order textons. The notion of higher-order textons is to extend the power of the first-order textons. Higher-order textons allow quantitative analysis of commonly occurring patterns of patterns, offering a mechanism for understanding more complex texture structure in images. In this thesis, textons and higher-order textons are used to distinguish mammograms from women having a high risk of breast cancer from women having a low risk of breast cancer.

A number of experiments were conducted to determine the best implementation of textons and higher-order textons for breast cancer risk assessment. Results indicate that texture analysis based on higher-order textons predicts risk at least as well as any method currently available for estimating breast cancer risk from mammograms. Risk of breast cancer can be measured using texture at least four years prior to the cancer becoming apparent mammographically.

In addition, a number of discoveries were made in the course of the study. Tex-

ture features from CC view mammograms (top view) perform better than texture features from MLO view mammograms (side view). Better risk assessment is obtained by measuring texture over the full breast than any particular local region of the breast. Texture features calculated from 3×3 local neighborhoods perform as good or better than texture features based on larger patches. Texture information relevant to breast cancer risk is more pronounced in the breast in which cancer eventually occurs than in the breast without known cancer of the same woman. These discoveries have potential impact on the fields of image analysis and computer-aided mammography and so form natural seeds for future work.

Publications arising from the Study

Referred Conference Paper

[1] Xi-Zhao Li, Simon Williams, and Murk J. Bottema. Intensity independent texture analysis in screening mammograms. In *11th International Workshop on Breast Imaging, IWDW2012, Philadelphia, PA, USA*, pages 474-481, July 2012.

[2] Xi-Zhao Li, Simon Williams, Gobert Lee, and Min Deng. Computer-aided mammography classification of malignant mass regions and normal regions based on novel texton features. In *12th International Conference on Control, Automation, Robotics and Vision, Guangzhou, China, ICARCV2012*, pages 1431-1436, December 2012.

[3] Xi-Zhao Li, Simon Williams, Peter Downey and Murk J. Bottema. Temporal breast cancer risk assessment based on higher-order textons. In *12th International Workshop on Breast Imaging, IWDW2014, Gifu, Japan*, pages 565-572, June-July 2014.

Referred Journal Paper

[4] Xi-Zhao Li, Simon Williams and Murk J. Bottema. Background intensity independent texture features for assessing breast cancer risk in screening mammograms. *Pattern Recognition Letters,* 34(9):1053-1062, Feb 2013.

[5] Xi-Zhao Li, Simon Williams and Murk J. Bottema. Texture and region dependent breast cancer risk assessment from screening mammograms. *Pattern Recognition Letters,* 36(15):117-124, Jan 2014.

[6] Xi-Zhao Li, Simon Williams and Murk J. Bottema. Constructing and applying higher order textons: Estimating breast cancer risk. *Pattern Recognition,* 47(3):1375-1382, Mar 2014.

Declaration

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

Xi-Zhao Li, Candidate

Murk J. Bottema, Principal Supervisor
Simon Williams, Co-supervisor

Acknowledgements

I would like to thank my supervisor, A/Prof. Murk J. Bottema for his support and encouragement throughout my PhD study. Importantly, he helped me improve my research skills in medical image analysis, academic writing and mathematical theories for my project. Specifically, I really thank him for giving me the opportunity to study in Flinders University.

I would like to also thank my co-supervisor, Dr. Simon Williams for his guidance and supervision throughout my PhD study. Specifically, he provided me great help in using Lyx for academic writing and opportunity for practicing teaching.

I thank Dr. Gobert Lee for her support in academic fundamental knowledge, suggestions for my research and guidance in using university resources for doing research. Special thanks to Dr. Mariusz Bajger for his technical help; Dr. Ray Booth and A/Prof. Alan Branford for giving me the opportunity of doing some mathematics and statistic related part-time jobs in the school. A special thank to Dr. Shu-Chuan Chu for her support in my study as a friend.

Special thanks to Dr. Adham Atyabi for his help in using the super computer for running my time consuming programs. Particularly, I thank Dr. Adham Atyabi for all his precious suggestions of doing research.

I would also like to thank BreastScreen South Australia (BSSA) for providing mammogram images for my final stage research.

Gratefully, I thank China Scholarship Council (CSC) and Flinders University for all their financial support for my whole study.

Finally, I thank all my family for their great support and love all the way through.

Chapter 1

Introduction

This introductory chapter is intended to provide the context for this thesis. The incidence of breast cancer is reviewed in Section 1.1 and the role and modalities of breast cancer screening are presented in Section 1.2. Widely accepted clinical breast cancer risk factors are discussed in Section 1.3. These discussions lead to the motivation and objectives of the thesis in Section 1.4. This is followed by an overview of the data sets used in the thesis in Section 1.5 and an overview of the remaining structure of the thesis in Section 1.6.

1.1 Breast Cancer Incidence

Breast cancer is one of the most common cancers and the cause of cancer death in the world. Breast cancer is the second most common cancer diagnosed in women worldwide. In 2008, approximate 1.38 million women across the world were diagnosed with breast cancer, accounting for nearly a quarter (23%) of all cancers diagnosed in women. Breast cancer incidence in women in developed countries is generally higher than women in developing countries. The incidence of breast cancer in most developed countries has increased worldwide in the last decades. Breast cancer is the most common cause of death from cancer in women worldwide and the mortality due to breast cancer ranked fifth in both sexes combined in 2008 (Boyle and Levin [2008], Ferlay et al. [2010]).

On average, 1 in 8 Australian females will develop breast cancer at some time during their life and 1 in 37 Australian females will die from breast cancer before the age of 85 (AIHW & AACR [2012a]). In 2009, $13,668$ new breast cancer cases were reported in Australia, which comprise 27.4% of all new reported cancers in females (AIHW & AACR [2012a]). The incidence of breast cancer in Australia increased from the year of 1982 to the year of 2009 (AIHW & AACR [2012a]). In 2010, $2,840$

Australian women died from breast cancer, which means that 8 Australian females died from breast cancer every day in 2010 on average (AIHW & AACR [2012a]). It is estimated that, by 2020, there will be $17,210$ new cases of breast cancer diagnosed in women (AIHW & AACR [2012b]). Breast cancer is the second leading cause of cancer-related death in Australian women, accounting for 15.3% of all cancer deaths in women in 2010 (AIHW & AACR [2012a]).

In the USA, breast cancer is the most common cancer among women excluding skin cancer, accounting for almost 1 in 3 cancers diagnosed in US women (Alteri et al. [2011]). About 1 in 8 American women will develop breast cancer during their lifetime (Alteri et al. [2011], Desantis et al. [2014]). In 2013, an estimated $232,340$ new cases of invasive breast cancer will be diagnosed among US women (Desantis et al. [2014]). It is expected that around $39,620$ women will die from breast cancer in 2013, which is the second most common cause of death from cancer (Desantis et al. [2014]).

In the UK, breast cancer is by far the most common cancer among women (2010), accounting for 31% of all new cancer cases in women (Cancer Research UK [2012]). The life-time risk of developing breast cancer in the UK is 1 in 8 in females (Cancer Research UK [2013a]). In 2010, it is reported that $49,564$ new female breast cancer cases occurred in the UK, indicating that 126 out of $100,000$ UK women were detected with cancer (Cancer Research UK [2013a]). In the last 10 years, female breast cancer incidence rates have increased by 6% (Cancer Research UK [2013a]). In 2010, about $11,600$ women in the UK died from breast cancer, which is approximately 32 per day (Cancer Research UK [2013a]). Breast cancer is currently the second most common cause of death of women in the UK (Cancer Research UK [2013a,b]).

1.2 Breast Cancer Screening

The objective of breast cancer screening is not to diagnose cancer but to ascertain whether there is enough evidence of breast cancer to call the subject back for more tests (Bottema et al. [2000]). Early detection through breast screening mammography is widely viewed as providing the best opportunity for reducing morbidity and mortality due to breast cancer. As a result, many countries provide population-based breast cancer screening to facilitate early detection (Zorbas [2003]). For example, in Australia, breast cancer screening is available to women aged 40 or over without charge to the client through BreastScreen Australia (AIHW & AACR [2012b]).

The practical value of this thesis is that for women participating in screening programs, the availability of screening mammograms provides additional information regarding breast cancer risk. In turn, estimates of breast cancer risk allow screening

strategies to adjust to the individual. Accordingly, screening programs are crucial to this study and merit some description.

1.2.1 Screen-film Mammography versus Digital Mammography

Currently, the most widely used method for early breast cancer detection is X-ray mammography. Mammography is a low-dose X-ray procedure that allows visualization of the internal structure of the breast. In modern society, high-quality mammography images with relatively low X-ray dose are used for early breast cancer detection.

Conventionally, X-ray mammography is screen-film mammography that is performed with the breast directly in contact with the screen-film cassette. Routinely, each breast will be imaged separately with two different views. The cranio-caudal (CC) view is taken from above a horizontally-compressed breast. The mediolateral-oblique (MLO) view is taken from the side and at an angle of a diagonally-compressed breast. Other views may be taken for a diagnostic mammography if necessary. Mammograms are read by one or more radiologists who decide if there is enough evidence of cancer to call the woman back for further tests. The protocol regarding the number of views and number of expert readers varies somewhat between countries. Screen-film mammography is widely available and is covered by nearly all the health insurance providers. It has been done for over 30 years and achieved high degree of success (Whitman and Haygood [2012]).

Full-field digital mammography is a newer technique that allows the X-ray image to be viewed on a computer screen without digitizing the film image. In computer-aided analysis, the quality of the digital image that is executable by the computer is improved. The dose of radiation used in full-field digital mammography is less than that in screen-film mammography, reducing the lifetime X-ray exposure (Obenauer et al. [2003], Gennaro and Maggio [2006], Hauge et al. [2011]). It is more accurate than screen-film mammography at finding cancer in young women and those with dense breast tissue (Bluekens et al. [2012], Skaane and Skiennald [2004]). Currently, this technique is not yet available everywhere because the cost of full-field digital mammography systems is around 1.5 to 4 times more than screen-film mammography systems (Pisano et al. [2005]). Generally, screen-film mammography and full-field digital mammography are considered equally effective (Lewin et al. [2002], Vinnicombe et al. [2009]).

1.2.2 Other Modalities

There are other modalities of breast cancer screening. Molecular breast imaging is a nuclear medicine breast cancer screening technique that is currently under study. It is promising for imaging people with dense tissue and may even have comparable accuracies as magnetic resonance imaging (Connor et al. [2009]). However, it involves higher doses of radiation, bringing greater risk of radiation damage than the two general breast cancer screening techniques above (Section 1.2.1). Ultrasonography is a significant adjunct to mammography and further clinical examination of suspicious abnormalities (Teh and Wilson [1998]). It aims to aid general mammography of women with dense tissue (Berg et al. [2008]). This modality increases the breast cancer detection rate but also increases the false positive rate (Berg et al. [2008, 2012]). Magnetic resonance imaging has been used to detect cancers that are not visible in mammograms. It is excellent for screening women with high genetic risk or dense breasts. However, it has been claimed to be less specific for women at average risk than mammography and the procedure of magnetic resonance imaging is more expensive (Hrung et al. [1999], Medical Advisory Secretariat [2010]). Computed tomography produces images of specific areas of the scanned object and allows the radiologist to see 3-dimensional full structure of the breast without occlusion (Herman [2009]). The cost of computed tomography examination is expensive and the dose of radiation delivered by computed tomography is high. Computed tomography is suggested only when no other test or procedure can supply the information needed (Fred [2004]). In the last several years, tomosynthesis has been proposed as an alternative to full-field digital mammography but is not yet widely used clinically (Helvie [2010], Teertstra et al. [2010]). None of these modalities are generally regarded as viable alternatives to screening mammography at this time.

1.3 Breast Cancer Risk Assessment

In this section, clinically identified breast cancer risk factors are reviewed. The significance of these factors in clinical practice for risk assessment is highlighted for making decisions regarding personalized screening programs.

1.3.1 Identified Risk Factors for Breast Cancer Risk Assessment

Even though the causes of breast cancer are not fully known, there are a number of factors associated with an increased chance of developing breast cancer in the future. Accordingly, a good understanding of risk factors for breast cancer is vital. Known

risk factors include age, body weight, alcohol consumption, diet, geographical location, family history, previous benign breast disease, cancer in other breast, exposure to ionizing radiation, taking exogenous hormones, and many more (McPherson et al. [2000]).

Some risk factors are particularly strong and consistent (relative risk (McPherson et al. [2000]) ≥ 2). Breast cancer incidence increases with age, doubling around every ten years until menopause when the rate of increase drops dramatically (McPherson et al. [2000]). Women who start menstruating early or have menopause late in life have increased breast cancer risk (McPherson et al. [2000]). Breast cancer risk increases for females who are nulliparous or have their first birth at a late age (McPherson et al. [2000]). Women with a genetic predisposition to breast cancer are associated with extremely high risk (McPherson et al. [2000]). The number of genes involved in predisposition to breast cancer is not yet clear but BRCA1 and BRCA2 are two genes closely linked to breast cancer (Miki et al. [1994], Wooster et al. [1995]). Ionizing radiation is found to increase breast cancer risk later in life (McPherson et al. [2000]). This is supported by the finding that teenage girls exposed to radiation during the Second World War were observed with double breast cancer risk (Boyce [2004]). Women with obvious atypical epithelial hyperplasia are associated with four to five times higher risk of developing breast cancer in the future than those who do not have any proliferative changes in their breasts (McPherson et al. [2000], National Breast and Ovarian Cancer Centre (NBOCC) [2009]). Women detected with invasive breast cancer in one breast are at 2 to 6 times the risk of developing breast cancer in the contralateral breast in the future (National Breast and Ovarian Cancer Centre (NBOCC) [2009]). Generally, the breast cancer risk for women from developed countries is higher than that of women from developing countries (National Breast and Ovarian Cancer Centre (NBOCC) [2009]).

Some other risk factors are not particularly strong or inconsistent. There is a correlation between breast cancer risk and dietary fat intake. Similarly, obesity, alcohol intake and smoking are associated with increased risk of breast cancer. There is a slight increase of relative breast cancer risk for females taking oral contraceptives and for 10 years after stopping these agents. Hormone replacement therapy is found to increase breast density and reduce the sensitivity and specificity of breast screening. As a result, it brings an increase of breast cancer risk (McPherson et al. [2000]).

Based on the above known risk factors, there are several well-known models for predicting risk clinically (Evans and Howell [2007]). The Gail model is a well-known risk prediction model which calculates a woman's risk of developing breast cancer within the next five years and within her lifetime (up to age 90) by translating a female's risk factors into an overall risk score through multiplying her relative

risks from several categories (such as age at menarche, number of breast biopsies, family history, ethnicity and age at first live birth). It focuses primarily on non-genetic risk factors with limited information on family history. It was designed by researchers at the National Cancer Institute and the National Surgical Adjuvant Breast and Bowel Project (Gail et al. [1989], Costantino et al. [1999]). The Clause model is another commonly used model for risk prediction, based on the prevalence of high-penetrance genes associated with breast cancer. This model incorporates more extensive family history information but excludes risk factors not related to family history (Claus et al. [1994]). The BRCAPRO model is a Bayesian model developed by Parmigiani and colleagues (Parmigiani et al. [1998]). It incorporates identified BRCA1 and BRCA2 mutation frequencies, cancer penetrance in mutation carriers, cancer status (affected, unaffected or unknown) and age of the female's first-degree and second-degree relatives who have breast cancer history. The Cuzick-Tyrer model integrates family history, surrogate measures of endogenous oestrogen exposure and benign breast disease in a comprehensive way (Tyrer et al. [2004]). The Cuzick-Tyrer model allows for the presence of multiple genes of different penetrance beyond what is allowed by the Clause and BRCAPRO models. In the literature, the Cuzick-Tyrer model was certified to provide the most consistently accurate risk estimation for women with high risk (Amir et al. [2003]).

In addition to the above risk factors, mammographic image appearance plays an important role in helping radiologists predict risk as well. Studies showed that breast density is one of the strongest predictors for the risk of developing breast cancer and is independent of other risk factors (Wolfe [1976a], Wolfe et al. [1987], Byng et al. [1994], Boyd et al. [1995], Byng et al. [1996, 1997], McCormack and Silva [2006]).

Wolfe was the first to study the relationship between mammographic appearance and breast cancer risk (Wolfe [1976a,b]). He proposed four breast pattern classes: $N1$, $P1$, $P2$, and DY, and demonstrated a substantial increase in breast cancer risk progressing from $N1$ patterns to DY. $N1$ denotes a breast comprising mostly fat; $P1$ denotes a breast with a prominent duct pattern but limited in extent; $P2$ denotes an extended and prominent duct pattern; and DY denotes an extremely dense duct pattern. The observations by Wolfe were reproduced by some studies (Wellings et al. [1975], Brisson et al. [1981], Boyd et al. [1984], Saftlas et al. [1989]). However, other studies did not reproduce odds ratios as great as those reported by Wolfe, and some even failed to find evidence of a relationship between Wolfe's breast patterns and breast cancer risk (Egan and Mosteller [1977], Whitehead et al. [1985], Mendell et al. [1977]). Later on, another parenchymal pattern classification was proposed by Tabár in 1997. Five Tabár patterns are based on anatomic-mammographic correlation with 3-dimensional, subgross (thick-slice) techniques and on the relative proportion of four "building blocks" (nodular densities, linear densities, homogeneous

fibrous tissue and radiolucent fat tissue) (Gram et al. [1997]). Similar to Wolfe patterns $N1$ and $P1$, the first three Tabár patterns (I - III) are grouped as the low risk group, and similar to Wolfe patterns $P2$ and DY, the last two Tabár patterns (IV and V) are grouped as the higher risk group.

The previous two classifications focus more on the structure of patterns while the following two classifications shift the focus to the amount and distribution of dense tissues. Boyd density classification was proposed in 1980s and was based on mammographic density percentage as assigned by radiologists. There are six categories of unequal intervals (none, $< 10\%$, $10-25\%$, $25-50\%$, $50-75\%$ and $\geq 75\%$) (Boyd et al. [1995]). Hereafter, this six category classification will be refereed to as SCC categories. In this pattern classification, breast cancer risk increases with the increase of density percentage. The American College of Radiology introduced BI-RADS classes (American College of Radiology [2003]). They are a modified version of Wolfe classes. BI-RADS I breasts are almost entirely fat, in which the fibrous and glandular tissue occupies less than 25% of the breast. BI-RADS II breasts have scattered fibroglandular densities in which fibrous and glandular tissue makes up from $25-50\%$ of the breast. Breasts in BI-RADS III are heterogeneously dense with $51-75\%$ areas of fibrous and glandular tissue. Breasts in BI-RADS IV are extremely dense, consisting of more than 75% fibrous and glandular tissue. From BI-RADS I to BI-RADS IV, the breast cancer risk increases.

In addition, the appearance of microcalcification/calcification clusters in mammograms is associated with a reasonably high risk of developing breast cancer (Thomas et al. [1993], Picca and Paredes [2003], Giger et al. [2013]). Calcifications are usually characterized by radiologists according to morphology, distribution, size, number, variability and stability from previous mammograms.

In practice, non-image risk factors are usually combined with factors related to mammographic appearance (such as density distribution) to help radiologists assess a woman's breast cancer risk more accurately. Schousboe et al. studied the cost-effectiveness of mammography with risk factors including age, breast density, history of breast disease, and family history of breast cancer. They suggested that personalized breast cancer risk should be estimated based on these risk factors for recommending breast cancer screening strategies to individuals (Schousboe et al. [2011]). A refinement of the Gail model was developed by Chen and colleagues to estimate breast cancer risk by adding a continuous measure of breast density in addition to the non-image risk factors used by the original Gail model (Chen et al. [2006]). The refined Gail model improves risk discrimination but it is not routinely available in clinical practice because it needs further validation with independent data and the continuous measure of breast density is not very convenient. In the study by Barlow et al., a model was developed (using Breast Cancer Surveillance

Consortium data and focus on one-year risk) that includes traditional risk factors and other recently identified factors - race, ethnicity, breast density, high body mass index, use of hormone therapy, type of menopause and previous mammographic result, and found that this model may identify high-risk women better than the original Gail model. This indicates that risk prediction models may be improved by adding other risk factors (such as breast density) (Barlow et al. [2006]). However, this model may overestimate a female's long term breast cancer risk by including incident cancers detected by the first mammogram. More recently, a new risk prediction model incorporating breast density with other traditionally used risk factors (age, race or ethnicity, family history and biopsy history) was developed to estimate a woman's five year risk of developing invasive breast cancer. Even though this model is convenient enough to be incorporated into routine breast cancer screening, its accuracy needs to be further evaluated in an independent population before being ready for clinical use (Tice et al. [2008]). Mealiffe et al. proposed another new breast cancer risk prediction model for combining genetic risk factors and clinically identified risk factors that obtained improved classification results in white non-hispanic postmenopausal women (Mealiffe et al. [2010]). This model is limited by the population based cohorts and clinical characteristics of women therein. In summary, currently, there are no breast cancer risk prediction models using both non-image and mammographic appearance risk factors available for clinical practice. However, the performance of risk estimation is able to be improved by combining diversified risk factors.

1.3.2 Benefits of Breast Cancer Risk Assessment

An important benefit of breast cancer risk assessment is to enable personalized breast cancer screening.

The idea of personalized breast cancer screening is to refine screening recommendations to women based on individual risk. In personalized screening programs, routine mammography for all women in their early 40s is not recommended but left to be an individual decision, taking into account patient context (National Cancer Institute [2011], Schousboe et al. [2011], Mandelblatt et al. [2011]). For young women, the small mortality reduction achieved from screening mammography does not justify the treatments resulting from false-positive findings (American Association for Cancer Research [2012]). With breast cancer risk assessment made according to individual risk factors, women and their doctors can make individual decisions about when to start to have mammograms and how often. The risk of breast cancer is not equal for all women and so adjusting breast cancer screening strategies according to the level of risk of an individual increases the efficiency and reduces the

cost of screening programs (Pashayan et al. [2011]). Results on the benefit of truly personalized screening for breast cancer have not been published, but personalzed screening is likely to reduce mortality in high risk women and reduce the chance of unnecessary intervention for low risk women.

1.4 Motivation and Objectives of the Thesis

The reduction of mortality and morbidity due to breast cancer is vital and it is facilitated by early detection (Surveillance Epidemiology and End Results (SEER) [2012], American Cancer Society (ACS) [2012a,b]). According to Section 1.2 and Section 1.3, optimal individualized strategies for early breast cancer detection, including breast cancer screening, depend, in part, on accurate breast cancer risk assessment. In the past, many well-known risk factors unrelated to breast images were used to construct models for breast cancer risk prediction and some of them have been applied in clinically practice (Section 1.3). Breast density derived from screening mammograms offers additional information regarding breast cancer risk and is a well-established breast cancer risk factor. Breast density can be estimated subjectively by radiologist or measured objectively using image analysis methods with the aid of computers. As presented in Section 1.3.1, breast cancer risk prediction models that combine non-image risk factors and breast density achieved some improvement in risk assessment compared with models using non-image risk factors alone.

In addition to breast density, patterns of tissue texture are also thought to be correlated to risk. However, the exact nature of texture patterns associated with breast cancer risk is not known and is difficult to quantify. Although some work on estimating risk from screening mammograms has appeared (Section 1.3 and Section 2.9), the full potential of risk assessment based on texture has not been fully explored. One problem in studying texture in screening mammograms is that the relationship between total attenuation of the X-ray beam and image intensity is non-linear. Hence the contribution to intensity of small components (such as ducts) results in an intensity signal that varies according to the local intensity of the background. Thus results reporting a positive contribution to risk assessment based on texture may be due to the indirect measurement of density.

In light of the above discussion, the objective of this thesis is to develop texture analysis methods suitable for assessing breast cancer risk with screening mammogram images independent of density. In addition, the contribution of texture independent of density to breast cancer risk assessment will be studied with newly developed texture analysis methods. Importantly, temporal breast cancer risk assessment is explored by using texture analysis methods to quantify breast cancer risk in a sequence of temporal screening mammograms.

1.5 Overview of the Data Sets

In order to avoid repeating common descriptions of the data sets used in the experimental chapters (Chapters 3 - 8), an overview of the data sets is presented in this introductory chapter. Two data sets are used for conducting the experiments reported in this thesis; the publicly available Digital Database of Screening Mammography (DDSM) (Heath et al. [1998, 2001]) and an in-house database of images sourced from the archives of BreastScreen SA (BSSA) , the organization that oversees breast screening in South Australia. Both sets of images play vital roles in the thesis.

The DDSM images were available from the onset of the study and each image in this set comes with a BI-RADS score. BI-RADS scores are used in Chapter 3 and Chapters 5 to 8 to insure a wide representation of mammographic appearance. In Chapter 4, BI-RADS classes are used as a surrogate for risk. However, the DDSM collection does not include temporal sequences of images. Hence, this set cannot be used to compare potential indicators of risk in image taken before the mammographic appearance of the cancer itself. This limits the value of this data set in estimating risk of breast cancer.

The BSSA database was not available at the onset of the study. The process of digitizing four images from 200 women over three visits each (2400 images in total) was conducted over the course of the study. The temporal structure of this data set allows a better measure of risk as described in Chapter 8. However, BI-RADS scores are not available for images in this database thus limiting the extent to which wide mammographic appearance could be used to train the algorithm for characterizing risk. BI-RADS scores were assigned to these images informally by the author (referred to in this thesis as in-house BI-RADS classes) to mitigate this deficiency.

Images from the DDSM database are acquired by Lumisys and Howtek machines at spatial resolutions ranging from 42μm to 50μm per pixel and depth ranging from 12 to 16 bit. Images were corrected for differences in acquisition parameters and machine characteristics before further processing steps were applied.

Film images from the BSSA database were digitized at 57.0 μm spatial resolution and 12 bit depth. Images were collected from three consecutive screening visits, nominally spaced two years apart and with the most recent visit being in 2005 or 2006. Here, a "case" will refer to the collection of images from one woman over all three visits. Cases were designated as cancer if anomalies found at screening during the $2005/6$ round were confirmed as cancer by histopathology but no evidence of cancer had been found in previous rounds. Cases were designated as normal if no cancer had been found in any round including at least one screening visit post $2005/6$.

For all the images in both data sets, the breast boundary was drawn manually by the author using *Image*J software. An image template of the breast region was generated based on the boundary.

1.6 Overview of the Thesis

This thesis contains 9 chapters including the current introduction chapter (Chapter 1) plus two appendixes. The remaining chapters describe the evolution of the texture methods developed for estimating breast cancer risk culminating in a longitudinal study described in Chapter 8.

Chapter 2 provides a literature review of related work in this area and reviews key methods used in this thesis for developing the final methods of texture analysis.

Methods for preprocessing images prior to texture feature extraction are developed in Chapter 3. This algorithm also separates texture information from density information in the image.

Breast cancer risk assessment with the surrogate true risk criteria of BI-RADS breast patterns is shown in Chapter 4. In this chapter, the effect of applying the proposed image preprocessing algorithm, using different view breast images and generating textons with different candidate methods and clustering methods are compared.

Starting in Chapter 5, the surrogate for true risk of breast cancer is changed. The breast contralateral to the breast in which cancer was detected in the current screening visit is defined as high risk while the breasts from women not found to have cancer in either breast are defined as low risk. In this chapter, different regions of interest (ROIs) within the breasts are compared in terms of texture information contributing to breast cancer risk assessment. The whole breast region is found to provide the most important texture information for risk classification and hence the full breast is used to conduct the experiment in subsequent chapters.

In Chapter 6, higher-order textons are introduced as a method for extending the power of conventional textons in analyzing texture. Higher-order textons are then used to calculate texture features for breast cancer risk assessment. The use of higher-order textons is shown to improve risk assessment and is adopted in subsequent chapters.

The contribution of texture and density to breast cancer risk assessment is analyzed separately and in combination in Chapter 7. The role of texture is found to be at least as important as that of density in breast cancer risk assessment.

Temporal breast cancer risk assessment is described in Chapter 8. The texture analysis methods developed in previous chapters, including the method of higher-order textons are applied in this chapter to determine the extent to which texture is able to predict the onset of cancer, two and four years prior to actual mammographic

detection. This temporal study indicates positive prediction ability of the developed texture analysis method in predicting future breast cancer.

Final remarks are made in Chapter 9. Feature indexing of the final set of features and some detailed supplementary experiment results are shown in Appendix A and B, respectively.

Chapter 2

Technical Background and Literature Review

This thesis comprises extensions and applications of computational texture analysis to estimate risk of breast cancer from screening mammograms. This chapter provides background on texture analysis generally (Section 2.1) and on the application of texture analysis to screening mammograms (Section 2.9). Textons are the most important component of texture analysis in this thesis and are reviewed separately in Section 2.2. Sections 2.3 to 2.8 describe existing methods that have been incorporated into the risk assessment system developed in this thesis. Section 2.3 introduces three general methods for extracting local feature vectors, the collection of which is used as the feature space for texton generation. Section 2.4 describes two clustering methods which can be used to generate textons from the feature space. Classifiers used in the thesis are introduced in Section 2.5. Validation methods for measuring the reliability of a classification system are described in Section 2.6. ROC analysis is reviewed in Section 2.7 and methods for selecting features from a large feature set are described in Section 2.8.

2.1 Texture Analysis

From the early days of image analysis, image texture has been recognized as an important attribute for understanding image content. In 1973, Haralick (Haralick et al. [1973]) introduced gray-tone spatial-dependence matrices (often named co-occurrence matrices (Davis et al. [1979])), which are tantamount to determining the joint probability that a pixel has value i and that a second pixel at distance d and orientation θ has value j. For each d and θ, a number of features are then derived from the joint probability distribution. Haralick suggested 14 such features including the mean, contrast, entropy, and difference entropy. In a similar vein, Galloway

(Galloway [1975]) introduced gray level run lengths. These methods have been used widely in many areas of image analysis (Dasarathy and Holder [1991], Tang [1998], Loh et al. [1988], Hu et al. [2012], Rath et al. [2012], Losson et al. [2013]), but suffer from two important drawbacks. First, if oriented texture is to be analyzed, then many different directions θ must be chosen in addition to several possible distances d. With 14 features for each of these combinations, the total number of features quickly explodes (Sahiner et al. [1998], Lee and Bottema [2006]). Second, each gray-tone spatial-dependence matrix has size $q \times q$ where q is the number of gray-tones used. So, either these matrices are very large, or the gray scale must be quantized. In applications to mammography, these problems are exacerbated by the fact that, typically, $q = 1024$ or $q = 4096$ and since small local intensity variation is often important, quantization may introduce errors. In addition, X-ray images are projection images, meaning that information of diagnostic value but low contrast is often superimposed on a highly fluctuating background representing various tissue types. As a result, a single object of constant X-ray attenuation (say a duct) may be represented by a wide range of intensities. Thus consistent pixel intensity values are not necessarily expected even for objects of constant X-ray attenuation.

Sum and difference histograms were introduced as an alternative to the above co-occurrence matrices (Unser [1986]). Instead of using co-occurrence matrices, texture features were computed directly from sum and difference histograms because they defined the principal axes of second-order probability functions of a stationary random process and therefore have equal effectiveness as co-occurrence matrices (Unser [1986]).

Fractal analysis is a representative model based texture analysis method for image analysis (Keller et al. [1989], Kaplan [1999], Quevedo et al. [2002], Myint [2003], Backes et al. [2012]). Fractal dimension as a scale insensitive ruggedness measure alone is not sufficient to classify natural textures (Medioni and Yasumoto [1984]). Hence other measures such as features based on the concept of lacunarity as the second-order statistic features of fractal surfaces were used together with fractal dimension features for texture analysis (Keller et al. [1989]). The Markov random field model is another representative texture model in image analysis (Cross and Jain [1983], Chen and Huang [1993], Rellier et al. [2004], Huawu and Clausi [2004]). Different texture features can be derived from Markov random field models with various settings of the model parameters, Markov random field models allow consideration of neighbors in all directions and images can be generated from features computed from Markov random field model (Cross and Jain [1983]). In addition, local binary patterns as a particular case of a texture spectrum model was proposed by Wang et al. in 1990 based on texture units and an image is characterized by its texture spectrum (Wang and He [1990], He and Wang [1990]). The term local

binary patterns was first recorded in 1994 (Ojala et al. [1994]). Afterwards, local binary patterns, which describe the relationship of pixels to their local neighborhoods, were used to compute texture measures for image classification (Schaefer and Doshi [2012], Guo et al. [2012a], Liu et al. [2012]).

Fourier transform as a method of texture analysis was studied early by Bajcsy (Bajcsy [1973]). She applied fan-shaped and ring-shaped filters to the Fourier power spectrum to calculate texture features. She found that the Fourier transforms work well on linear periodic textures as well as on linear regular but not periodic textures (Bajcsy [1973]). Afterwards, Matsuyama et al. used Fourier transforms to extract texture features by combining the frequency domain and the picture space (Matsuyama et al. [1983]). Traditionally, in this method, texture features are extracted from Fourier spectrum of the texture image. In the study by Hsu et al., multi-resolution Fourier transforms were used to analyze natural textures consisting various levels of structures (Hsu et al. [1993]). Multi-resolution texture features were generated by the wavelet and windowed Fourier transforms, where scale and frequency were varied independently. The Fourier transform plays a dominant role in signal analysis but a much smaller role in image analysis. This is because most signals of interest are generated by systems with natural periodic signatures. This includes the human voice, music, starlight, etc. In these cases, the Fourier transform serves to focus attention on the key information content, namely the frequencies. In images, most of the information is carried by edges (often curved ones) and regions of high or low contrast of various shapes but periodic features are not common. In image texture analysis, the Fourier transform may be used to determine periodic textures, but is seldom optimal for characterizing texture in general. However, in the study by Li et al. (Li et al. [2008]), discrete Fourier transforms were used to extract features for classifying BRCA1/BRCA2 gene mutation carriers (high risk) and non carriers (low risk). This is a very special case, because the very strong risk indicator (gene mutation) was used and a fixed small square ROI was chosen from the central region behind the nipple for risk assessment. Periodic textures are not expected in screening mammograms and so Fourier methods have not been investigated for texture analysis in this thesis.

Texture features are naturally derived from filtered images. More, and throughout the thesis, the world filter is used in the signal processing sense. A filter is the same as a convolution kernel. In 1979, Laws introduced a collection of 5×5 filters and texture energy features were derived from the resulting filtered images (Laws [1979]). Many different filters have been used since then, often tailored to the application in mind. Filters are routinely designed to search for specific shapes and at the same time manage noise, reduce aliasing and subtract background fluctuations. For example, Gabor wavelets (Gabor and Ing [1946], Lee [1996], Liu et al.

[2012]) are filters designed to detect oriented textures, manage aliasing, etc., and are a better option than run length or spatial-dependence matrices for quantifying oriented texture patterns in most applications. For similar reasons, the wide range of wavelet filters (Chang and Kuo [1993], Laine and Fan [1993], Livens et al. [1997], Van et al. [1999], Khouzani and Soltanian [2005], Nascimento et al. [2013]) are natural choices for multi-resolution texture analysis. Further more, textons defined as the texture primitives were generated from filtered image feature space for extracting texture features (Section 2.2).

2.2 Textons

As early as 1981 the notion of texture primitives was introduced by Julesz to refer to local features that allow human perception to distinguish between iso-second-order textures (Julesz [1981]). Malik et al. (Malik et al. [1999, 2001]) calculated textures from the output of multi-dimensional filters (filter bank). They performed vector quantization, or clustering, in the high-dimensional feature space of filter responses to find texture prototypes. They called these prototypes textons, which is an operational definition for textons in gray-level images. By using K-means clustering on a large number of images, they constructed a universal texton dictionary. By mapping each pixel to the texton nearest to its vector of filter responses, the image can be analyzed into texton channels. These foundations are built on models for human perception of texture in images. Combinations of grooves, spots, ridges, and hollows are thought to be perceived as a finite number of textures up to equivalence under changes of scale, orientation and lighting. This motivated the idea of representing pixels by vectors of texture primitives and then clustering these vectors to determine a finite number of representative patterns - textons. Julesz (Julesz [1981]) proposed that the first-order statistics (density) of textons can be used to discriminate textures, rather than second- or third-order statistics. In 1983, Julesz et al. (Julesz and Bergen [1983]) claimed that only differences in textons or their density can be pre-attentively detected and pointed out that the focus should be on texton differences.

The word texton was later re-invented to refer to co-occurrences of filter outputs (Cula and Dana [2004], Leung and Malik [2001], Schmid [2001], Varma and Zisserman [2005]). A common realization of this idea is to create a feature vector comprising the outputs at a pixel of a filter bank and to search for clusters in the resulting feature space. The clusters are called textons. Zhu et al. (Zhu et al. [2005]) discussed the definition of textons and argued that the set of textons must be learnt from, or best tuned to, an image ensemble. The texton boost model is a texton based model proposed by Shotton et al. to learn features from texton maps by incorpo-

rating texture, layout and context information efficiently (Shotton et al. [2009]). In their study, the term texture-layout filters were described and used to extract features from texton maps. At the same time, semantic texton forests were proposed by Shotton et al. for image categorization and segmentation as well (Shotton et al. [2008]). Semantic texton forest is another texton based model built on randomized decision forests and skips the time consuming steps of filtering, k-means clustering and k-nearest neighbor assignment. However, it suffers from the large dimensionality of the bag of semantic textons.

Varma and Zisserman (Varma and Zisserman [2003]) proposed constructing feature vectors from pixel values in $N \times N$ neighborhoods instead of standard filter responses. They compared the performance of using filter banks to using $N \times N$ neighborhoods for natural texture classification. They found that filter banks were not necessary because classification based on local $N \times N$ neighborhoods outperformed classification based on filter banks. Obviously, the computation time of this method without filtering was less than that using filter banks. Petroudi and Brady (Petroudi et al. [2003]) applied filter bank classifiers to mammographic pattern classification. Later, they used features from local $N \times N$ neighborhoods for classification (Gong and Petroudi [2006], Petroudi and Brady [2011]). These latter features based on local $N \times N$ neighborhoods performed as well as using filter banks.

Textons continued to be applied widely in image classification. Textons based on global patterns were applied to classify dermoscopic images (Sadeghi et al. [2012]). A texton dictionary was generated from two descriptors: the first named continuous maximum responses used the maximum of the filter responses and the second rectifies the filter responses to calculate principal curvatures of the image surface. This process was used to classify Brodatz images (Zhang et al. [2013]). In the study by Li et al. (Li et al. [2012d]), the local binary pattern based on scale-adaptive textons was found to be promising for texture description and scale invariant texture classification. A multi-stage Bayesian level sets algorithm based on textons learnt from a filter bank similar to maximum response 8 (MR8) (Section 2.3.1) was used to classify human embryonic stem cell colonies (Lowry et al. [2012]). Textons generated from Google Earth Satellite images with the Leung-Malik (LM) filter bank (Section 2.3.1) were used to estimate population density (Javed et al. [2012]). Guo et al. (Guo et al. [2012b]) proposed complex textons generated from the MR8 filter bank for texture image classification of the Outex database (Ojala et al. [2002]).

Generally, the procedure of texton generation, feature extraction and classification can be described by the following five steps (Figure 2.1); (1) extracting local feature vector, (2) clustering into textons, (3) creating texton map, (4) constructing histogram of textons and (5) classification. Operationally, feature vectors of texture primitives (multi-dimensional feature vectors) are usually filter responses obtained

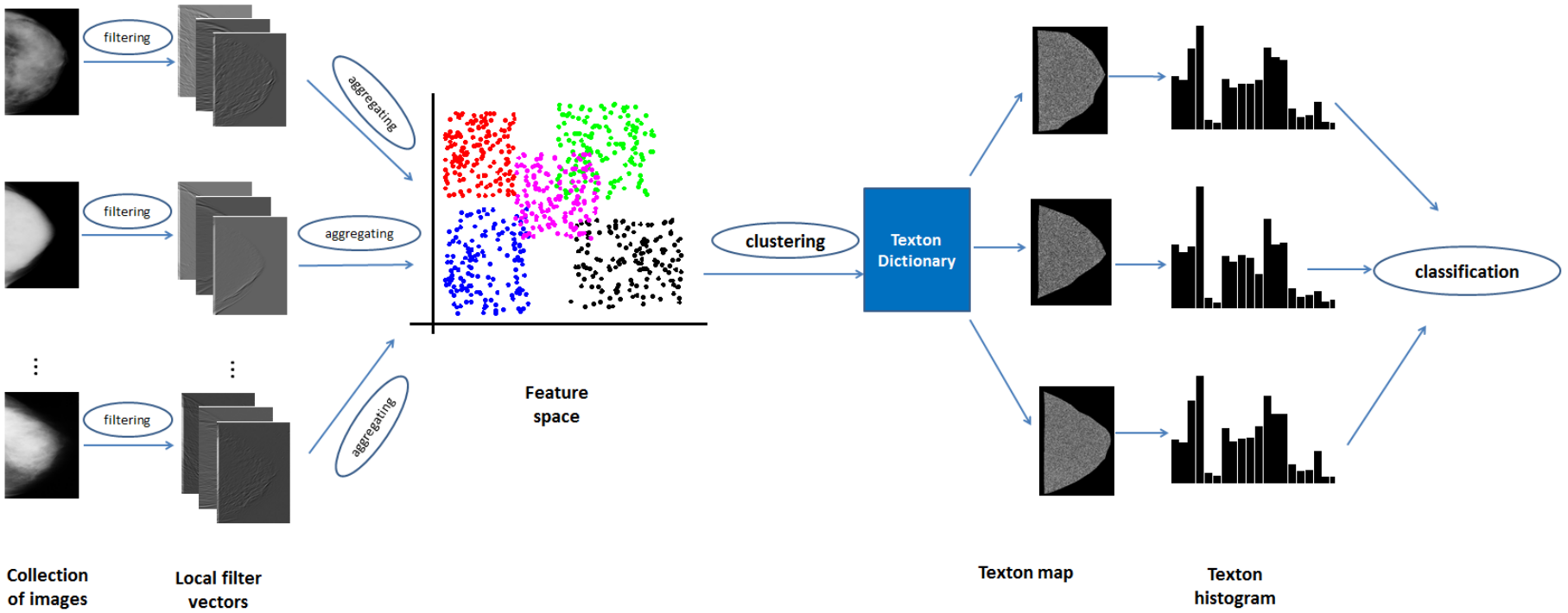

Figure 2.1: Framework of texton generation, feature extraction and classification described in five steps: (1) extracting local feature vector, (2) clustering into textons, (3) creating texton map, (4) constructing histogram of textons and (5) classification. Operationally, feature vectors of texture primitives (multi-dimensional feature vectors) are usually filter responses obtained by applying filter bank on a number of images.

by applying filter bank on a number of images. By applying a clustering method on feature vectors from a set of images, a universal texton dictionary for these images is constructed. By mapping each pixel to the texton nearest to its feature vector of filter responses, an image can be analyzed into texton channels, which are called texton maps. The histogram of each texton map represents the texture features of the corresponding image and is used for final classification. Each of these steps will be described in detail in the ensuing sections.

2.3 Methods for Extracting Local Texture Feature Vectors

Many methods could, in principle, be used to construct feature vectors that capture local texture information. The three most prominent methods in the literature are reviewed below. Local feature vectors from these three methods can be used for texton generation. These methods are also the ones considered during the development of this thesis.

2.3.1 Standard Filter Banks

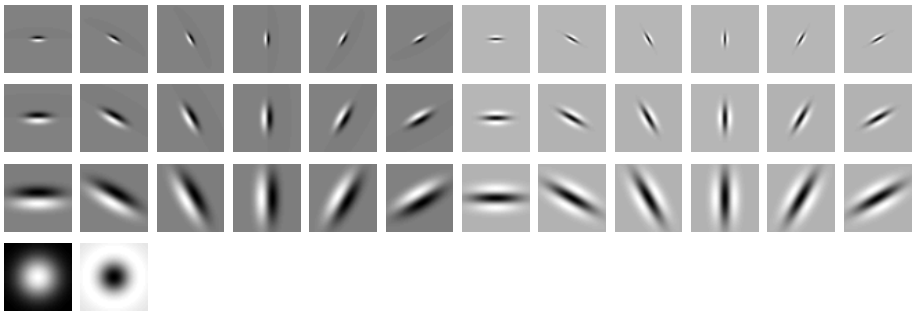

Figure 2.2: Root filter set.

The common root filter set (Varma et al. [2007]) consists of 38 filters: a Gaussian and a Laplacian of Gaussian both with $\sigma = 10$ pixels, an edge filter at 3 scales $((\sigma_x, \sigma_y) = \{(1, 3), (2, 6), (4, 12)\})$ and a bar filter at the same 3 scales (Figure 2.2). The first two filters are rotational symmetric while the latter two are oriented. The MR8 filter bank is a reduced filter bank set derived from the common root filter set. However, the MR8 filter bank is rotationally invariant because only the maximum filter response across all six orientations of every scale for the two anisotropic filters is recorded. As a result, though based on 38 initial filters in the root filter bank,

the MR8 filter bank has only 8 filter responses. If the edge filter and the bar filter, respectively is only at one scale $(\sigma_x, \sigma_y) = (4, 12)$ and only the maximum filter response across all six orientations of this scale for the two anisotropic filters is recorded, the root filter set turns out to be the MR4 filter bank.

The LM (Leung and Malik [2001]) filter set has 48 filters: first and second derivatives of Gaussians at 6 orientations and 3 scales, 8 Laplacian of Gaussian filters and 4 Gaussians (Figure 2.3). The scale of the LM filter bank is between $\sigma = 1$ and $\sigma = 10$ pixels. Due to the structure of LM filter bank, it is rotationally variant.

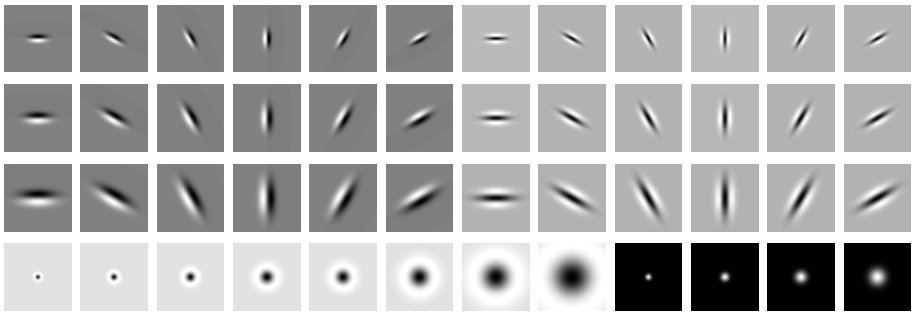

Figure 2.3: LM filter bank.

The Schmid (S) (Schmid [2001]) filter bank contains 13 rotationally invariant and isotropic filters (Figure 2.4) of the form

$$F(r, \sigma, \tau) = F_o(\sigma, \tau) + \cos(\frac{\pi\tau r}{\sigma})e^{-\frac{r^2}{2\sigma^2}},$$

where $F_o(\sigma, \tau)$ is the DC component (the mean value of the waveform). The values of (σ, τ) pair are $(2, 1)$, $(4, 1)$, $(4, 2)$, $(6, 1)$, $(6, 2)$, $(6, 3)$, $(8, 1)$, $(8, 2)$, $(8, 3)$, $(10, 1)$, $(10, 2)$, $(10, 3)$ and $(10, 4)$. Because each element filter of the filter bank is rotationally invariant, on the whole, the S filter bank is rotationally invariant as well.

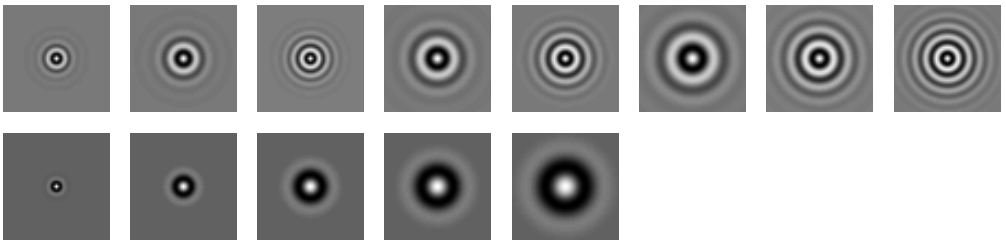

Figure 2.4: S filter bank.

The MR8 filter bank is more widely used than LM and S. Varma and Zisserman (Varma and Zisserman [2005]) showed that better results were obtained by the MR8 filter bank on classifying natural images from the Columbia-Utrecht texture database (Dana et al. [1999]) of 61 natural textures.

2.3.2 $N \times N$ Neighborhoods

In this thesis, the name $N \times N$ neighborhoods will be used to refer to the method introduced by Varma and Zisserman (Varma and Zisserman [2003]) to replace the use of filter banks. In this method, a feature vector is associated to a pixel p by listing the raw image intensity values in the $N \times N$ neighborhood of p. The order of listing the raw image intensity values in terms of the relative positions to p must be the same for each pixel. Studies using local $N \times N$ neighborhoods vary on whether all the pixel values in the neighborhood should be used (resulting in a feature vector of length N^2) or if the central pixel should be excluded (resulting in a feature vector of length $N^2 - 1$). Sometimes, the central pixel is replaced by the mean of its neighborhood. Varma et al. (Varma and Zisserman [2003], Gong and Petroudi [2006]) showed that there is no significant difference between the performance if the central pixel is included or not.

Although the $N \times N$ neighborhood method for generating local texture feature vectors is generally viewed as a departure from extracting feature vectors using filter banks, the method can be realized as a filter bank. The filter bank of size N^2 where filter i comprises an $N \times N$ array with zeros in every position except at position i where the value is one, produces the same feature vectors as the $N \times N$ neighborhood method. However, the fact that these feature vectors may be formed without implementing a filter bank is useful and will play an important role in this thesis (Chapter 6).

In the study by Varma and Zisserman [2003], the performance of $N \times N$ neighborhoods for texture image classification was shown to be at least as good as the MR8 filter bank but with less computation.

2.3.3 Gabor Filters

The Gabor filter is a linear filter used in image processing for edge detection and texture feature extraction. Frequency and orientation representations of Gabor filters have been found to be particularly appropriate for texture representation and discrimination. In the spatial domain, a 2-dimensional Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. 2-dimensional Gabor functions were first proposed by Daugman (Daugman [1985]) for modeling simple cells in the visual cortex of mamalian brains. All Gabor filters can be generated from one mother filter by dilation and rotation. There are three commonly used Gabor filter functions (Equations 2.1, 2.2 and 2.3),

$$\text{Complex: } g(x,y,\lambda,\theta,\sigma,\gamma) = \exp(-\frac{x'^2+\gamma^2y^2}{2\sigma^2})\exp(i(2\pi\frac{x'}{\lambda}+\varphi)) \tag{2.1}$$

$$\text{Real: } \mathrm{Re}(g(x,y,\lambda,\theta,\sigma,\gamma)) = \exp(-\frac{x'^2+\gamma^2y^2}{2\sigma^2})\cos(2\pi\frac{x'}{\lambda}+\varphi) \tag{2.2}$$

$$\text{Imaginary: } \mathrm{Im}(g(x,y,\lambda,\theta,\sigma,\gamma)) = \exp(-\frac{x'^2+\gamma^2y^2}{2\sigma^2})\sin(2\pi\frac{x'}{\lambda}+\varphi) \tag{2.3}$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$. The filter has a real and an imaginary component representing orthogonal directions. The two components can be formed into one to use together or used individually.

In order to extract texture features efficiently from Gabor filters, the parameters of Gabor filters need to be decided according to the particular task. Parameters for Gabor filters constructed from the real Gabor filter function in Equation 2.2 are: (1) the wavelength, λ, in the cosine factor of the Gabor filter kernel, (2) the standard deviation of the Gaussian factor of the Gabor function, σ, (3) the orientation of the normal to the parallel strips of a Gabor function, θ, (4) the phase offset, φ, in the argument of the cosine factor of the Gabor function, (5) the spatial aspect ratio, γ, which specifies the ellipticity of the support of the Gabor function, (6) the half-response spatial frequency bandwidth, b. The parameter b is related to the ratio σ/λ (Equation 2.4). The value of σ is usually specified through the value of b. Similarly, six parameters are defined for the remaining two Gabor filter functions.

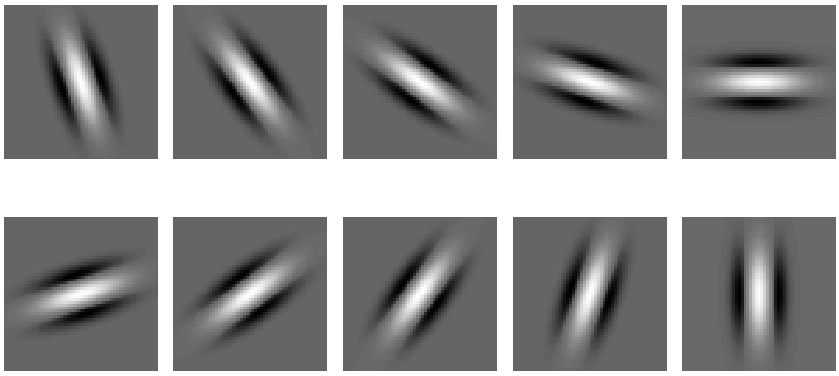

Figure 2.5: An example of a Gabor filter bank consisting of 10 Gabor filters with $\lambda = 20$, $\sigma = 4.2$, $\theta = k\pi/10$ ($k = 1, 2, \ldots, 10$), $\varphi = 0$, $\gamma = 0.4$ and $b = 4$.

$$b = \log_2 \frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{\ln 2}{2}}}, \quad \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\cdot\frac{2^b+1}{2^b-1} \tag{2.4}$$

An example of Gabor filter bank comprising of 10 Gabor filters, constructed according to the real part of the Gabor filter (Equation 2.2) is shown in Figure 2.5.

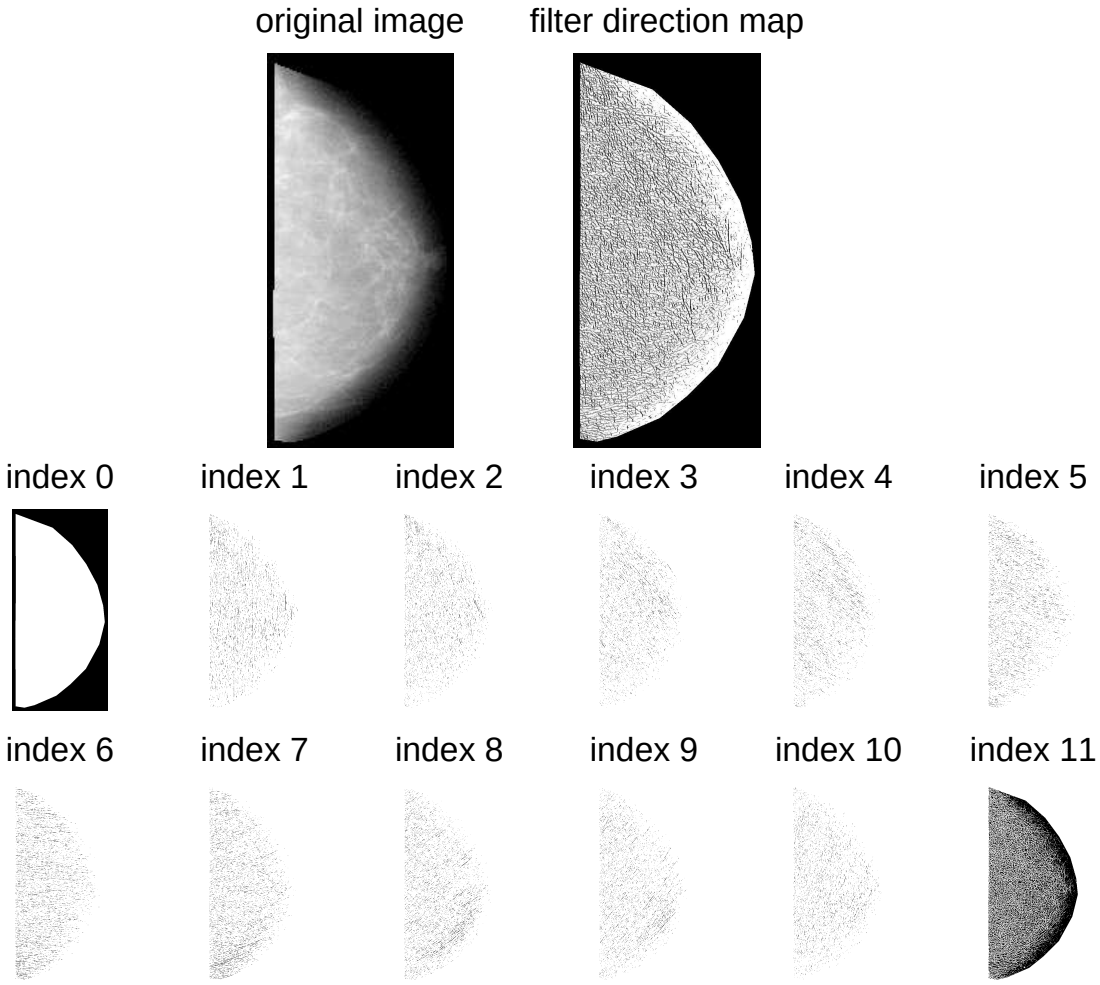

Figure 2.6: Texture structures from a Gabor filter bank. The original image is a cropped screening mammogram. The filter direction map shows the index of maximum orientation as a gray scale image. In this example, pixels outside the breast region are assigned index 0 and pixels with maximum response less than the preset threshold are assigned index 11. The index images 1 - 10 show pixels with maximum response at the orientation corresponding to that index.

In the Gabor filter bank method, feature vectors are elements of filter responses obtained by applying a Gabor filter bank to an image.

In addition, if the focus is on determining the orientation of textures in an image, then the maximum filter response among all the filter responses can be used. Representing each pixel by the index of its maximum filter response direction generates a map of maximum filter directions. In order to show clearer texture information, a threshold for the maximum filter response T should be applied to the filter direction

map and the elements of the map whose corresponding maximum filter responses are under the threshold are ignored. The reason is that if the maximum filter response is very low, there is little or no orientation information and the highest response may be random. Figure 2.6 shows an example of a filter direction map generated by using the example Gabor filter bank in Figure 2.5.

In the maximum Gabor response method, feature vectors are constructed by applying the $N \times N$ neighborhood method to the filter direction map. The reason for applying the $N \times N$ neighborhood method is that the filter direction map consists of maximum filter direction labels, which carry no rank information.

2.4 Clustering Methods

Textons are generated from the array of local feature vectors by clustering. In pattern recognition, three widely used clustering methods are K-means, fuzzy C-means and parametric methods such as Gaussian mixture methods. Only K-means and fuzzy C-means clustering methods will be reviewed here since Gaussian mixture models were not considered in this thesis. The reasons are: (1) In the literature on textons, only K-means and fuzzy C-means have been used. (2) The EM (Expectation-Maximization) algorithm for estimating a certain Gaussian mixture model is closely related to the K-means clustering (Dempster et al. [1977]). Gaussian mixture can be taken as soft K-means clustering (Hastie et al. [2008]). (3) Time constraints did not allow the exploration of the possible effect of using Gaussian mixture model for clustering.

2.4.1 K-means

K-means clustering is a simple exclusive clustering method for learning K cluster centroids from a given data set (MacQueen [1967]). Given an initial set of centroids, the first step is to associate each point to its nearest centroid by minimizing the objective function (a squared error function)

$$O = \sum_{n=1}^{K} \sum_{m=1}^{j_n} ||p_m^n - c_n||^2,$$

where p_m^n denotes the mth element of cluster n, c_n denotes the center of cluster n and j_n is the number of elements in cluster n. In the second step, centroids are re-calculated as the means of points in each cluster.

The above two steps are iterated until there are no changes in class membership. This clustering algorithm is sensitive to the initial selected clustering centroids but this effect can be reduced by running K-means clustering multiple times with differ-

ent initial cluster centers and choosing the cluster centers yielding the smallest value of the objective function.

2.4.2 Fuzzy *C*-means

Fuzzy *C*-means is an overlapping clustering method, which allows one data point to belong to two or more clusters. It was initially proposed by Dunn (Dunn [1973]) and then improved by Bezdek (Bezdek [1973]). This clustering method aims to minimize the object function

$$O_i = \sum_{m=1}^{N} \sum_{n=1}^{C} (u_{mn})^i ||p_m - c_n||^2,$$

where i is the fuzziness index and $i \in [1, \infty)$, u_{mn}^i is the degree of membership of point p_m in the cluster c_n, p_m is the mth multi-dimensional measured data, c_n is the nth multi-dimensional center of the cluster. The dimension of p_m and c_n is the same. $\| * \|$ is any norm expressing the similarity between any measured data and the center. N is the total number of measured data and C is the total number of clusters.

Given a set of data, it is initialized into a single array $U^0 = [u_{mn}]$ with the selected values for C and i.

In the first iteration step ($h = 1$), firstly cluster centers $C^{(h)} = [c_n^h]$ are updated from U^0 according to

$$c_n^h = \frac{\sum_{m=1}^{N} (u_{mn})^i p_m}{\sum_{m=1}^{N} (u_{mn})^i}.$$

Secondly, based on the calculated cluster centers, the initial array U^0 is updated to $U^h = [u_{mn}^h]$ by

$$u_{mn}^h = \frac{1}{\sum_{g=1}^{C} (\frac{||p_m - c_n^h||}{||p_m - c_g||})^{\frac{2}{i-1}}},$$

where c_g is the gth multi-dimensional center of the cluster. In the third step, if $\| U^1 - U^0 \| \geq \delta$, the first two steps are repeated.

The iteration will stop when $\| U^{h+1} - U^h \| < \delta$ or the local minimum is reached, where δ is a termination criterion ($0 < \delta < 1$). For this algorithm, different initializations will lead to different evolutions of the algorithm. Different initialization may require different numbers of iteration to reach the stopping criterion.

2.5 Classifiers

In this thesis, classifiers are used to allocate mammogram images among different classes according to texture features calculated from textons. In this section, three classifiers are described and these will be used to conduct the experiments in the following chapters.

The k-nearest neighbor classifier and the Fisher classifier are at opposite ends of the spectrum of classifiers in terms of bias and variance (Hastie et al. [2008]) and so together provide an overview of classification potential of the features. The SVM classifier was chosen because of the general popularity of this classifier.

2.5.1 Ensemble k-nearest Neighbor Classifier

The k-nearest neighbor classifier does not require a model to be fit. In this classifier, there are a fixed number of classes and known examples of objects (training objects). The goal is to find a decision rule to classify a query point x into the corresponding class. Let n_m be the number of data points in class ω_m and let k be a fixed positive integer. The k-nearest neighbor method works as follows.

For every x in the feature space, find the sphere of volume V that just contains k training members of the data set closest in distance to x. Let k_m denote the number of these elements that belong to class ω_m. An estimate of the conditional probability $P(x \mid \omega_m)$ is

$$\hat{P}(x \mid \omega_m) = \frac{k_m}{n_m V}$$

and similarly, an estimate of the probability $P(\omega_m)$ and $P(x)$ are

$$\hat{P}(\omega_m) = \frac{n_m}{n} \qquad \text{and} \qquad \hat{P}(x) = \frac{k}{nV},$$

where $n = \sum n_m$. Using Bayes rule

$$\hat{P}(\omega_m \mid x) = \frac{\hat{P}(x \mid \omega_m)\hat{P}(\omega_m)}{\hat{P}(x)} = \frac{k_m}{k}.$$

The resulting decision rule is completely intuitive. For a point x in the feature space, find the k data points that are closest to x and classify x as belonging to the class ω_m which has the most representatives among these k nearest neighbors.

Dasarathy and Sheela (Dasarathy and Sheela [1979]) seem to be the first to discuss partitioning the feature space using 2 or more classifiers, thus initiating the notation on ensemble methods. Then, Hensen and Salamon (Hansen and Salamon [1990]) claimed that invoking ensembles of similar neural network classifiers helped reduce the remaining residual generalization errors. Schapire (Schapire [1990])

proved that converting weak classifiers into one strong classifier achieved higher accuracy. Since then, research in ensemble systems has expanded rapidly and numerous ensemble algorithms have been proposed. There are several reasons for using ensemble based systems (Polikar [2006]): (1) Statistically, combining the outputs of several classifiers by averaging may reduce the risk of selecting a poorly performing classifier. (2) Ensemble methods are suitable for both large and small volumes of data. (3) A divide-and-conquer approach dividing the data space into smaller and easier-to-learn partitions means each classifier needs only learn one of the simpler partitions. (4) Ensemble based approaches can be successfully used for applications, where data from different sources are combined to make a more informed decision (referred to as data fusion).

For the application to *k*-nearest neighbor classifiers, random subspace ensembles are used to improve the accuracy. The advantage of random subspace ensembles is that less memory is used.

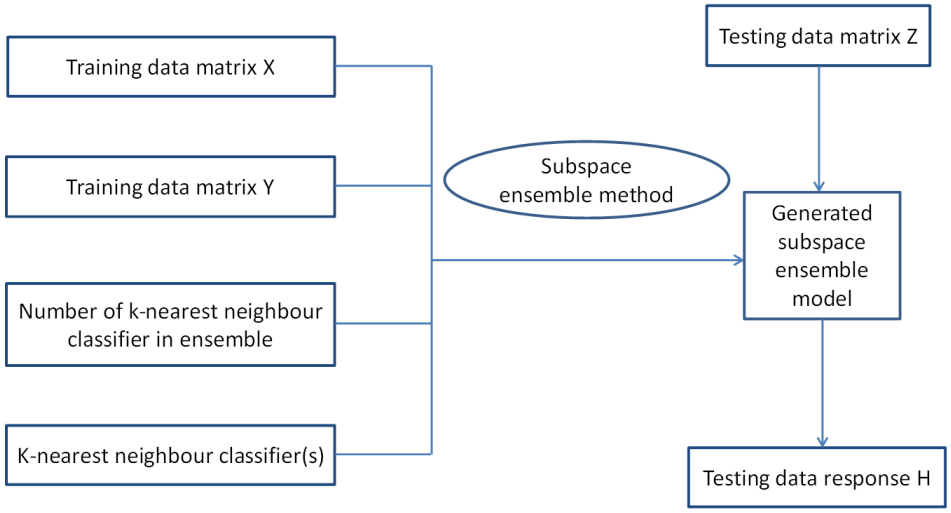

Figure 2.7: Framework for the process of subspace ensemble application on *k*-nearest neighbor classifier.

The random subspace method is a parallel learning algorithm, that is, the generation of each decision tree is independent. For the subspace ensemble *k*-nearest neighbor classifier, each *k*-nearest neighbor classifier is a decision tree and the combined classifier is taken as a decision forest (Ho [1998]). For this method, in each pass, a subspace is obtained by randomly selecting a small number of dimensions from the given feature space and assigning unselected dimensions with a constant value. Projecting training images according to the selected subspace, a decision tree (*k*-nearest neighbor classifier) is constructed with projected training images. An unknown image projected onto the subspace can be classified using the decision tree. Individual trees are combined by averaging the conditional probability of assigning

an unknown datum to each class at the leaves. In the end, the unknown data will be assigned to the class obtaining the maximum average conditional probability.

Practically, the process of classifying novel testing images with generated subspace ensemble model can be displayed by Figure 2.7.

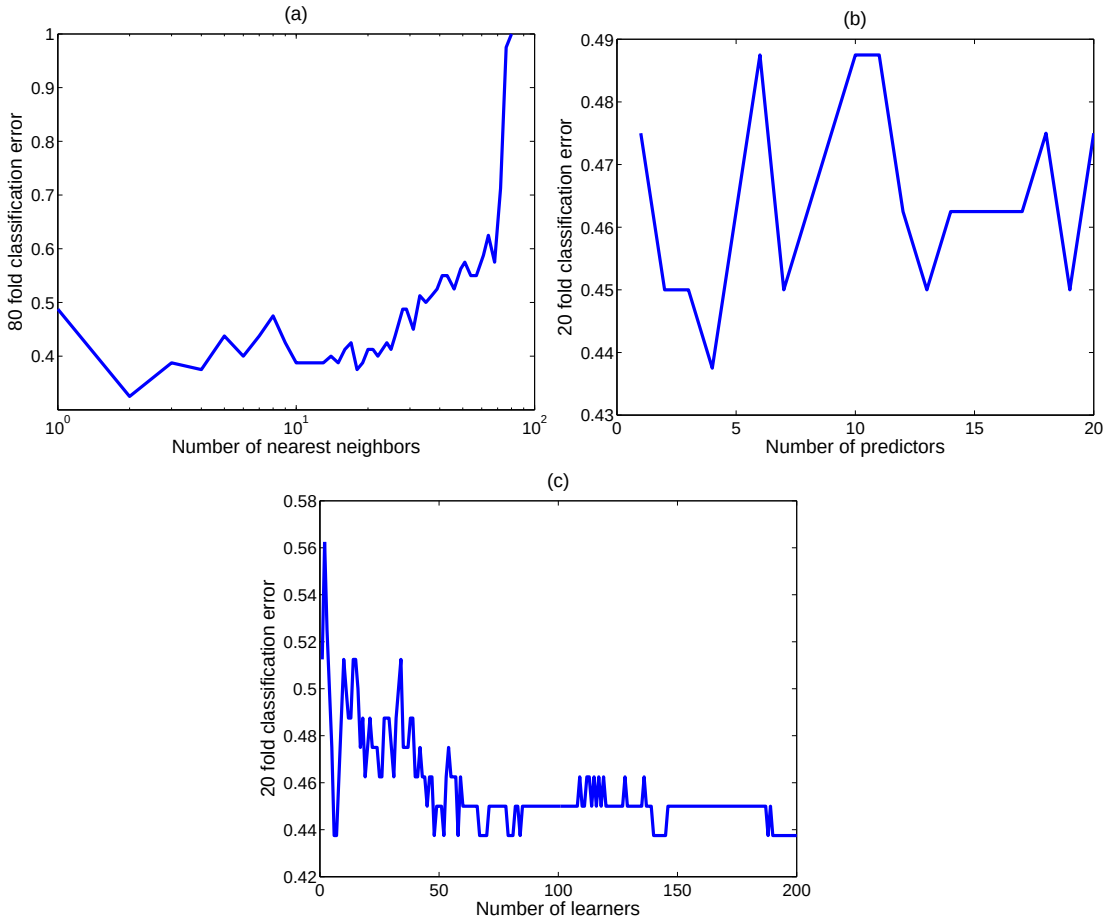

Figure 2.8: The process of choosing three parameters for the subspace ensemble k-nearest neighbor classifier: (a) the cross validation errors for different numbers of nearest neighbors in the k-nearest neighbor classifier, k, (b) the cross validation errors for different numbers of predictors, m (how many features were used), (c) the cross validation errors for different numbers of k-nearest neighbor classifiers, n. From these figures, the number of nearest neighbors k is chosen to be 2, the number of predictors m is chosen to be 4 and the number of weak learners n is chosen to be 69 since reasonable low evaluation errors were obtained at these values.

Three parameters need to be set before applying the final subspace ensemble k-nearest neighbor classifier: (1) the number of nearest neighbors which is the parameter needed to be set for any k-nearest neighbor classifier, k, (2) the number of predictors leading to the smallest cross-validation error, m, (3) the number of learners which is the smallest number of learners in the ensemble that still obtain good classification performance, n. One example of choosing these three parameters is shown in Figure 2.8.

Let D be the total number of predictors of the input data for classification. Implementing the subspace ensemble k-nearest neighbor classifier requires the following five steps:

1. Decide the three basic parameters; k, m and n.
2. Randomly choose m predictors from the D possible total predictors.
3. Train a k-nearest neighbor classifier with the chosen m predictors.
4. Repeat steps 2 and 3 until there are n k-nearest neighbor classifiers.
5. Predict by taking an average of the prediction score of the k-nearest neighbor classifiers and classify the testing input data to the class with the highest average score.

2.5.2 Fisher Classifier

For two ideal spherically shaped clusters of equal radius, optimal classification may be achieved by finding the hyperplane perpendicular to the line connecting the centers of the two group points that lies half way between the two centers of the two groups. If the two clusters are not spherical and not of equal size, finding the best hyperplane to separate the groups is not as obvious. The Fisher classifier finds the optimal hyperplane in this case.

The idea of Fisher classifier as a linear classifier is to find the orientation vector v that maximizes the separation between the two training groups (g_1 and g_2) after standardizing for within group variance. This means the objective is to maximize the ratio

$$R = \frac{\text{distance between groups}}{\text{std within samples}} = \frac{v'\overline{g_1} - v'\overline{g_2}}{\sqrt{v'Sv}},$$

where if $g_1 = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$, $g_2 = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$, then $\overline{g_1} = \begin{pmatrix} \overline{z_1} \\ \overline{z_2} \\ \vdots \\ \overline{z_n} \end{pmatrix}$, $\overline{g_2} = \begin{pmatrix} \overline{y_1} \\ \overline{y_2} \\ \vdots \\ \overline{y_m} \end{pmatrix}$, and S is the variance-covariance matrix.

Differentiating R with respect to v and setting the result equal to zero yields

$$\overline{g_1} - \overline{g_2} = \frac{v'(\overline{g_1} - \overline{g_2})Sv}{\sqrt{v'Sv}}.$$

Further, since the quantities $v'(\overline{g_1} - \overline{g_2})Sv$ and $\sqrt{v'Sv}$ are both scalars, the orientation v can be found by solving $\overline{g_1} - \overline{g_2} = Sv$. Thus, the Fisher direction is given by $v = S^{-1}(g_1 - g_2)$, which is perpendicular to the discriminant plane (Figure 2.9).

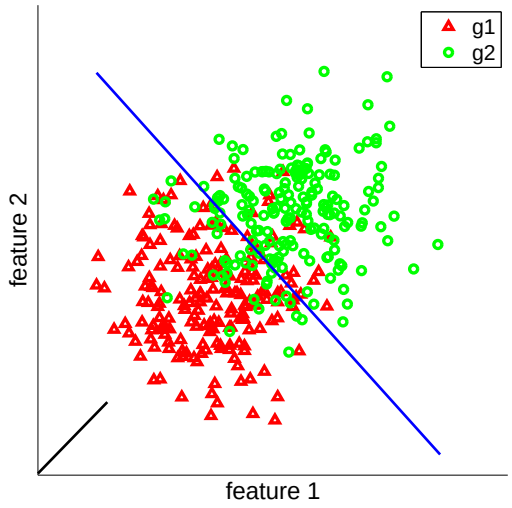

Figure 2.9: An example of Fisher classifier used for classifying two groups. The black line is the Fisher orientation vector and the blue line is the discriminant surface of Fisher classifier.

2.5.3 Support Vector Machine (SVM)

Two groups in a feature space are called perfectly separating if there is a hyperplane such that all the elements of one group lie on one side and all the elements of the other group lie on the other side. Any hyperplane that separates the two groups is called a separating hyperplane. If there is one, there are infinitely many. The hyperplane that maximizes the distance to the closest point in each group is called the optimal separating hyperplane. The Fisher classifier applied to perfectly separating groups does not necessarily find a separating hyperplane and rarely finds the optimal separating hyperplane (Hastie et al. [2008]). Finding the optimal separating hyperplane requires solving the optimization problem given by

$$\max_{\beta_0, \|\beta\|=1} M$$

subject to

$$y_i(x_i^T \beta + \beta_0) \geq M,\ i = 1,2,\ldots,N,$$

where (x_i, y_i), $i = 1, 2, \ldots, N$ constitutes the training data with x_i denoting the ith input vector and $y_i = \pm 1$ indicating the true class membership.

As it stands, this problem is difficult to solve, but is equivalent to the problem (called the Wolfe dual) (Hastie et al. [2008])

$$\max L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

subject to

$$\alpha_i \geq 0.$$

Here the condition $\|\beta\| = 1$ has been replaced by $\|\beta\| = 1/M$. In this version, the objective function is quadratic but the constraints are linear. Standard convex methods provide numerical solutions.

The derivation of the Wolfe dual version also shows that the following conditions (Equations 2.5, 2.6 and 2.7) must be satisfied by the α_i.

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{2.5}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i \tag{2.6}$$

$$0 = \alpha_i \left(y_i (x_i^T \beta + \beta_0) \right), i = 1, 2, \ldots, N. \tag{2.7}$$

These conditions imply the following crucial facts.

1. If $\alpha_i > 0$, then $y_i(x_i^T \beta + \beta_0) = 1$ which means that x_i lies on boundary of the slab of maximal width parallel to, and containing the optimal separating hyperplane that does not contain any observations. The width of the slab is $M = 1/\|\beta\|$.

2. If $y_i(x_i^T \beta + \beta_0) > 1$, then x_i is not on the boundary of the slab and $\alpha_i = 0$.

These facts show that β in Equation 2.5 is defined only in terms of the points x_i on the boundary of the slab. These x_i are called the support vectors.

SVM are classifiers designed in the spirit of the discussion above but for cases where the two groups are not perfectly separating. This is done by introducing variables ξ_i that measure the amount by which $x_i^T \beta + \beta_0$ lies on the wrong side of the margin of the slab of width M. In the language of linear programming, the ξ_i are slack variables. With $\|\beta\| = 1/M$, the resulting optimization problem is

$$\min \|\beta\|$$

subject to

$$y_i(x_i^T \beta + \beta_0) \geq M(1 - \xi_i), i = 1, 2, \ldots, N, \xi_i > 0, \sum \xi_i \leq K,$$

where K is a constant (Hastie et al. [2008]). Bounding the sum of the ξ_i limits the number of points on the wrong side of the slab boundaries.

Following steps similar to the case of perfectly separating groups, a vector $\hat{\beta}$ is

found of the form

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i,$$

where α_i is nonzero only for vectors x_i that lie on margin of the slab or on the wrong side of their respective margin. For those vectors on the margin, $\hat{\xi}_i = 0$. These vectors can be used to find $\hat{\beta}_0$. The decision function is then given by

$$\text{sign}(x^T \hat{\beta} + \hat{\beta}_0).$$

2.6 Validation

The literature does not clearly distinguish validation and cross validation. As a result, the following descriptions of validation and cross validation are based on personal understanding and preference.

The easiest method of validation is resubstitution validation. In this validation, the whole data set is used for both training and testing. This method suffers from the problem of over-fitting, because the algorithm may perform extremely well on the available data but relatively poor on an unseen data set. This method does not predict performance on new data and so it is not commonly used.

Hold-out validation splits the whole data set into two parts: one for training and the remaining for testing. To some extent, this avoids over-fitting since there is no overlap between the two folds. However, the procedure does not use the available data efficiently and the performance may be highly dependent on the choice of training/testing partition. On the other hand, if the partitioned test fold is favourable for training, then the prediction performance may suffer, leading to skewed results. The above problems could be partly solved by repeating hold-out validation multiple times, each time choosing half the data randomly for training and the rest for testing and then measuring the average. This procedure is also named repeated random sub-sampling validation. However, repeated random sub-sampling validation may cause some observations never be selected as validation or selected disproportionately often. It is known that two factors will affect the performance measure; the training set which affects the performance measure indirectly through the learning algorithm and and the test set which affects the performance directly. Cross validation is proposed to reach a compromise.

Cross validation is a method for estimating how well classifier performance generalizes to independent (unseen) data. Cross validation was introduced in the 1930s (Larson [1931]) where one sample was used for regression and another for prediction. This idea was further developed by Mosteller and Wallace (Mosteller and Wallace [1963]). A clear statement of cross validation in the modern sense appeared

in Mosteller and Tukey [1968]. Cross validation is used to measure the reliability of models in the setting of classification (Hastie et al. [2008]).

Cross validation is described in the form of k-fold cross validation. In k-fold cross validation, the data are firstly divided randomly into k equal (or almost equal) folds. Subsequently, there are k iterations of classification. In each iteration, a different fold of data is held-out for validation while the remaining $k-1$ folds of data are used for training. Sometimes, in order to increase the reliability of estimated performance, the k-fold cross validation is run multiple times (Refaeilzadeh et al. [2008]), each time with different assignments of data to folds.

When $k = 2$, the whole data set is divided into two folds (one for training and the remaining one for testing), which is almost the same as hold-out validation. However, in 2-fold cross validation, the fold used for training the first time will be used for testing the second time and vise versa (the fold used for testing at the beginning will be used for training at the second time). The performance is measured as the average of the two runs. So 2-fold cross validation is also called hold-out cross validation.

Leave-one-out cross validation is a special case of k-fold cross validation, where k is exactly the number of instances in the data set. In other words, only one observation in each iteration is used for validation of the classifier trained by the remaining $k-1$ observations. This type of cross validation is unbiased but has high variance, and so may result in unreliable estimates (Efron [1983]). However, this disadvantage may be ignored especially when the data set is very small.

Generally, cross validation methods depend on the value of k used. Cross validation methods with different values of k were compared with each other and 10-fold cross validation was recommended by Kohavi [1995] since the performance was reasonably good and the estimate was nearly unbiased. Hastie et al. recommended 5-fold and 10-fold cross validation as a good compromise between variance and bias (Hastie et al. [2008]).

2.7 Accuracy and ROC Analysis

Accuracy is a commonly used measure of classification performance. By definition, this is the proportion of correct assignments over the total number of assignments. The higher the accuracy, the better the classification performance. However, in the field of medical diagnosis, the application of accuracy for measuring performance is limited and the conclusions made based on accuracy alone should be considered with caution. This is because the consequences of an error in assigning a patient as having no disease when the disease that is actually present is not the same as assigning the patient as having disease when no disease is present (Metz [1978]).

Accordingly, concepts such as sensitivity and specificity were proposed to overcome the limitation of accuracy. Sensitivity is the accuracy of classifying the group of subjects with disease, which is the proportion of correct positive assignments of disease over the total number of the group of subjects with disease. Similarly, specificity is the accuracy of classifying the group of subjects with no disease, which is the proportion of correct assignments of no disease over the total number of the group of subjects with no disease. Instead of using a single measure of accuracy, sensitivity and specificity are used together to represent the medical classification performance.

The receiver operating characteristic (ROC) curve is commonly used to measure medical diagnostic classification performance more meaningfully. For constructing the ROC curve, two indices - true positive fraction (TPF) and true negative fraction (TNF) are considered. TPF is the same as sensitivity and TNF is the same as specificity. Corresponding to TPF, there is another index, named false negative fraction (FNF) . Corresponding to TNF, there is another index, named false positive fraction (FPF) . The relationship between these four indices is

$$\text{TPF} + \text{FNF} = 1 \text{ and } \text{TNF} + \text{FPF} = 1.$$

As illustrated in Figure 2.10 for two overlapping groups, increasing the threshold will reduce both TPF and FPF but increase both TNF and FNF. As a result, it is necessary to select a confidence threshold to achieve a appropriate compromise among sensitivity and specificity. In the original foundation of ROC analysis, the confidence threshold was decided by a human decision maker. This is still true in many medical applications. In machine learning, the threshold is varied at small increments to quantitatively record the balance between sensitivity and specificity.

An ROC curve is a plot of TPF (vertical axis) versus FPF (horizontal axis). The points on the ROC curve are obtained by moving the decision threshold along the decision axis (Figure 2.11). The ROC curve displays a trade-off between the sensitivity and specificity. No matter what form of the two group distributions are, TPF and FPF always increase or decrease together as the decision threshold is changed (Metz [1978]) (illustrated in Figure 2.11). A ROC curve must include the lower left corner (TPF = 0 and FPF = 0) of the graph since, by setting the threshold to ∞, all the subjects will be classified as negative, and across the upper right corner (TPF = 1 and FPF = 1) of the graph since, by setting the threshold to $-\infty$, all the subjects will be classified as positive. In addition, a proper ROC curve should be above the major diagonal of the ROC space and the slope of the ROC curve will steadily decrease as it goes up and to the right on the curve. ROC analysis provides a description of medical image classification ability that is independent of both disease prevalence

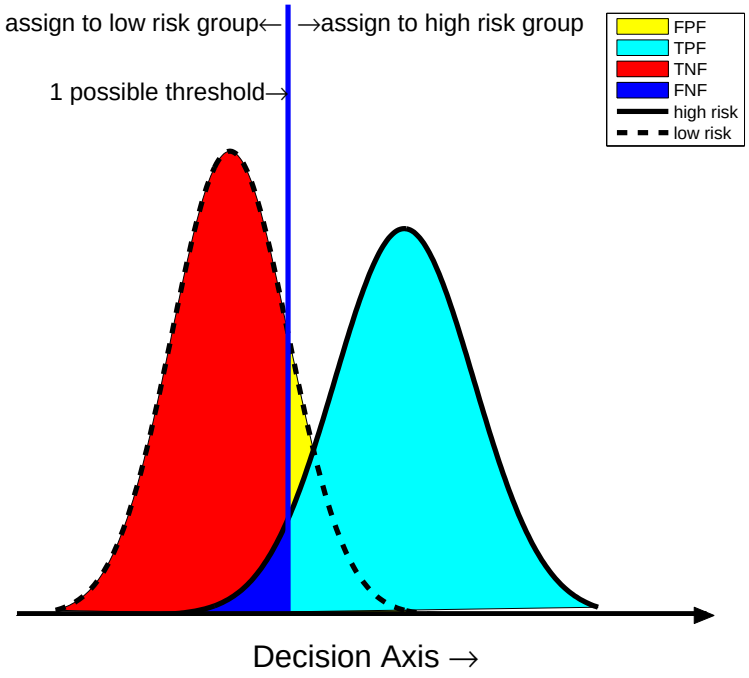

Figure 2.10: Illustration of four decision fractions defined by a possible decision threshold. The group with solid line represents high risk and the group with dashed line represents low risk. The blue line perpendicular to the decision axis is one possible decision threshold. The cyan colour patch indicates the TPF, the blue colour patch indicates the FNF, the red colour patch indicates the TNF and the yellow colour patch indicates the FPF. By moving the decision threshold line along the decision axis, different four decision fractions are defined.

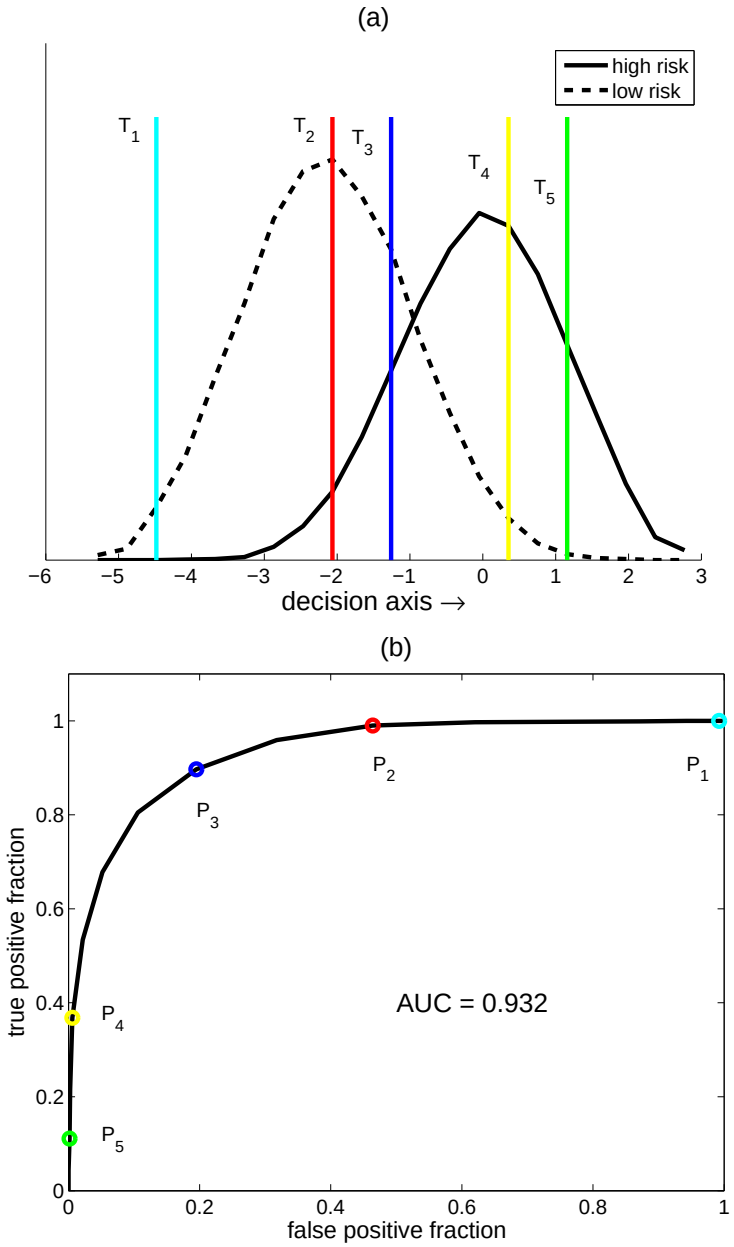

Figure 2.11: Illustration of the process of generating the ROC curve: (a) Shows the four decision fractions for each of five different decision thresholds. (b) Shows five points on the ROC curve corresponding to the five decision thresholds in (a). P_1 corresponding to T_1, P_2 corresponding to T_2, P_3 corresponding to T_3, P_4 corresponding to T_4, P_5 corresponding to T_5.

and decision threshold effects because both TPF and FPF are independent of disease prevalence.

On one hand, if the prevalence of the disease is very low, the FPF needs to be maintained low, otherwise nearly all positive decisions will be false positive ones. Thus the decision maker should focus on the lower left part of the curve to keep FPF small. On the other hand, if the prevalence of the disease is very high, the accuracy of finding true positive cases is meaningful and TPF needs to be maintained reasonable high. Thus the decision maker should operate on the higher right part of the curve to allow a high FPF and low FNF.

The above description of varying the decision threshold for obtaining a number of pairs of TPF and FPF values is generally the initial method of generating ROC curve in practice. An ROC curve can be generated from subjective Yes/No response data by asking the human observer to read all the cases several times with a different decision threshold each time. It is suitable for diagnostic tests that yield a single quantitative value for each case (Yes or No). However, it is impractical for diagnostic tests that cannot be interpreted objectively by human observers because they can not associate continuously numerical values according to their subjective impressions of certainty. Therefore, in practice, the "rating method" was developed in experimental psychology (Green and Swets [1974]). In this method, human observers are required to give their judgment by selecting one of several ratings/categories of confidence and then a pair of TPF and FPF value is calculated for each rating confidence. The ROC curve is generated by fitting these few points statistically on the graph.

Even though the rating method is practical, it tends to be subject to statistical error because the number of cases for plotting the ROC curve is limited and human decisions are not always reproducible. As a result, the procedure of curve fitting is used to help draw a smooth curve that almost goes through the plotted points. The maximum likelihood curve fitting procedure is widely used for this purpose (Metz [1978]). This curve assumes that the underlying distribution is normal.

In order to summarize classification performance over all operating points, the area A bounded above by the ROC curve and below by the horizontal axis is often reported. If classification is perfect, $A = 1$ and if classification is random $A = 1/2$. There are two methods commonly used to estimate A. The theoretical ROC curve based on the normal assumption of the data can be used to compute A directly. Thus the constant A is denoted by A_z where the letter z refers to the normal deviate of the decision variable (Swets [1979], Metz et al. [1998]). Alternatively, A is estimated by using the operating points to perform a numerical integrations. In this case, estimate of A is called the area under the curve (AUC). The higher the AUC score, the better the general classification performance.

A_z is the natural choice in human observer studies since only very few operating

points can be set. The AUC is natural in machine learning situations since there is essentially no limit to the number of operating points. Thus numerical integration gives accurate values for the area under the ROC curve and does not require assumptions regarding the underlying distributions.

In this thesis, all ROC curves are generated by machine classification and so the AUC is used to report classifier performance.

2.8 Feature Selection

The purpose of feature selection is to select a small number of valuable features for a classification task rather than using the full set of features. There are several reasons for doing feature selection. First of all, features may be expensive to obtain and fewer features mean less computational cost. In other words, feature selection provides faster and more cost-effective predictors. Secondly, the presence of ineffective features often cripple the performance of a classifier on test observations. For example, when the classifier is over trained with a large dimensional feature space and limited number of training instances, the performance of classifying unknown test samples can be poor (Fukunaga [1990], Chan et al. [1998, 1999], Bottema et al. [2000]). As a result, feature selection can improve the performance of predictors. In addition, in order to reach a trade off between high classification performance and a small number of input features, it is reasonable and important to ignore features with little effect on the output because the classification goal is to approximate a underlying function between the input and output. So feature selection can provide better understanding of the underlying process that generated the data.

In the field of statistics, feature selection is referred to as feature subset selection. Given a feature set $F = \{F_i | i = 1 \ldots n\}$, feature selection is used to find a subset $f = \{F_{i_1}, F_{i_2}, \ldots, F_{i_m}\}$ with $m < n$, that optimizes an objective function.

Generally, feature selection requires a search strategy for selecting candidate subsets and an objective function to evaluate the performance of these candidate methods.

Search strategies can be grouped into three categories: exponential algorithms such as exhaustive search, sequential algorithms such as sequential forward selection and sequential backward selection, and random search algorithms such as genetic algorithms.

Objective functions (also called evaluation functions) are sometimes categorized into the method of filters, evaluating feature subsets by their information content such as interclass distance, statistical dependence or information-theoretic measures and the method of wrappers, evaluating feature subsets by their predictive performance through statistical resampling or cross validation (Langley [1994]). In the

study by Dash et al. (Dash and Liu [1997]), objective functions were grouped into five categories: distance (or separability), information (or uncertainty), dependence, consistency and misclassification rate. No matter which category of the objective function is applied, the purpose is to evaluate the goodness of a feature subset produced by a certain search strategy. Commonly, the criterion used for classification models belongs to the method of wrappers, which is usually the misclassification rate.

Two feature selection methods used in this thesis are reviewed below. These two feature selection methods are named according to the search strategies.

2.8.1 Exhaustive Search Feature Selection

As the name implies, exhaustive search feature selection methods evaluate all possible feature combinations in order to determine the one leading to the best value of objective function. If the original set of features is of size n, then there are $2^n - 1$ possible feature subsets. For this reason, exhaustive search feature selection is practical only for relative small values of n. However, exhaustive search feature selection is one of the few methods that guarantees finding the optimal feature subset.

2.8.2 Sequential Feature Selection

Sequential feature selection methods can be divided into four general categories: sequential forward feature selection, sequential backward feature selection, bidirectional feature selection, and sequential floating feature selection. Search algorithms for sequential feature selection employ the technique of stepwise regression for searching candidate subsets.

Sequential forward feature selection starts from an empty candidate set. Features are sequentially added to the candidate set until the addition of further features does not improve the objective function. Sequential forward feature selection generally performs best when the optimal feature subset is small. In contrast, sequential backward feature selection starts from a full candidate set and features are sequentially removed from the candidate set until the removal of further features does not improve the objective function. Sequential backward feature selection generally works best when the optimal feature subset is large.

Bidirectional feature selection is a parallel implementation of sequential forward and backward feature selection. Sequential forward selection is performed from the empty set at the same time sequential backward is performed from the full set. Both directions will converge to the same solution on the condition that features already selected by sequential forward selection are not removed by sequential backward

selection and features already removed by sequential backward selection are not selected by sequential forward selection.

Sequential floating feature selection contains two methods: sequential floating forward selection and sequential floating backward selection. For the former, after each forward selection step, backward steps are performed until the evaluation of the objective function improves. As for the later, after each backward selection step, forward steps are performed until the evaluation of the objective function improves.

2.9 Computer-aided Breast Cancer Risk Assessment

Many of the techniques of pattern recognition and texture analysis reviewed in Sections 2.1 - 2.8 have been used for computer-aided breast cancer risk assessment. Computer-aided breast cancer risk assessment is the ultimate objective of the thesis. This section introduces the motivation, the role and the main steps of computer-aided risk assessment (Sections 2.9.1 and 2.9.2) and reviews the history of computer-aided risk assessment (Section 2.9.3), serving as the context of the thesis (Section 2.9.4).

2.9.1 Motivation for and Role of Computer-aided Risk Assessment

Radiologists are highly trained and skilled, so why develop automatic techniques for predicting breast cancer risk by analyzing screening mammograms? First of all, decisions made by humans are generally subjective and qualitative while computers provide objective and quantitative analysis (Rangayyan [2005]). Importantly, decisions and analysis made by radiologists can vary from person to person or from time to time. This may be due to the difference in knowledge, variations in training, and level of understanding. With computers, quantitative analysis becomes possible, computed results are consistent and computers can perform repetitive tasks independently. Secondly, radiologists can be tired or affected by environmental factors and personal circumstances, leading to human error. Third, while the true cost for human and machine readers are both difficult to estimate (Taylor et al. [2010]), a system using a human reader working with a computer has the potential to be more cost effective than two human readers.

Computer-aided breast cancer risk assessment may become an important component of risk assessment but it is unlikely to replace existing risk estimates. Mammogram images are important for risk assessment, but there is some other significant information that is not amenable to quantification or logic rule-based processes, including the mental state of the female, family history, and socio-economic factors. The results of image analysis obtained through computers from screening mammo-

grams should be integrated together with other patient information for comprehensive risk assessment. Certainly, the aggregation of quantitative and objective analysis facilitated by computers and qualitative and subjective analysis realized by human experts will lead to a more accurate breast cancer risk assessment.

2.9.2 Steps for Conducting Computer-aided Risk Assessment

Computer-aided breast cancer risk assessment is conducted by taking advantage of advanced pattern recognition and image analysis techniques (Sections 2.1 - 2.8). Generally, there are seven main steps. The first step is making a decision on the criterion for defining high risk and low risk groups. The second step is acquiring data based on the surrogate of true risk decided in the first step. In the third step, image preprocessing is used to correct the original mammogram images before further image analysis. Examples of image preprocessing step include removing labels and artifacts, adjusting the resolution of the image and normalizing the image. ROIs for risk assessment are extracted in the fourth step. This step is optional depending on the actual task. Features used for classifying images into different risk groups are extracted in the fifth step. The sixth step is needed only if feature selection is necessary. In the last step, classifiers are designed to automatically distinguish high and low risk.

2.9.3 History of Computer-aided Risk Assessment

There is no definition of risk that is both general and practical in the context of developing algorithms for estimating risk of developing breast cancer. Developing an algorithm to estimate risk is very different than developing an algorithm to detect breast cancer, for example. There are reliable methods for establishing the true disease state of women (gold standard) against which results of an algorithm for automatic detection can be compared. Estimates of risk provided by an algorithm must be tested against the true risk of developing breast cancer. Risk only applies to populations, but in validating algorithms, must be assigned per individual. Any study assessing breast cancer risk suffers from the problem of identifying those truly "at risk". A definitive statement is not possible as subjects free of cancer at the end of a study may still be at risk of developing cancer at a later time. Thus the true risk of breast cancer for the purpose of validation is never known and is not even well defined. Instead, a surrogate for the true risk of developing breast cancer must be adopted in order to test estimates of risk generated by the algorithm. In the absence of a gold standard, various surrogates (Table 2.1) have been developed by a

Table 2.1: Surrogates of risk used in the literature on computer-aided breast cancer risk assessment and in this thesis. "*" denotes surrogates for risk used in this thesis.

risk criteria	existing risk surrogates
1. breast patterns	Wolf pattern classes
	SCC categories
	Four density pattern classes corresponding to percentages of density of $< 5, 5-25, 25-75$ and $75-100$
	BI-RADS classes *
	Tabár classes
2. genetics	ER subtype specific classes (high risk) and control cases (low risk)
	BRCA 1/2 (high risk) and control cases (low risk)
3. Disease state	breast images from cancer cases (high risk) *
	breast images from non-cancer cases (low risk) *

number of researchers as computer-aided risk assessment evolved using established risk factors described in Section 1.3.

As the connections between breast tissue structure, breast density and cancer risk emerged (Section 1.3), computer-aided systems were developed to classify mammogram images into different breast patterns for predicting breast cancer risk. In an early work, Magnin et al. (Magnin et al. [1986]) used the spatial gray-level dependence method and gray level difference method, both based on co-occurrence matrices, to quantify density variations in mammograms to characterize images into Wolfe pattern classes (Section 1.3.1) for breast cancer risk evaluation. Caldwell et al. (Caldwell et al. [1990]) computed fractal dimensions in mammograms to classify them into the four Wolfe pattern classes. Tahoces et al. (Tahoces et al. [1995]) extracted texture features based on the Fourier transform, spatial relationships among grey levels and absolute values of the grey levels from three different ROIs in CC view images to categorize mammogram images into Wolfe pattern classes. Byng et al. (Byng et al. [1996]) extracted texture features from fractal dimensions and grey-level histograms to classify images according to SCC categories (Section 1.3.1). The year after, the same group used texture features calculated from regional skewness and fractal dimension to characterize images into corresponding SCC categories (Byng et al. [1997]). Li et al. predicted breast cancer risk by classifying mammogram images into their modified SCC density classes with mammographic density computed by a computer program called Cumulus (Byng et al. [1994]) and their automated measure, which mimics Cumulus (Li et al. [2012a]). Karssemeijer (Karssemeijer [1998]) applied two classifiers with and without pectoral features computed from grey-level histograms to categorize images into four density pattern classes, corresponding to percentages of density of $< 5, 5-25, 25-75$ and $75-100$. An automated image analysis tool was developed for classifying breast images into

four density classes based on characteristics of gray level histogram (Zhou et al. [2001]). Later, the same group used this tool to evaluate the accuracy of using mammograms for estimating breast density (Wei et al. [2004]). Their study was carried out by analyzing the correlation between the percent mammographic dense area and the percent glandular tissue volume estimated from MR images. In 2003, Petroudi and Brady (Petroudi et al. [2003]) proposed the application of features extracted from textons generated from filter bank responses to classify mammograms into the four BI-RADS classes (Section 1.3.1). Later, the same group used texton features generated from local $N \times N$ neighborhoods to characterize images into BI-RADS classes (Petroudi and Brady [2006]). Texton features from local $N \times N$ neighborhoods outperformed the filter banks in their previous paper. Texture features calculated from co-occurrence matrices were used to classify mammogram images into BI-RADS classes (Oliver et al. [2008]). Two measures of breast cancer risk; the Gail and Clause risk estimates and mammographic breast density calculated from digital breast tomosynthesis and digital mammography, respectively were compared in classifying images into slightly modified SCC categories for estimating breast cancer risk (Kontos et al. [2009, 2011]). More recently, He et al. (He et al. [2012]) used textons to generate grey-level histograms to classify images into Tabár and BI-RADS classes. More detail can be found in their previous work (He et al. [2009]), where the term "cluster center" is used to describe what is now commonly called "texton". In addition, texture features calculated from the matrices consisting of the frequencies or the probabilities of the texton co-occurrences were used to classify images into BI-RADS classes (Petroudi and Brady [2011]).

In addition to using breast patterns as the criteria for defining true risk, surrogate true risk criteria related to genetics have been used in computer-aided risk assessment as well. Karemore et al. (Karemore et al. [2012]) calculated texture features from Gaussian derivatives at four different scales to classify Estrogen-Receptor (ER) subtype specific classes (ER-positive vs ER-negative, high risk) and control cases (low risk). They obtained a best performance of AUC score $= 0.71$ by combining breast density and texture features. In addition, genetic markers such as mutations in BRCA 1/2 have been used in this context to define high risk group (Huo et al. [2000, 2002], Li et al. [2004, 2006, 2007, 2008, 2010]). An AUC score of 0.88 was achieved for classifying 256×256 square ROIs extracted from full-field digital mammogram images (Section 1.2.1).

All of the above criteria are reasonable but none measure risk directly. Recently, the surrogate true risk criteria for breast cancer in some studies was based not on the mammogram appearance but the chance of developing breast cancer in the future. As a result, two risk groups were defined: a high risk group consisting of images of breasts unaffected by cancer from cases in which cancer was found in the other

breast, and a low risk group consisting of breast images with no cancers from non-cancer cases. For example, mammographic parenchymal pattern measure was computed and compared with age and percent density in assessing breast cancer risk in a case-control study (Wei et al. [2011]). An AUC score of 0.78 ± 0.04 was obtained by the model of six combined measures. Keller et al. (Keller et al. [2012]) studied breast cancer risk temporally with features calculated from density and morphology. A combined area-volumetric model for density resulted in an AUC score of 0.70. An automated and objective measurement of the grayscale value variation within a mammogram was compared with the percent density in estimating breast cancer risk in a case-control study (Heine et al. [2012]). The best AUC score they achieved was 0.76. Zheng et al. calculated bilateral mammographic density asymmetry related features to classify positive cases (high risk of developing breast cancer) and negative cases (not recalled) (Wang et al. [2011], Zheng et al. [2012]). The highest AUC score they achieved was 0.781 ± 0.023. A fully automated software pipeline was developed by quantitatively measuring both breast density and texture properties in the case-control breast cancer risk assessment study of Zheng et al. (Zheng et al. [2013]). This study extracts texture features for points on a spatial regular lattice and from a surrounding window of each lattice point, resulting features that characterize the local mammographic appearance throughout the whole breast. An AUC score of 0.75 was achieved by combining percentage density and texture measures calculated from post-processed images with a window size of $6.3mm^2$.

In this thesis, various surrogates for risk are used as dictated by: (1) the data available, (2) the specific objectives of a particular experiment, (3) the necessity for comparing results to previously published work, and (4) to examine the dependence of risk estimates on choice of surrogate (Table 2.1).

2.9.4 Context of the Thesis

The problem of establishing a criterion for true breast cancer risk in a study on estimating breast cancer risk has been addressed in various ways as described in the previous section. Studies based on surrogates for risk such as carriers of BRCA 1/2 mutations, for example, produce good results but generate methods valid only for this cohort. The method has not been tested on women at high risk for other reasons and so the benefit to the general population is not known.

In this thesis, the view is taken that the criterion for assigning true risk should be based on whether a particular woman actually develops breast cancer at some time in the future or not. While many preliminary experiments in this thesis use data from current screening rounds only (for practical reasons), the final experiments are longitudinal and compare estimates of risk to whether the woman developed

mammographically detectable breast cancer two or four years later (Chapter 8).

In addition, the focus of this thesis is on the contribution of image texture information to quantify risk. Breast density is a known risk factor and standard methods exist for measuring density from screening mammograms. Density and texture are not necessarily independent and so in this thesis steps are taken to separate density from texture in order to evaluate the contribution of texture separately.

Finally, in the course of the study, the scope of textons (Section 2.2) as a device for quantifying texture is extended to higher order textons designed to capture patterns of local texture patterns (Chapter 6).

Chapter 3

Local Normalization: A Preliminary Study

A study preliminary to the main thesis project was conducted to ascertain if textons could be used, in principle, to distinguish tissue associated with cancer from normal tissue (Li et al. [2012c]). This study was conducted before the methods used in the main body of the thesis (Chapters 4 - 8) were fully developed. The work was also conducted on a limited data set of images. This study is not strictly a risk study since the focus is on distinguishing ROIs with known cancer from ROIs without known cancer. However, to some extent, ROIs with cancer can be taken as the absolutely high risk group while ROIs without cancer can be regarded as low risk group (the third risk criteria in Table 2.1). The study confirmed that the use of textons to distinguish tissue types is feasible and hence using textons to assess risk could be worthwhile. Thus this study provided the green light for the investigation leading to this thesis.

The reason for including this preliminary study here is that an important discovery was made that impacted both the processing steps and the focus of the main project. In particular, a naive implementation of textons lead to reasonably good classification of ROIs with cancer and ROIs without cancer. Upon further inspection, it transpired that the alleged texture features corresponded to breast density despite the fact that zero sum filters were used to extract preliminary texture features. This is explained by the fact that intensity response curve of the image acquisition process is not linear (the digitization process is also nonlinear). Accordingly, local variation due to noise or low contrast structure is intensity dependent. Hence, simply subtracting local background does not suffice to remove the signal in the image due to breast density.

Thus, a simple method for local normalization was devised for removing both the local mean and local variation and so arrive at density independent texture features. The focus of the study was revised to measure the potential of texture in assessing

breast cancer risk independent of density.

The experiment leading to these conclusions is described in Section 3.1, the local normalization method inspired by this experiment is described in Section 3.2 and, for completeness, the performance of the resulting density independent texture features in classifying ROIs as cancer or non-cancer is presented in Section 3.3.

3.1 Classifying ROIs as Cancer or Non-cancer

This preliminary study is presented in four sections. Section 3.1.1 describes the data used in this experiment, Section 3.1.2 describes the experimental details, results are presented in Section 3.1.3 and a brief discussion and conclusion are provided in Section 3.1.4.

3.1.1 Data

In total, 89 cancer cases with known malignant cancer were selected from the DDSM database (Section 1.5). Cases were selected only if the manifestation of the cancer was in the form of a mass. Cases were not included if microcalcifications associated with the cancer were present.

Of the original 89 cases, 49 were randomly selected for training and 40 were reserved for testing. Annotation information from the DDSM database was used to extract cancer ROIs from both CC and MLO views of the breasts with malignant mass. This resulted in 100 training and 92 testing cancer ROIs (in some breasts, more than one malignant mass was detected). For each cancer ROI, a corresponding non-cancer ROI was extracted from the contralateral unaffected breast at the location symmetric to the cancer ROI (Figure 3.1), resulting in 100 training and 92 testing non-cancer ROIs. Templates for ROIs were saved for future use.

3.1.2 Experimental Details

In the image preprocessing step, selected images from DDSM database were read in automatically and preprocessed by applying image contrast enhancement. Here, the image contrast enhancement was realized by mapping the values of the input intensity image to new values such that, 1% of the data was saturated at low and high intensities of the input data.

MR8 filter bank (Section 2.3.1) was applied to the preprocessed images and so each pixel was represented by an 8-dimensional primitive feature vector. ROIs were extracted from filtered images by applying ROI templates (Section 3.1.1) and removing redundant areas. The reason for applying the filter bank before ROI extraction

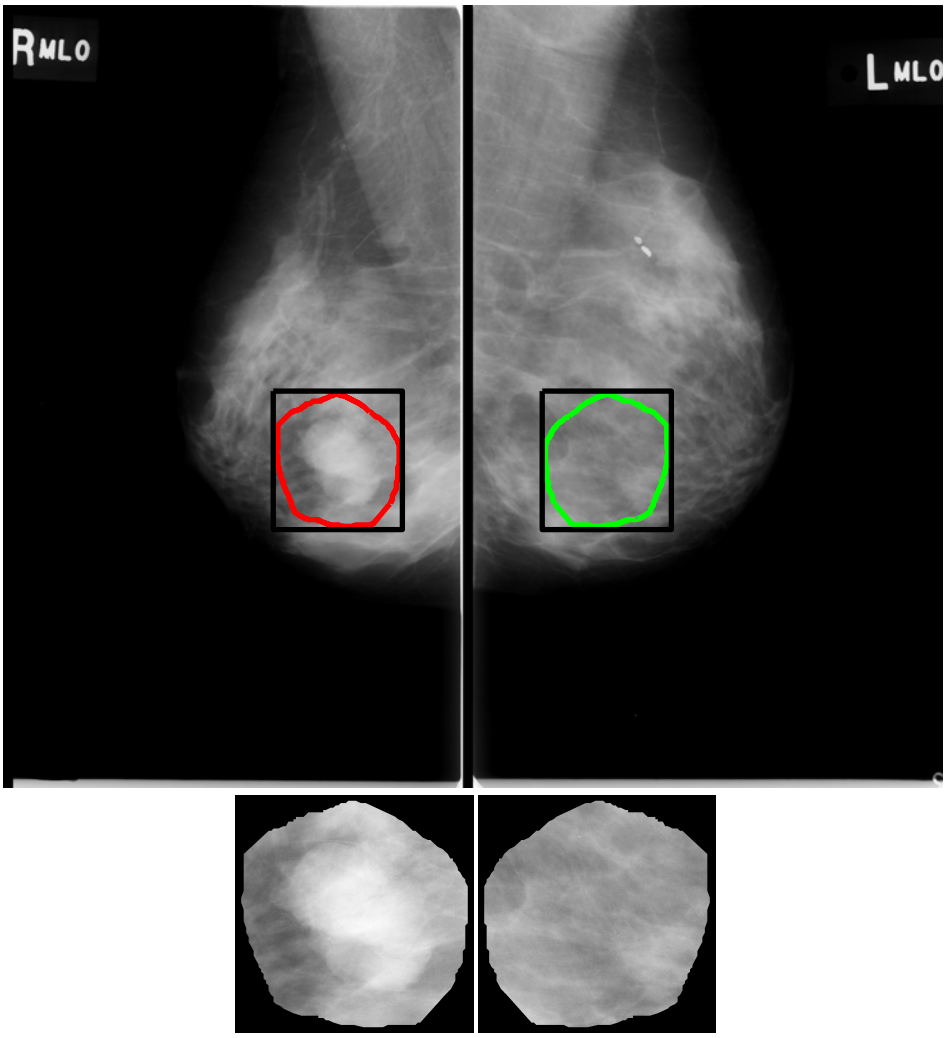

Figure 3.1: An example of choosing cancer and non-cancer ROIs from the MLO view breast images. On the left of the top row is the right MLO view breast, the circled region indicates a malignant mass region located by an experienced radiologist and the square box is the cancer ROI. On the right of the top row is the left MLO view breast from the same woman. The circled region is the corresponding non-cancer region obtained by symmetry and the square box is the non-cancer ROI. The bottom row shows the extracted cancer and non-cancer ROIs.

was to avoid the edge effects.

In the step of texton dictionary generation, the number of feature vectors was reduced by a factor of 25 so that each region of 5×5 pixels in the ROI was represented by a single 8-dimensional feature vector. Separate feature spaces were constructed from cancer and non-cancer feature vectors.

K-means clustering (Section 2.4.1) was applied separately to the feature space of non-cancer ROIs and the feature space of cancer ROIs, resulting in two texton sub-dictionaries. The number of textons (K) needed to separate the two classes of ROIs was not known ahead of time and so this experiment was repeated for seven values of K ($K = 4$, 8, 10, 13, 15, 17, 20), resulting in seven texton sub-dictionaries for non-cancer ROIs and seven texton sub-dictionaries for cancer ROIs. Non-cancer and cancer sub-dictionaries of the same size were combined to form seven final texton dictionaries of sizes 8, 16, 20, 26, 30, 34 and 40.

Next, the texton map was generated for each ROI and the normalized histogram of textons over the texton map was taken as the final texture representation of each ROI (Section 2.2 in Chapter 2). The components of the normalized texton histogram were adopted as features for classification. The Fisher classifier (Section 2.5.2) was used to generate ROC curves and compute AUC scores (Section 2.7, Figure 3.2). The set of training ROIs was used to determine classifier parameters (the Fisher vector direction) and the set of testing ROIs was used to determine classifier performance on unseen data.

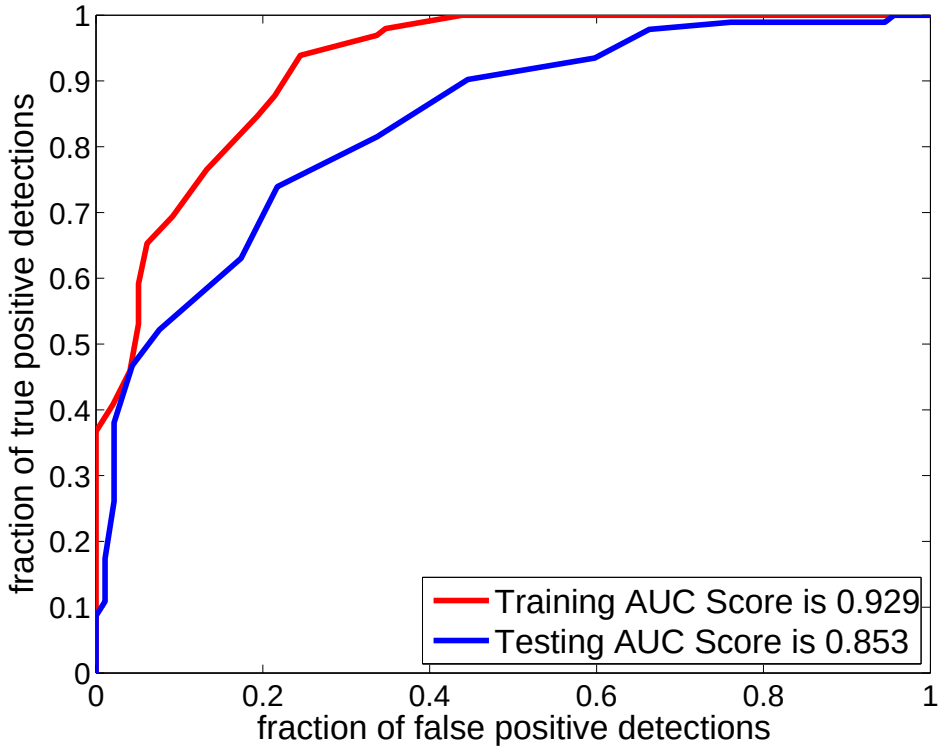

Figure 3.2: ROC curves for classifying ROIs as cancer or non-cancer with 40 textons.

3.1.3 Results

AUC scores for classifying ROIs as cancer or non-cancer with seven different sized texton dictionaries described above are shown in Table 3.1. Since the final texture features used for ROIs classification are the texton frequencies of the texton map, it is useful to inspect the texton map. Figure 3.3 shows an example of texton maps for the whole cancer breast image and the contralateral non-cancer breast image with 16 textons.

Table 3.1: AUC scores for classifying ROIs as cancer or non-cancer described in Section 3.1.2.

texton dictionary size	training AUC scores	testing AUC scores
8 textons	0.844	0.764
16 textons	0.842	0.703
20 textons	0.794	0.731
25 textons	0.884	0.807
30 textons	0.910	0.802
34 textons	0.923	0.821
40 textons	0.929	0.853

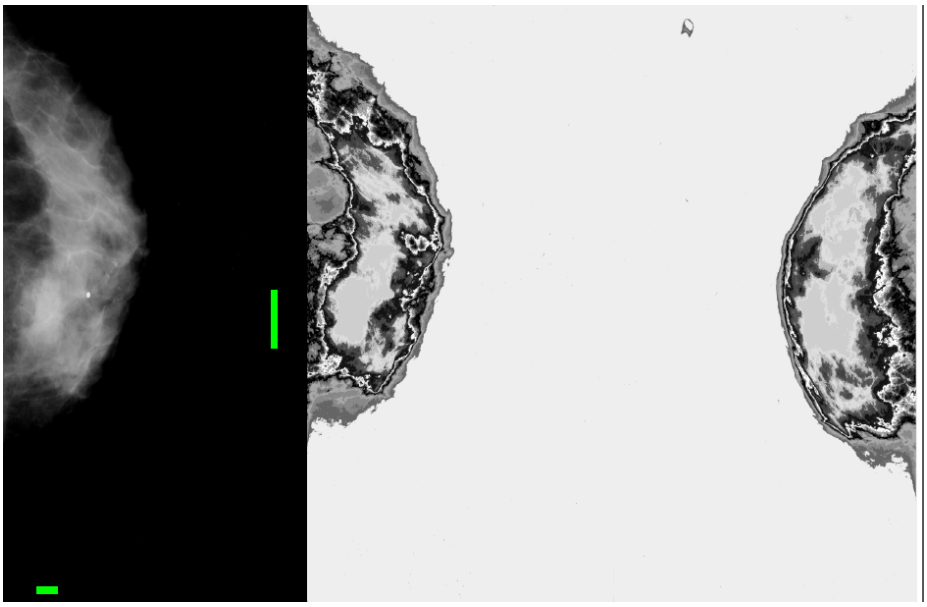

Figure 3.3: On the left is the original mammogram. The bars show the horizontal and vertical extent of a cancer location. The middle panel shows the texton map of the original mammogram (left panel) obtained by replacing each pixel by the texton label. The result shown is for a final texton dictionary of size 16 learnt from aggregated cancer and non-cancer ROIs. The right panel shows the texton map of the contralateral non-cancer breast mammogram.

3.1.4 Discussion and Conclusion

According to Table 3.1 and Figure 3.2 in Section 3.1.3, it seems that the algorithm distinguishes tissues associated with cancer and non-cancer. However, the pair of texton maps (Figure 3.3) indicate that the texture features identified with high intensity regions are not specific to cancer/absolutely high risk. The classification AUC scores in Table 3.1 may appear satisfactory just because the selected cancer ROIs are mostly associated with high density while non-cancer ROIs are mostly associated with low density (Figure 3.1). This means that density might be playing the major role in classifying cancer and non-cancer tissue and so the role of texture alone is unclear. The remaining two sections (Sections 3.2 and 3.3) of this chapter describe work aimed at separating the contributions of texture and density in classifying ROIs as cancer or non-cancer.

3.2 Local Mean and Variance Normalization

One method for testing if texture features can provide information about the presence of breast cancer independent of background intensity (density) is to compute texture

features on "flattened" images. An example of a flattened image is D_r defined by

$$D_r(p) = X(p) - \text{mean}(X(B(p, r))), \tag{3.1}$$

where X is the original image and $B(p,r)$ is the disk of radius r centered at pixel p. However, due to the nonlinearity of the imaging process, the local variation is also a function of background intensity (Figure 3.4). Thus texture measures extracted from D_r will still reflect local background intensity (related to density) as seen in Section 3.1.

In order to remove dependence on local variation, the normalized image N_r is defined by

$$N_r(p) = \frac{D_r(p)}{S_r(p)}, \qquad \text{where } S_r(p) = \text{std}(X(B(p, r))). \tag{3.2}$$

To explore textures based on these normalized images, N_r was computed using radii $r = 3n + 1$ for $n = 0, 1, \ldots, 7$ (pixels) on full resolution ($\approx 50\mu$m per pixel) mammograms. No structure could be seen for low values of r, but significant linear structures appeared for larger values of r (Figure 3.5).

An experiment was conducted to discover if the linear structures found in the normalized images N_r (textures independent of density) could be used as features to distinguish non-cancer and cancer tissues (Section 3.3).

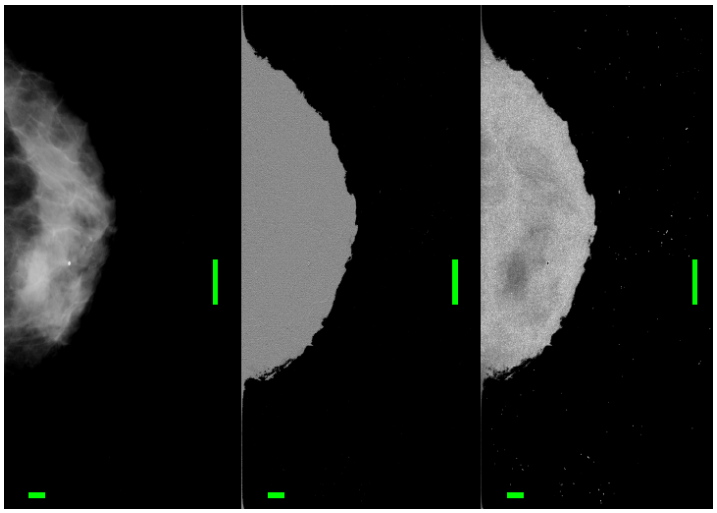

Figure 3.4: An example of a flattened image. On the left is the original mammogram X shown in Figure 3.1 (MLO view) and Figure 3.3 (CC view). The bars show the horizontal and vertical extent of a cancer location. The middle panel is the local mean subtracted image D_r (Equation 3.1). In this panel, the background has been set to the minimum value of the image to facilitate the display. The right panel is the local standard deviation image S_r. Due to the nonlinearity of the imaging process, the brightest region in X appear as a relatively dark region in S_r. For this example, $r = 5$ pixels.

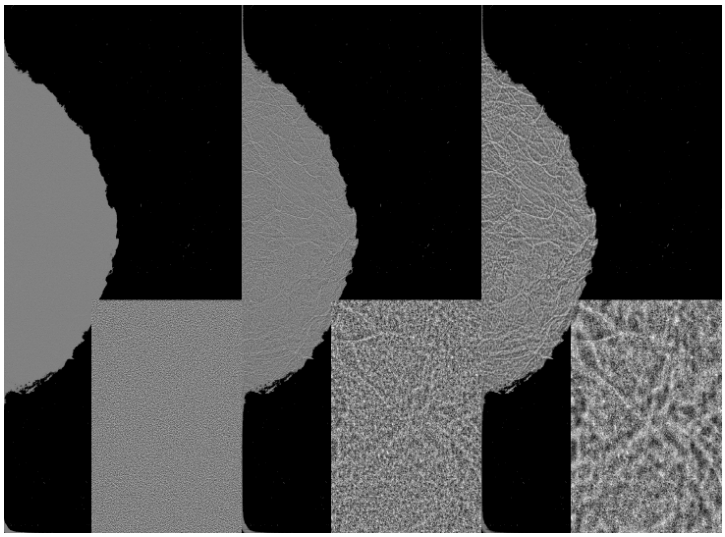

Figure 3.5: Mammograms normalized using N_r (Equation 3.2). Each panel shows the normalized image N_r obtained from the image X in Figure 3.4 for values of $r =$ 1, 10, 22 respectively (left to right). The insets in the lower right of each panel show the region of the known malignant mass (cancer ROI) indicated by the bars in Figure 3.4. The left panel shows essentially no structure for $r = 1$, but structure emerges with increasing r. In each panel, the background has been set to the minimum image value to facilitate display.

3.3 Application of the Local Normalization to Classify ROIs as Cancer or Non-cancer

In order to determine if texture alone retains any ability to distinguish tissues associated with cancer from non-cancer breast tissues, the experiment described in Section 3.1 was repeated but with the application of local normalization described in Section 3.2 as an image preprocessing step prior to the step of applying MR8 filter bank. The same protocol was used to construct the feature space, generate textons, determine texton maps and construct normalized texton histograms for each ROI. Classification performance was evaluated using the Fisher classifier in terms of AUC scores.

AUC scores for cancer and non-cancer ROIs classification are shown in Table 3.2 and Figure 3.6.

Classification scores based on texture features independent of density in this section (Table 3.2) were generally lower than classification scores based on texture features associated with density in Section 3.1 (Table 3.1). However, this classification performance was still quite satisfactory - close to 0.80 for 40 textons. By applying a paired t-test on the testing AUC scores for classification with prior normalization and AUC scores for classification without prior normalization, there is no significant difference between them (mean of difference $= -0.053$, $p = 0.06$,

Table 3.2: AUC scores for the application of local normalization to classify cancer and non-cancer ROIs.

texton dictionary size	training AUC scores	testing AUC scores
8 textons	0.757	0.697
16 textons	0.788	0.714
20 textons	0.856	0.768
25 textons	0.833	0.659
30 textons	0.869	0.729
34 textons	0.851	0.747
40 textons	0.856	0.796

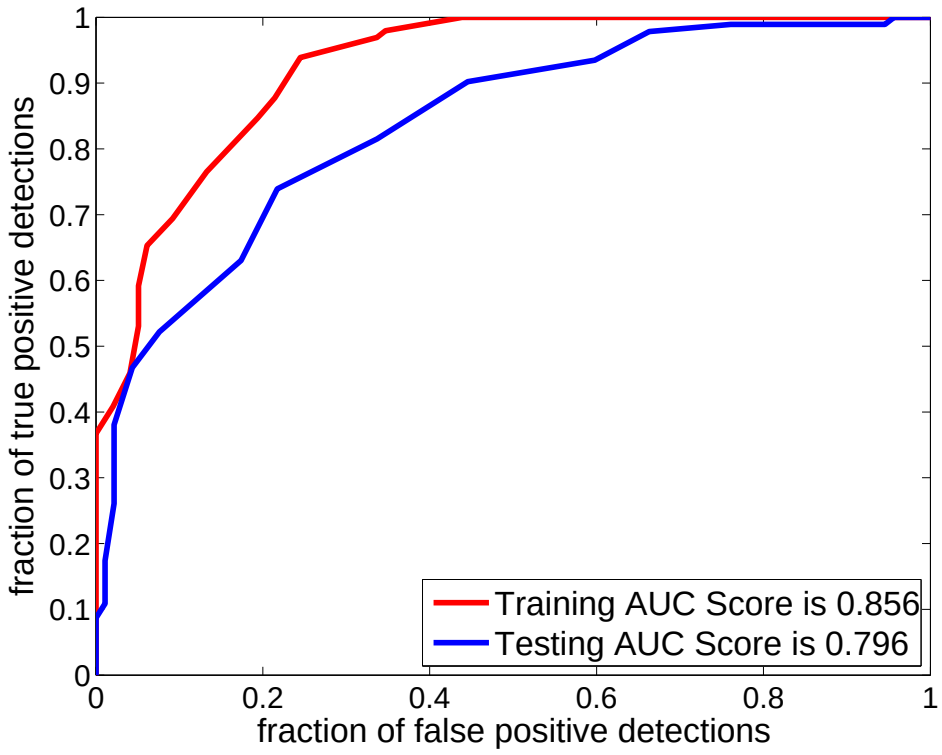

Figure 3.6: ROC curves for the application of local normalization to classify cancer and non-cancer ROIs with 40 textons.

$n = 7$). This indicates that the role of texture is significant in classifying cancer and non-cancer tissue. In addition, this indicates the application of textons for texture analysis to cancer risk assessment is feasible. Therefore, it is promising to apply a texton related method to extract texture features independent of density for breast cancer risk assessment in future studies.

Chapter 4

Variations of Texton Implementation for Risk Assessment

The central theme of this thesis is to develop texture analysis methods, and in particular, texton based methods, to estimate risk of breast cancer. The general idea of textons described in Section 2.2 of Chapter 2, requires many choices for implementation. In this chapter, several key variations of textons are considered in order to determine the most promising implementation. Section 4.1 describes three image data subsets used for conducting experiments in this chapter. In Section 4.2, the significance of texture alone realized by applying the local normalization (Section 3.2) in estimating risk is established. This section also compares performance based on CC, MLO and the combination of both view images. In Section 4.3, various widely applied candidate methods for generating textons are considered and two methods for clustering - the key step in extracting textons from the feature space - are compared in Section 4.4.

The main algorithm for estimating breast cancer risk used in the remainder of this thesis derived from studies in this chapter. In particular, it was found that: textures independent of density achieved surprisingly good risk assessment performance compared with textures dependent of density; textons based on local intensities (in $N \times N$ neighborhoods) outperform other methods for generating textons; and clustering based on K-means is somewhat better than fuzzy C-means clustering (Section 2.4.2). In addition, better estimates of risk were obtained by applying these methods to CC view images than either MLO ones or the combination of both (Section 4.2).

The work reported in Chapter 3 indicates that good results are obtained if textons are computed subsequent to applying the normalization step and they are not statistically different from those without normalization. But the experiments in Chapter 3 are, strictly speaking, not studies in risk assessment. Hence further verification of the merit of the normalization step was sought and so the algorithm without normal-

ization was included in the experiments reported in this chapter as another variation of implementing textons.

Although the major objective of the thesis is to study risk assessment based on images in years prior to the diagnosis of breast cancer (Chapter 8), longitudinal data were not available until late in the study. Accordingly, the preliminary experiments reported here were based on classifying mammograms according to BI-RADS classes (Section 1.3 of Chapter 1) identified by expert radiologists. Thus, in this chapter, BI-RADS classes act as surrogates for breast cancer risk (the first risk criteria in Table 2.1).

4.1 Data Set

Three image data subsets were selected from the DDSM data set (Section 1.5) for conducting the experiments in this chapter; a CC data set, a MLO data set and a combined CC and MLO data set. Details of these data sets are provided below. Equal numbers of images were taken from each of the four BI-RADS classes for each of these three data sets.

Although in this chapter images are classified into BI-RADS classes, the four classes BI-RADS I - BI-RADS IV may be viewed as representing four classes of increasing risk. In addition, density classes BI-RADS I and BI-RADS II can be taken as a low density class while density classes BI-RADS III and BI-RADS IV are categorized as high density class. Example CC view BI-RADS images are shown in Figure 4.1.

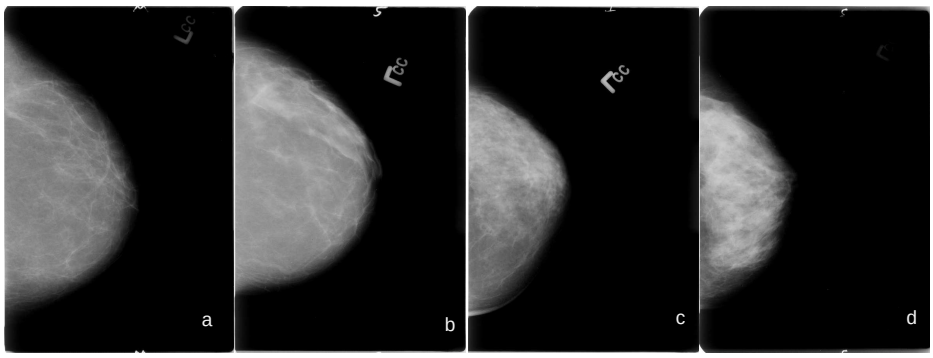

Figure 4.1: Examples of CC view mammogram images from four BI-RADS density classes; (a) BI-RADS I, (b) BI-RADS II, (c) BI-RADS III, (d) BI-RADS IV.

The CC data set comprises 40 right CC view images in each BI-RADS class (160 images from 160 different women). Of these 20 in each class were used for training and the remaining 20 for testing. This data set was used for conducting the experiment on texture alone with the application of normalization (Section 4.2),

comparing three methods for texton generation (Section 4.3) and comparing two clustering methods (Section 4.4).

The MLO data set comprises 40 right MLO view images in each BI-RADS class (160 images from 160 different women). Of these 20 in each class were used for training and the remaining 20 for testing. This data set was used for conducting the experiment on texture alone with the application of normalization (Section 4.2).

The Combined CC and MLO data set comprises both the CC and MLO views of the right breasts from 80 women (160 images) equally divided between the four BI-RADS classes. Within each class, 10 CC images and 10 MLO images were used for training and the remaining 10 CC and 10 MLO images were used for testing. This data set was used for conducting the experiment on texture alone with the application of normalization (Section 4.2). In particular, these three image data subsets are used to study the significance of textures from different view mammogram images to risk assessment (Section 4.2).

4.2 Application of the Local Normalization to BI-RADS Classification

The normalization step introduced in Chapter 3 has the additional effect of increasing the noise in the non-breast region relative to the tissue related intensity variation within the breast region and also produces a band of anomalous texture at the boundary of the breast (Figure 4.2 (c)). To avoid training textons to recognize this region instead of texture difference between high and low risk tissue, an additional preprocessing step was included to remove the band of anomalous texture (Figure 4.2 (d)). This was done by eroding the templates described in Section 1.5 (Figure 4.2 (a)) by a circular structure element of 50 pixels ($\approx$ 2.5 mm). In order to compare all methods fairly, this extra preprocessing step was adopted for all studies reported in this chapter and all studies reported in the thesis from this point on. These final templates (Figure 4.2 (b)) were saved for use throughout.

In this chapter, BI-RADS class classification with and without normalization is compared with the study by Petroudi et al. (Gong and Petroudi [2006]). The study by Petroudi et al. used the Oxford mammography data set which is not available generally. In order to do the comparison, the method described by Petroudi et al. was implemented on the CC data set, the MLO data set and the combined CC and MLO data set. For ease of exposition, these three algorithms will be referred to as follows. The term "algorithm with normalization" will be used to refer to the method proposed in Section 4.2.1.1 with the application of local normalization (Section 3.2). The term "algorithm without normalization" will be used to refer to the

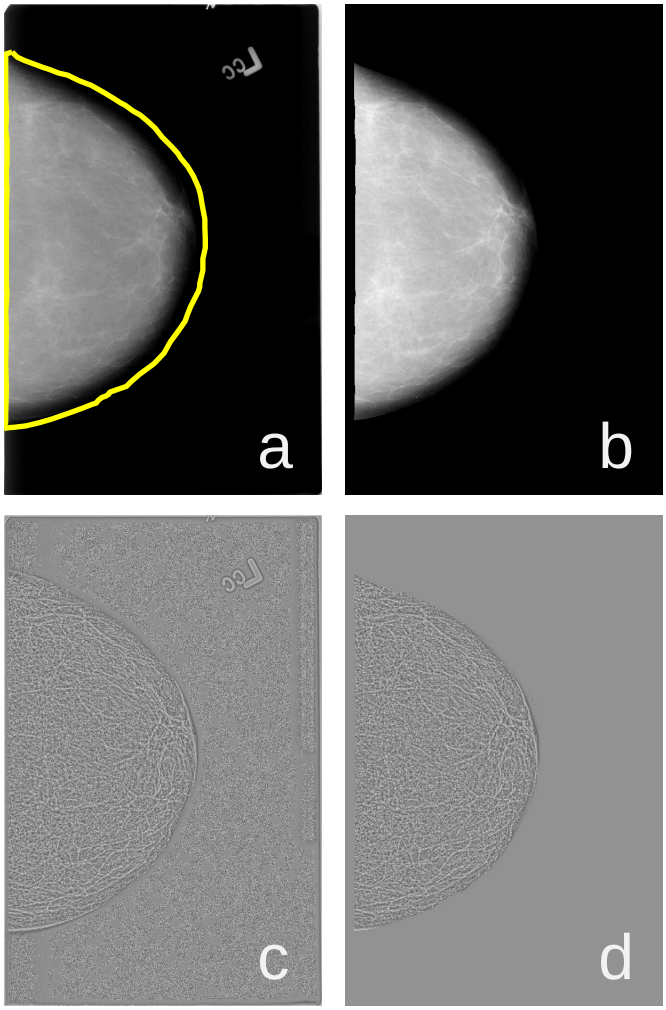

Figure 4.2: Image preprocessing steps: (a) original breast image with initial breast boundary, (b) the image in (a) after applying the final image template, (c) the image in (a) after normalization, (d) the image in (c) after applying the final image template. The apparent increase in brightness of the breast in (b) is a display artifact. The brightest pixels in (a) comprise anomalies near the edge of the image outside the breast and within the LCC label. These are removed in applying the final template and the intensities within the breast region are rescaled to cover the full range of display values.

algorithm without the application of local normalization (Section 4.2.1.1). The term "Petroudi's algorithm" will be used to refer to the algorithm in the study of Petroudi et al. (Section 4.2.1.2).

4.2.1 Experimental Details of Three Algorithms

In this section, experimental details of the three algorithms described above: algorithm with normalization, algorithm without normalization, and Petroudi's algorithm in BI-RADS classification are introduced.

4.2.1.1 Algorithm with and without normalization

The algorithm with normalization requires setting three parameters: (1) the value of the radius r used in the local normalization step, (2) the value of N that specifies the size of the local $N \times N$ neighborhood for computing texture features, (3) the value of K in K-means clustering (equivalently, the number of textons in the texton dictionary). Increasing r results in poorer resolution of local structure (Chapter 3). The value $r = 22$ was determined empirically (Chapter 3) to provide a balance between retaining texture information and sufficient local focus. The value $N = 3$ was chosen for the local $N \times N$ neighborhood to match previous work reported in the literature (Gong and Petroudi [2006]). For K-means clustering, K was empirically set to $K = 5$ for each BI-RADS class, resulting in 20 textons (clusters) in total. The value of K and the method for applying K-means clustering was also chosen to match the work by Petroudi et al. (Gong and Petroudi [2006]).

First, every image was normalized (Section 3.2) and then each pixel in the image of each data subset was replaced by a vector of length eight comprising the eight normalized intensity values in the 3×3 neighborhood of the pixel (but omitting the central pixel p). The breast region was reduced to avoid texture distortion as described in Section 4.2. The array was sub-sampled by $8 \times 8 \rightarrow 1$ so that every patch breast tissue of 64 pixels in the original image was represented by a single vector of length 8. Thus, texture primitives were computed at full resolution, but represented at the subsampled rate to manage the computational and storage loads. Next, the local mean of the 3×3 image patch from the normalized image was included as the 9th component of the vector representing the patch.

The second step was to generate the texton dictionary. K-means clustering with $K = 5$ was applied to the collection of vectors of length 9 representing the image patches from the training images within a BI-RADS class. With 5 clusters generated per class, a texton dictionary of size 20 was obtained.

In the third step, every image patch represented by a vector of length 9 was assigned a label according to the closest cluster center (texton) in the feature space.

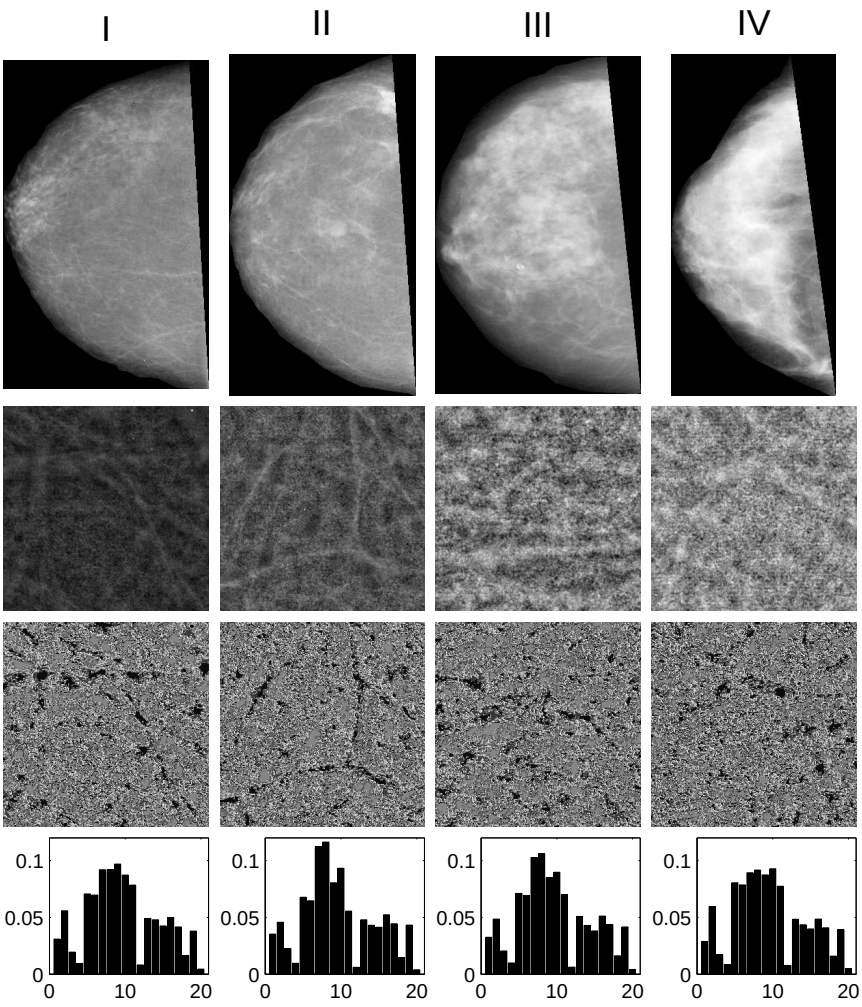

Figure 4.3: Examples of CC view BI-RADS images (the first row), normalized image patches (the second row), detailed texture features in texton map patches (the third row) and texton histograms (the fourth row) of the algorithm with normalization. Each column corresponds to one of the BI-RADS pattern classes (I - IV from left to right). Patches in the second and third rows were chosen from the same positions in the original BI-RADS images.

Pixels outside the breast and on the boundary of the breast were assigned value 0. This resulted in a texton map for each image (Figures 4.3 and 4.4). Each image was subsequently represented by the normalized histogram of textons comprising the associated texton map excluding texton 0.

Finally, ensemble k-nearest neighbor classifiers (Section 2.5.1) were used to classify the images into the four BI-RADS pattern classes. Because the classification is for more than 2 groups (4 density classes) , the rule for breaking a tie in each k-nearest neighbor classifier learner is "nearest". That is, if two or more classes are tied in having the greatest numbers of neighbors of a particular point P, then the class assigned to P is the one having the closest neighbor to P. The method of searching the nearest neighbors is exhaustive since there are 20 different texture features. The

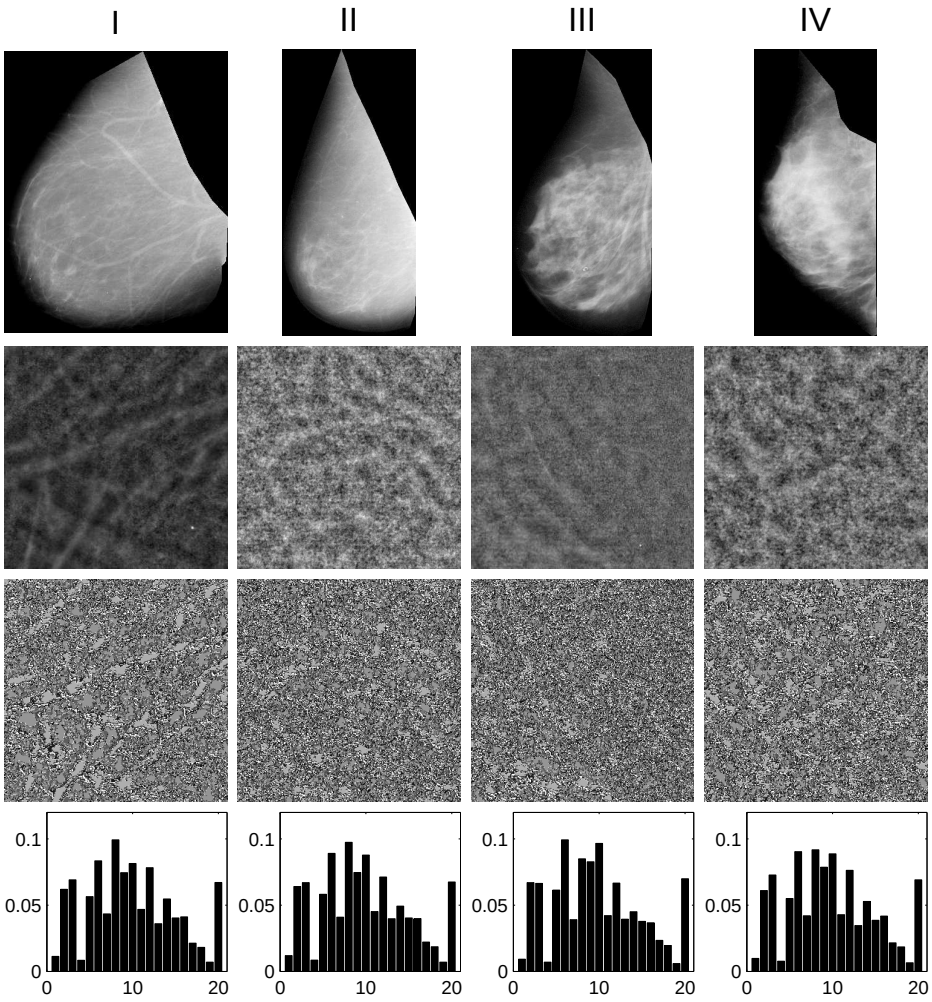

Figure 4.4: Examples of MLO view BI-RADS images (the first row), normalized image patches (the second row), detailed texture features in texton map patches (the third row) and texton histograms (the fourth row) of the algorithm with normalization. Each column corresponds to one of the BI-RADS pattern classes (I - IV from left to right). Patches in the second and third rows were chosen from the same positions in the original BI-RADS images.

three parameters introduced in Section 2.5.1 for the ensemble k-nearest neighbor classifier were set to be $k = 2$, $m = 14$, $n = 50$, then $k = 2$, $m = 11$, $n = 100$ and $k = 7$, $m = 16$, $n = 163$ for the CC data set, the MLO data set and the combined CC and MLO data set, respectively. The rule for setting these three parameters is given in Section 2.5.1 of Chapter 2.

In addition to four BI-RADS class classification described above, a separate two-class classification was conducted. Images from BI-RADS I and BI-RADS II pattern classes were combined to form a single low risk group and images from BI-RADS III and BI-RADS IV pattern classes were combined to form a single high risk group. Note that this process is not the same as sub-sampling the confusion matrix resulting from four BI-RADS classification.

The whole process of the algorithm with normalization was repeated on the CC data set, the MLO data set and the combined CC and MLO data set to explore the influence of different views in breast cancer risk assessment.

The whole process of the algorithm without normalization is similar to the algorithm with normalization except that the normalization step was omitted. The feature vector associated with each pixel in the image of each data set was of length nine comprising the eight raw pixel intensity values in the 3×3 neighborhood instead of normalized pixel intensity values and the mean of the 3×3 image patch of the original image.

4.2.1.2 Petroudi's algorithm

Petroudi's algorithm works similarly to the algorithm with normalization. The only difference lies in the first step, a low-pass filter was applied thus removing local background but not necessarily local variation. Images were subsampled by a factor of 25 to 1 (5×5 patch to 1) in order to match the spatial resolution of the images used in (Gong and Petroudi [2006]). All the remaining steps were the same as in the two algorithms above.

4.2.2 BI-RADS Classification Results for Three Algorithms

For the CC data set, the classification performance and confusion matrix for testing images for the algorithm with normalization are shown in Table 4.1 (a). The classification performance and confusion matrix for testing images for Petroudi's algorithm are shown in Table 4.1 (b). The classification performance and confusion matrix for testing images for the algorithm without normalization are shown in Table 4.1 (c).

For the MLO data set, the classification performance and confusion matrix for testing images for the algorithm with normalization are shown in Table 4.2 (a). The classification performance and confusion matrix for testing images for Petroudi's

Table 4.1: Classification performance tables and confusion matrices for CC view testing mammograms; (a) the algorithm with normalization, (b) Petroudi's algorithm, (c) the algorithm without normalization.

(a)

BI-RADS category	I	II	III	IV
accuracy	0.700	0.20	0.95	0.70
	0.675		0.975	

	I	II	III	IV
I	14	6	0	0
II	2	4	0	0
III	2	8	19	6
IV	2	2	1	14

(b)

BI-RADS category	I	II	III	IV
accuracy	0.70	0.15	0.55	0.55
	0.80		0.85	

	I	II	III	IV
I	14	12	0	0
II	5	3	4	2
III	0	5	11	7
IV	1	0	5	11

(c)

BI-RADS category	I	II	III	IV
accuracy	0.70	0.15	0.35	0.60
	0.775		0.775	

	I	II	III	IV
I	14	10	0	0
II	3	3	6	5
III	2	7	7	3
IV	1	0	7	12

algorithm are shown in Table 4.2 (b). The classification performance and confusion matrix for testing images for the algorithm without normalization are shown in Table 4.2 (c).

For the combined CC and MLO data set, the classification performance and confusion matrix for testing images for the algorithm with normalization are shown in Table 4.3 (a). The classification performance and confusion matrix for testing images for Petroudi's algorithm are shown in Table 4.3 (b). The classification performance and confusion matrix for testing images for the algorithm without normalization are shown in Table 4.3 (c).

4.2.3 Discussion and Conclusion of Three Algorithms

Comparing the classification performance tables and confusion matrices in Table 4.1 (a), (b), (c) with Table 4.2 (a), (b), (c) indicates that the performance of the algorithm with normalization is comparable to Petroudi's algorithm. For CC view images (Table 4.1 (a) and (b)), the algorithm with normalization performs substantially better in all four risk groups and performs particularly well in assigning images correctly to the high risk groups. For MLO view images (Table 4.2 (a) and (b), the same conclusion was obtained. When it comes to the combined CC and MLO view images (Table 4.3 (a) and (c)), the algorithm with normalization outperforms Petroudi's

Table 4.2: Classification performance tables and confusion matrices for MLO view testing mammograms; (a) the algorithm with normalization, (b) Petroudi's algorithm, (c) the algorithm without normalization.

(a)

BI-RADS category	I	II	III	IV
accuracy	0.70	0.05	0.75	0.70
	0.775		0.925	

	I	II	III	IV
I	14	13	1	0
II	4	1	1	0
III	1	5	15	6
IV	1	1	3	14

(b)

BI-RADS category	I	II	III	IV
accuracy	0.65	0.05	0.25	0.75
	0.825		0.850	

	I	II	III	IV
I	13	13	2	0
II	5	1	3	0
III	1	5	5	5
IV	1	1	10	15

(c)

BI-RADS category	I	II	III	IV
accuracy	0.75	0.00	0.50	0.55
	0.825		0.800	

	I	II	III	IV
I	15	13	1	2
II	3	0	4	1
III	0	5	10	6
IV	2	2	5	11

Table 4.3: Classification performance tables and confusion matrices for combined CC and MLO view testing mammograms; (a) the algorithm with normalization, (b) Petroudi's algorithm, (c) the algorithm without normalization.

(a)

BI-RADS category	I	II	III	IV
accuracy	0.80	0.00	0.70	0.35
	0.800		0.875	

	I	II	III	IV
I	16	12	2	2
II	3	0	0	0
III	0	6	14	11
IV	1	2	4	7

(b)

BI-RADS category	I	II	III	IV
accuracy	0.60	0.00	0.35	0.35
	0.70		0.75	

	I	II	III	IV
I	12	16	0	0
II	0	0	5	4
III	6	4	7	9
IV	2	0	8	7

(c)

BI-RADS category	I	II	III	IV
accuracy	0.45	0.00	0.35	0.35
	0.700		0.675	

	I	II	III	IV
I	9	17	0	2
II	4	0	7	6
III	6	3	7	5
IV	1	0	6	7

algorithm as well.

Comparing Table 4.1 (a) and (c), Table 4.2 (a) and (c), and Table 4.3 (a) and (c) indicates that normalizing the images before measuring texture features does improve the classification performance. For CC view images (Table 4.1 (a) and (c)), the algorithm with normalization performs substantially better than the algorithm without normalization in every individual class and performs substantially better in the overall categories of high density. For MLO view images (Table 4.2 (a) and (c)), normalization seems to have little effect on overall performance but is noticeably worse in the highest risk class, BI-RADS IV, and noticeably better in the class BI-RADS II. For combined CC and MLO view images (Table 4.3 (a) and (c)), normalization provided much better performance in each density class.

These results establish that texture features alone retain information relevant to assess the risk of breast cancer independently of image density. In fact, since the performance of the algorithm with normalization tended to be better than the algorithm without normalization and Petroudi's algorithm, texture may be more important than background image intensity in BI-RADS class classification. This is perhaps not surprising since image intensity depends not only on breast density and breast size (both of which are correlated to breast cancer risk) but also on image acquisition parameters.

A striking result is that the algorithm with normalization performs better on CC view images than MLO view images (Table 4.1 (a) and Table 4.2 (a)). In addition, the performance of the algorithm with normalization is better on MLO view images than the combined CC and MLO view images (Table 4.2 (a) and Table 4.3 (a)), especially in the fourth density class. This trend is seen in Petroudi's algorithm (Table 4.1 (b), Table 4.2 (b) and Table 4.3 (b)), but the trend is not as strong as the one of the algorithm with normalization. The reason for this is not clear, nor is it clear if this difference is due to sampling only or if it reflects an inherent difference between CC and MLO view images. The latter is plausible. The natural structure of the breast does seem to vary more in the vertical direction than lateral direction although this has not been quantified to our knowledge. If this is true, then capturing the natural variation plus anomalies might require more textons for MLO images than CC images and suggests that optimal parameters for these views may not be the same. The notion of developing different algorithms for CC and MLO views may be important.

4.3 Comparison of Candidate Methods for Texton Generation

Three methods for generating textons were compared: MR8 filtering, $N \times N$ neighborhoods and Gabor filtering (Section 2.3). In these experiments, only CC view images were used since they provided better results in the previous section. Since texture independent of density was shown to play an important role in mammogram image classification in Section 4.2 above, all the following experiments on breast cancer risk assessment will focus on texture analysis independent of density as realized by local normalization presented in Section 3.2.

4.3.1 Experimental Details of Three Candidate Methods

In this section, experimental details for applying three candidate methods: MR8 filtering, $N \times N$ neighborhoods and Gabor filtering for texton generation are described.

4.3.1.1 MR8 filtering

The experimental details for the method of MR8 filtering were the same as the process described above in Section 4.2.1 for the algorithm with normalization using the CC data set except that the MR8 filter bank (Section 2.3.1) was applied to the normalized images to construct 8-dimensional feature vectors for texton generation. The final BI-RADS classification results for testing images for the method of MR8 filtering are shown in Table 4.4 (a).

4.3.1.2 $N \times N$ neighborhoods

Experimental details for the method of $N \times N$ neighborhoods are exactly the same as the process described above in Section 4.2.1 for the algorithm with normalization using the CC data set. The results for the method of $N \times N$ neighborhoods are shown in Table 4.4 (b). They are not the same as the results in Table 4.1 (a), because the three parameter settings in these two subspace ensemble k-nearest neighbor classifiers are not the same (Section 4.3.2).

4.3.1.3 Gabor filtering

Three versions of using Gabor filters (Section 2.3.3) are presented here: Gabor filter textons, Gabor oriented features and Gabor oriented textons. As the name suggest, the first and third methods generate textons. The second method generates features related to the orientation of tissue texture but does not follow the protocol for textons in that there is no clustering step.

Table 4.4: Classification performance tables and confusion matrices for the candidate methods for the generation of textons: (a) MR8 filtering, (b) $N \times N$ neighborhood method, (c) Gabor filter texton method, (d) Gabor oriented feature method, (e) Gabor oriented texton method.

(a)

BI-RADS category	I	II	III	IV
accuracy	0.35	0.1	0.7	0.4
	0.65		0.675	

	I	II	III	IV
I	7	9	3	6
II	0	2	2	0
III	2	6	14	6
IV	11	3	1	8

(b)

BI-RADS category	I	II	III	IV
accuracy	0.7	0.1	0.95	0.6
	0.625		1	

	I	II	III	IV
I	14	6	0	0
II	3	2	0	0
III	2	12	19	8
IV	1	0	1	12

(c)

BI-RADS category	I	II	III	IV
accuracy	0.55	0.15	0.75	0.8
	0.625		0.825	

	I	II	III	IV
I	11	7	2	1
II	3	3	2	0
III	3	8	15	3
IV	3	2	1	16

(d)

BI-RADS category	I	II	III	IV
accuracy	0.3	0.15	0.4	0.4
	0.55		0.675	

	I	II	III	IV
I	6	5	5	2
II	4	3	3	2
III	6	9	8	8
IV	4	3	4	8

(e)

BI-RADS category	I	II	III	IV
accuracy	0.65	0	0.55	0.7
	0.75		0.725	

	I	II	III	IV
I	13	10	5	2
II	4	0	3	0
III	2	7	11	4
IV	1	3	1	14

In the Gabor filter texton method, Gabor filters (Section 2.3.3) were used as the common filters and so the whole process of conducting the experiment was the same as the process described above in Section 4.2.1 for the algorithm with normalization. Gabor filters at 10 orientations were used on the CC data set to construct 10-dimensional feature vectors. The final results for the Gabor filter texton method are shown in Table 4.4 (c).

In the Gabor oriented feature method, after the filter direction map was obtained with the application of the threshold $T = 0.2$ and the image template, texture features were calculated from the filter direction map as the frequencies of maximum filter response directions omitting the direction labeled with 0. Classification results for the Gabor oriented feature method are shown in Table 4.4 (d).

In the Gabor oriented texton method, after the filter direction map was obtained, the 3×3 neighborhood method was applied on the filter direction map to construct 9-dimensional feature vectors (including the central filter direction). Similar to the algorithm with normalization using CC data set for texton generation, a total of 20 textons were generated from four BI-RADS classes. For each image, the frequency of textons generated in this way was taken as the final texture feature excluding the out of breast region texton label 0. Classification results for testing images for the Gabor oriented texton method are shown in Table 4.4 (e).

4.3.2 Discussion and Conclusion of Three Candidate Methods

In the classification step using an ensemble k-nearest neighbor classifier, in order to make comparison, the same parameters were used for three methods described above. These parameters were different from those used in the algorithm with normalization applied to the CC data set. Because the Gabor oriented feature method had 11 texture features, unlike the remaining methods of texton generation which had 20 texture features. In this study, in order to compare the three candidate methods of texton generation, the three parameters introduced in Section 2.5.1 of the ensemble k-nearest neighbor classifier were set to be k =2, $m = 4$, $n = 20$. The limitation here was that the parameter m (the number of features) could not be over 11.

According to the results above (Table 4.4), the $N \times N$ neighborhood method seems to be the best for texture feature calculation with textons. Gabor filtering is in second place but the performance depends on how Gabor filtering is applied. In comparison with the two methods above, MR8 filtering does not perform as well. Similar trends were reported in the literature. Varma and Zisserman found that performance obtained from $N \times N$ neighborhoods was as good or better than standard textons based on filter banks in texture image classification with Columbia-Utrecht database (Varma and Zisserman [2003]). Petroudi et al. applied 3×3 neighborhood method to classify mammogram images and, like Varma and Zisserman, found that performance matched that of standard filter banks (Gong and Petroudi [2006]).

In summary, for further study of risk assessment in this thesis, $N \times N$ neighborhoods and Gabor filtering will be considered. But the way of applying Gabor filtering needs to be explored.

4.4 Comparison of Two Clustering Methods

From the conclusions obtained above (Sections 4.2.3 and 4.3.2), CC view mammograms seem better suited for risk assessment of BI-RADS class classification and

3×3 neighborhood method is the best texton related method for texture feature calculation from the normalized images. Thus the data set of CC view images and 3×3 neighborhood method will be used to conduct the experiments of comparing K-means and fuzzy C-means clustering.

The experiment using 3×3 neighborhood method, K-means clustering and the CC data set was conducted in Section 4.2.1 already and the results obtained were shown in Table 4.1 (a). The whole process of this experiment was repeated by using fuzzy C-means (Section 2.4.2) instead of K-means. In order to compare these two clustering methods, the same number of clusters (overlap parameter) used in K-means experiment were generated from the training images within a BI-RADS class using fuzzy C-means clustering. The remaining parameters were optimized for fuzzy C-means clustering and were not the same as for K-means clustering. The three parameters in the ensemble k-nearest neighbor classifier were the same in the two experiments. Thus, in the last classification step, an ensemble k-nearest neighbor classifier was used to show the risk classification performance for testing images for each clustering method with the same parameter settings as in Section 4.2.1. The results for BI-RADS class classification with fuzzy C-means clustering are shown in Table 4.5.

Comparing the results in Table 4.1 (a) of K-means clustering with the results in Table 4.5 of fuzzy C-means clustering, generally K-means clustering works better than fuzzy C-means clustering in BI-RADS classification for risk prediction. In addition, K-means clustering has been used more commonly in the literature than fuzzy C-means clustering in texton generation in texture image classification. Thus only K-means clustering will be considered in the following study on texton related risk assessment.

Table 4.5: BI-RADS classification performance table and confusion matrix for testing images using fuzzy C-means clustering instead of K-means clustering (Table 4.1 (a)).

BI-RADS category	I	II	III	IV
accuracy	0.50	0.25	0.70	0.65
	0.65		0.95	

	I	II	III	IV
I	10	4	0	0
II	6	5	1	1
III	3	9	14	6
IV	1	2	5	13

Chapter 5

Texture and Region Dependent Risk Assessment

A key question in computer-aided risk assessment is whether patterns relevant to breast cancer risk are concentrated in a particular region or spread throughout the breast. Huo et al. (Huo et al. [2000, 2002]) consistently selected a ROI of 256×256 pixels from the central breast region behind the nipple, regardless of the breast size, to classify images into high or low risk groups. Texture features were extracted from local gray-level variation analysis and an average AUC score of 0.91 was obtained in classifying BRCA1/BRCA2 mutation carriers and non carriers. They also found that high risk images tended to be dense and mammographic patterns appeared as a coarse low contrast texture. Choosing the central region behind the nipple is reasonable since this region is usually the densest part of the breast and density is a significant indicator of breast cancer risk (Huo et al. [2000], Li et al. [2008]). In 2004, the same group, studied the effect of ROI size and location on breast cancer risk again using the BRCA1/BRCA2 mutations to assign high and low risk groups (Li et al. [2004]). Five ROIs were selected manually from left CC view images: (A) the central breast region immediately behind the nipple, (B) the upper central breast region, (C) the lower central breast region, (D) the center of the central breast regions, and (E) the central left breast region. Their results showed that the size of the ROI was not important but there was a statistically significant decrease in classification performance as the ROI location varied from the central region behind the nipple (A) to other locations (B, C, D, and E). In 2008, they applied power law spectral analysis to mammograms to distinguish BRCA1 and BRCA2 mutations carriers from non carriers (Li et al. [2008]). Their power spectral analysis was based on the power spectrum obtained from discrete Fourier transforms. The central region (A) was found to provide the best performance. They achieved an AUC score of 0.9 in differentiating 30 BRCA1/BRCA2 gene mutation carriers (high risk) from 60 age-matched non carriers (low risk).

The main mammographic indicator of breast cancer is the amount and distribution of the dense tissue. In addition to the density (or its surrogate, intensity), texture is thought to provide information relevant to risk assessment (Wolfe [1976b]). Several studies have appeared on the use of texture for classifying risk (Petroudi et al. [2003], Gong and Petroudi [2006], Petroudi and Brady [2011]). However, in these studies, texture and density were considered together (not considered separately). Although breast density is usually most pronounced in the region just behind the nipple, whether this holds for texture is not known. Accordingly, in this chapter, several regions of the breast as well as the full breast were examined separately.

The two best candidate methods for texton generation found in Chapter 4; $N \times N$ neighborhoods and Gabor filtering, were tested on all regions. The former is a texton feature and the latter ties texture features directly to biological structure and therefore has the potential to deliver a causal result rather than just an observational one.

The region just behind the nipple is found to be the most significant local region for estimating risk, but estimates based on the entire breast perform better. Texton features are found to perform better than features based on oriented tissue structures.

5.1 Data Set

In this chapter, high risk mammograms were taken to be images of the unaffected breast from women identified to have cancer (benign or malignant) in the contralateral breast at screening. Low risk mammograms were randomly selected left or right breast images from women not found to have cancer at screening in either breast (the third risk criteria in Table 2.1). Images were taken from DDSM database (Section 1.5 of Chapter 1). Only CC view images were selected according to the conclusion made in Section 4.2 of Chapter 4. Here, the BI-RADS scores were not used as an indicator of risk. The BI-RADS scores were used only to ensure that the data set represents a wide spectrum of mammographic appearance. To do this, 40 low risk and 40 high risk images (by the criterion described above) were taken from each of the four BI-RADS categories (320 images in total). Within each group and each BI-RADS class, half the images were randomly selected for training and the remaining half were reserved for testing.

The preprocessing steps were the same as in Chapter 4, non-breast objects were removed and erosion of the template by a circular structure element of radius 50 pixels was used to further reduce the breast region to generate the final image template for future use.

5.2 Delineating Local Regions

In order to study the contribution of different regions of the breast to risk assessment, six regions were defined. The first three were annular sections centered at the nipple. These shapes were chosen because breast cancer risk is thought to be associated with the region just behind the nipple (Huo et al. [2000, 2002], Li et al. [2004, 2008]). The annular shape was chosen since this seems to be a more natural shape with respect to the distribution of tissue in the breast than the square regions used in Li et al. [2004]. To define these regions, three landmark points were selected manually: the nipple and two points (called extreme points) on the boundary of the breast were chosen so as to maximize the area of the triangle formed by the nipple, and the extreme points that lie fully within the breast region (Figure 5.1). The two extreme points were constrained to lie on the same vertical line. Let R denote the distance between the nipple and the vertical line containing the extreme points. For $n = 1, 2, 3$, region Ω_n was taken to be the region within the breast and lying at a distance between R_n and R_{n+1} from the nipple where $R_1 = \frac{1}{10}R$, $R_2 = \frac{7}{15}R$, $R_3 = \frac{5}{6}R$ and $R_4 = \frac{6}{5}R$. These values were chosen to avoid the nipple and separate the remaining central region of the breast (away from the breast boundary) into annular regions of equal width ($\frac{11}{30}R$). The value $R_4 = \frac{6}{5}R$ was set to include a substantial part of the breast beyond the vertical line containing the extreme points, but the region within Ω_3 outside the breast was, of course, not included. Region Ω_6 was the full breast region within the template, and the remaining two regions were the full central region defined by $\Omega_5 = \Omega_1 \cup \Omega_2 \cup \Omega_3$ and the non-central or boundary region defined by $\Omega_4 = \Omega_6 \backslash \Omega_5$. Since these regions are defined in terms of breast specific landmarks, the regions scale according to the size of the breast.

5.3 Texture Features and Classification

As stated in Chapter 4, the focus of the thesis is texture analysis based on textons in breast cancer risk assessment. Thus the local normalization proposed in Section 3.2 was used to separate texture from density of images for all the following chapters. Two classes of texture features were extracted from the normalized images; texton features based on pixel intensities in local $N \times N$ neighborhoods with $N = 3$ and features based on oriented tissue structures.

5.3.1 Texton Features

The value $N = 3$ was chosen for three reasons. Firstly, texture analysis based on $N \times N$ neighborhoods with $N = 3$ is well established in the literature. Varma and

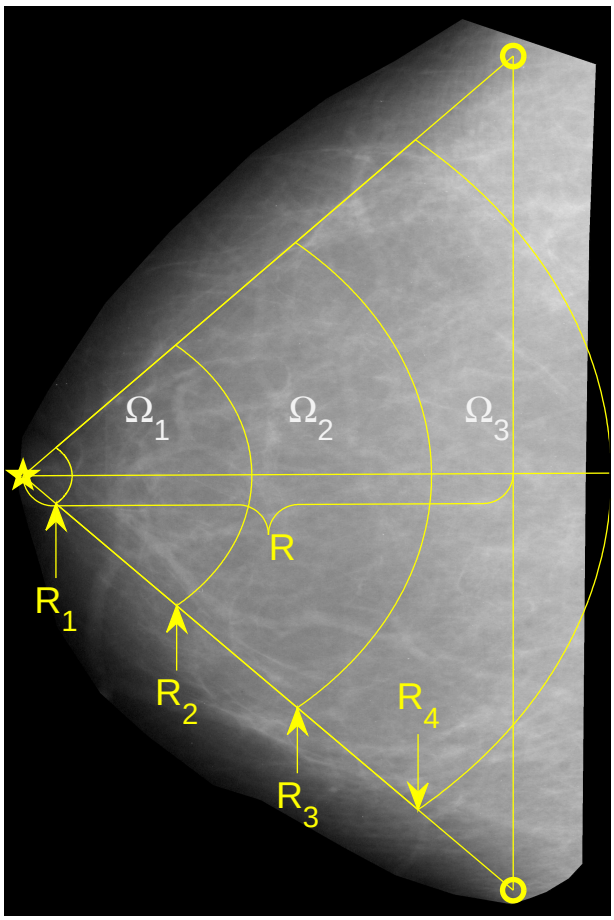

Figure 5.1: An example of delineating local regions with three landmark points; the star on the left is the nipple and the two circles on the right are the two extreme points described in the text.

Table 5.1: Risk classification performance for different size $N \times N$ local neighborhoods for six different regions of the breast from Ω_1 to Ω_6; (a) total accuracies of ensemble k-nearest neighbor classifier, (b) total accuracies of SVM classifier, (c) testing AUC scores from the Fisher classifier.

(a)

	3×3	5×5	7×7
Ω_1	0.650	0.656	0.631
Ω_2	0.551	0.531	0.550
Ω_3	0.682	0.538	0.581
Ω_4	0.613	0.563	0.588
Ω_5	0.575	0.557	0.563
Ω_6	0.694	0.663	0.650

(b)

	3×3	5×5	7×7
Ω_1	0.688	0.588	0.538
Ω_2	0.607	0.500	0.444
Ω_3	0.681	0.569	0.575
Ω_4	0.594	0.594	0.631
Ω_5	0.619	0.631	0.469
Ω_6	0.713	0.656	0.650

(c)

	3×3	5×5	7×7
Ω_1	0.702	0.597	0.567
Ω_2	0.599	0.517	0.423
Ω_3	0.694	0.523	0.560
Ω_4	0.634	0.612	0.661
Ω_5	0.601	0.625	0.425
Ω_6	0.763	0.648	0.634

Zisserman introduced textons based on $N \times N$ neighborhoods and included a comparison of performance with values of $N = 3, 5, 7, \ldots, 19$ on the Columbia-Utrecht database. They found that $N = 7$ was optimal but only slightly better than $N = 3$ (96.19 percent accuracy compared to 95.33 percent) and at the expense of a much larger computational load (Varma and Zisserman [2003]). Petroudi et al. (Gong and Petroudi [2006], Petroudi and Brady [2011]) used $N = 3$ to classify mammograms and they compared the performance based on 3×3 neighborhoods to that based on 5×5 neighborhoods. They found no significant difference between the two neighborhood sizes, but found that computation time for 5×5 neighborhoods was significantly longer. Secondly, the performance of using $N = 3, 5, 7$ was tested by the author and the best results were achieved with $N = 3$ (Tables 5.1 and 5.2 in this chapter and Table B.1 and Table B.2 in Section B.1 of Appendix B). Thirdly, by taking the image as a discretization of a differentiable surface, the first and second partial derivatives at a point suffice to classify all quadratic surfaces, for example. Since three points allow estimates of both first and second partial derivatives, a 3×3 neighborhood encompasses all the information needed to assign the best local quadratic approximation of the image at the central point.

A separate set of textons was constructed for every region Ω_n ($n \in \{1, 2, 3, 4, 5, 6\}$) from 160 training images. Let $S_{n,B}$ denote the set of 40 training images (20 low risk and 20 high risk) of BI-RADS class $B \in \{\text{I, II, III, IV}\}$ restricted to region Ω_n. For a pixel p_i in one of the member images of $S_{n,B}$, the feature vector associated with p_i was defined as $v_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,8})$, where $p_{i,1}, p_{i,2}, \ldots, p_{i,8}$ denote the image intensity values of the eight pixels sharing either a vertex or edge with p_i (pixels at the edge of an image were not included). The collection of these feature vectors from the pixels forming the member images of $S_{n,B}$ formed an 8-dimensional feature space. K-means clustering with $K = 5$ was applied to this feature space resulted in 5 clusters, identified by their centers T_j, $j = 1, 2, \ldots, 5$. These 5 centers are the textons associated with the set of images $S_{n,B}$. Repeating this process for the

four BI-RADS classes resulted in a total of 20 textons representing region Ω_n. These are the same steps for generating textons as described in Section 4.2.1.1 except that the textons are specific to each region.

Once the set of 20 textons was established for a particular region Ω_n ($n \in \{1, 2, 3, 4, 5, 6\}$), each feature vector v_i in Ω_n was associated to a texton T_m where m is such that $\|T_m - v_i\| \leq \|T_j - v_i\|$, $j = 1, 2, \ldots, 20$. Texton map images were formed by replacing every pixel p_i by the associated texton index m. Pixels that were not part of the breast region of the image (outside of the image template) were assigned a texton index 0. The histogram of texton indices of the texton map restricted to region Ω_n constituted the texture representation of the region. The histogram did not include the texton index 0.

5.3.2 Oriented Structure Features

The second class of texture features was derived from the oriented structure of breast tissue by applying Gabor filters. Filter parameters were chosen to match characteristics of the perceived tissue structure at the limit of spatial resolution (Figure 5.2). Oriented structures are derived from the filter direction map introduced in Section 2.3.3. By this criterion, the wavelength (λ) of the cosine factor of the Gabor filter kernel was set to be $\lambda = 20$, the size of Gaussian envelope (σ) was chosen to be $\sigma = 4.2$, the phase offset (φ) in the argument of the cosine factor of the Gabor function was set to be $\varphi = 0$, spatial aspect ratio (γ) specifies the ellipticity of the support of the Gabor function and its value was chosen to be $\gamma = 0.4$, and the half-response spatial frequency bandwidth (b) was set to be $b = 4$. With these parameters, 10 filter orientations at $\theta = k\pi/10$, $k = 1, 2, \ldots, 10$ sufficed to resolve orientations of the tissue structure (Figures 5.2 and 5.3). The Gabor filter bank and associate oriented structures used here are the same as those introduced in Section 2.3.3 of Chapter 2.

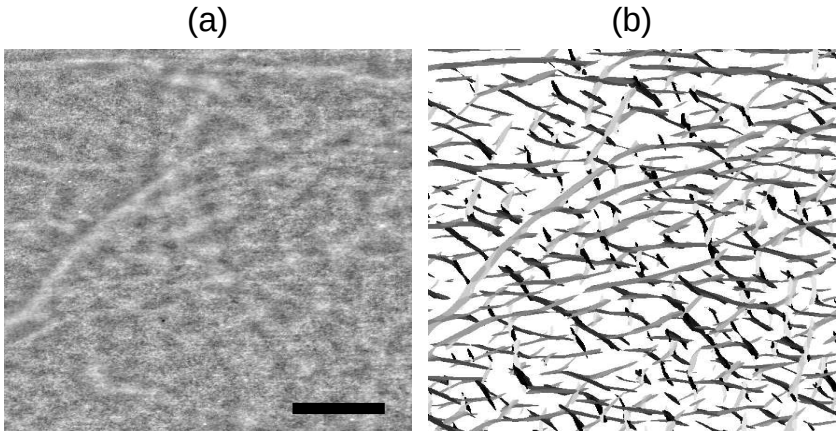

Figure 5.2: An example of oriented tissue structures in an image patch of Figure 2.6: (a) oriented tissue structures in the normalized image patch, scale bar represents 5*mm*, (b) connected components after thresholding the responses of oriented Gabor filters. Features are extracted from individual Gabor filter responses (after thresholding) but in this figure, for illustration only, the connected components from all the responses are shown together with gray levels indicating the various orientations.

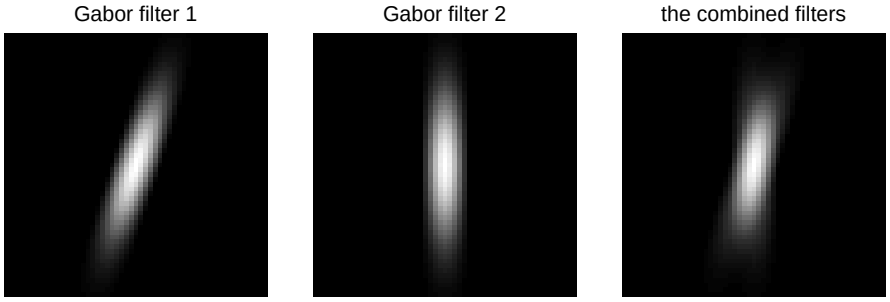

Figure 5.3: Example of Gabor filters in two consecutive orientations ($\frac{9}{10}\pi$ and π) of showing only positive intensity parts: (1) From left to right, the first picture is the Gabor filter at orientation $\frac{9}{10}\pi$. (2) The second picture is the Gabor filter at orientation π. (3) The last picture is the aggregation of the above two Gabor filters.

An empirically defined threshold of $T = 0.20$ was applied to the output image of each oriented Gabor filter (same as Section 4.3.1.3). Connected components in the resulting binary oriented structure image (Figure 5.2) were extracted. For each connected component, four features were recorded (Figure 5.4): (1) Feature f_1 is the distance from the geometric center of the component to the nipple. (2) Feature f_2 is the angle between the line joining the geometric center and the nipple and the horizontal axis. (3) Feature f_3 is the angle between the line joining the geometric center and the nipple and the major axis of the connected component (the orientation of the connected component relative to its angular location with respect to the nipple). (4) Feature f_4 is the area of the connected component.

The feature f_3 was chosen instead of the orientation of the connected component

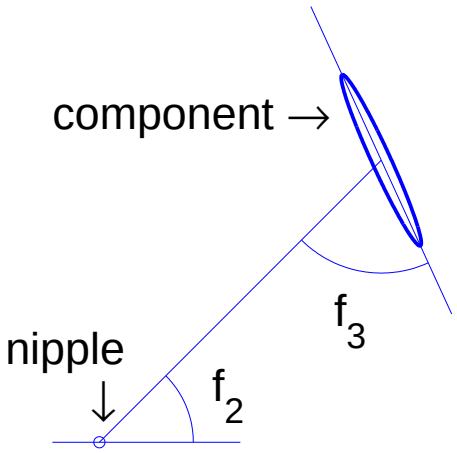

Figure 5.4: Features for oriented tissue structure texture. Feature f_1 (not indicated) is the distance between the nipple and the component which together with feature f_2 gives the location of the component relative to the nipple. Feature f_3 is the angle between the major axis of the elliptical approximation of the component and the line connecting the centroid of the component to the nipple. Feature f_4 is the area of the component (not indicated).

relative to the horizontal axis subsequent to a preliminary study. A histogram of connected component orientations with respect to the horizontal axis of the image resulted in a bimodal distribution while a histogram of connected component orientations with respect to the line connecting the centroid of the component to the nipple resulted in a distinctly mono-modal distribution centered at, and symmetric with respect to, the direction from the component to the nipple (Figure 5.5). This observation is consistent with the general description of oriented breast structure as favoring alignment towards the nipple (Resier et al. [2011, 2012]). The major axis of the connected component was taken to be the direction of the largest eigenvalue obtained by applying principal component analysis (PCA) to the coordinates of the pixels comprising the connected component.

5.3.3 Risk Classification

The two types of texture features calculated from the 160 training images (20 high risk and 20 low risk cases for all four BI-RADS classes) were used to train three classifiers each: an ensemble k-nearest neighbor classifier, a support vector machine (SVM) with the linear kernel (Section 2.5.3), and a Fisher classifier. Then the trained classifier was used to classify testing images into high or low risk group, respectively to show the texture based performance of different regions in risk assessment.

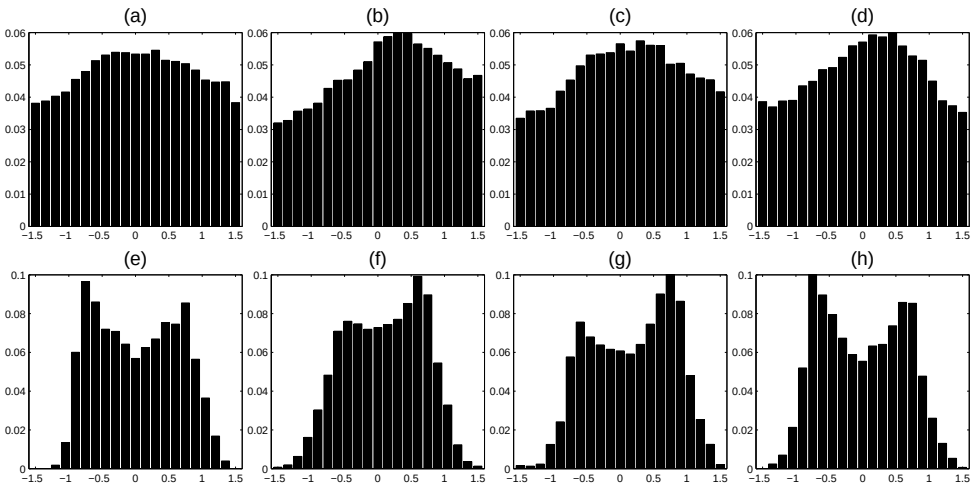

Figure 5.5: Example histograms of angle features for connected components. Top row (a), (b), (c) and (d) are four example histograms of feature f_3. Bottom row (e), (f), (g) and (h) are four example histograms of the orientation of the connected component relative to the horizontal axis.

Table 5.2: Classification performance for texton features with different classifiers; ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier.

(a) Texton features of k-nearest neighbor

region	low risk	high risk	total accuracy	rank
Ω_1	0.650	0.650	0.650	3
Ω_2	0.488	0.613	0.551	6
Ω_3	0.750	0.613	0.682	2
Ω_4	0.588	0.638	0.613	4
Ω_5	0.550	0.600	0.575	5
Ω_6	0.775	0.613	0.694	1

(b) Texton features of SVM

region	low risk	high risk	total accuracy	rank
Ω_1	0.788	0.588	0.688	2
Ω_2	0.700	0.513	0.607	5
Ω_3	0.800	0.563	0.681	3
Ω_4	0.425	0.763	0.594	6
Ω_5	0.675	0.563	0.619	4
Ω_6	0.850	0.575	0.713	1

(c) Texton features of Fisher

region	training AUC	testing AUC	rank
Ω_1	0.861	0.702	2
Ω_2	0.682	0.599	6
Ω_3	0.879	0.694	3
Ω_4	0.832	0.634	4
Ω_5	0.809	0.601	5
Ω_6	0.890	0.763	1

Table 5.3: Classification performance for oriented tissue structure features with different classifiers; ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier.

(a) Oriented tissue features of k-nearest neighbor

region	low risk	high risk	total accuracy	rank
Ω_1	0.538	0.513	0.526	5
Ω_2	0.575	0.425	0.500	6
Ω_3	0.563	0.613	0.588	2
Ω_4	0.600	0.675	0.638	1
Ω_5	0.500	0.588	0.544	3
Ω_6	0.575	0.513	0.544	3

(b) Oriented tissue features of SVM

region	low risk	high risk	total accuracy	rank
Ω_1	0.525	0.438	0.481	6
Ω_2	0.588	0.413	0.500	5
Ω_3	0.588	0.525	0.556	4
Ω_4	0.635	0.563	0.599	2
Ω_5	0.625	0.500	0.563	3
Ω_6	0.638	0.563	0.600	1

(c) Oriented tissue features of Fisher

region	training AUC	testing AUC	rank
Ω_1	0.500	0.438	6
Ω_2	0.538	0.450	5
Ω_3	0.638	0.513	3
Ω_4	0.625	0.650	1
Ω_5	0.673	0.513	3
Ω_6	0.625	0.600	2

5.4 Results

Classification performance based on texton features for local regions and the full breast using ensemble k-nearest neighbor classifier, SVM and the Fisher classifier are shown in parts (a), (b) and (c) of Table 5.2, respectively. Classification results based on oriented tissue structure features for local regions and the full breast using ensemble k-nearest neighbor classifier, SVM and the Fisher classifier are shown in parts (a), (b) and (c) of Table 5.3, respectively. Results in part (a) of Tables 5.2 and 5.3 list the accuracy scores of low risk and high risk group image classification found by ensemble k-nearest neighbor classifier. In part (a) of Table 5.2, the values of k (the number of nearest neighbors) was $k = 3$ for each region. The remaining parameters for the ensemble k-nearest neighbor classifier are the number of features and the number of learners, which were set to 3 and 60, respectively. Details on the implementation of the ensemble classifier including setting of parameters are in Section 2.5.1 of Chapter 2. In part (a) of Table 5.3, the value of k was set to be 5 for each region. The number of predictors and number of learners, were set to 9 and 79, respectively. In parts (a) and (b) of Tables 5.2 and 5.3, the accuracies for the low and high risk groups are listed separately as well as the total accuracy for each region. Total accuracy is the average of low risk and high risk image classification accuracies (proportion of total correct assignments) since the number of testing images in each risk group is the same. The last column of parts (a) and (b) of Tables 5.2 and 5.3 gives the rank of each region in terms of total accuracy. In part (c) of Tables 5.2 and 5.3, the AUC scores are listed for the training and testing sets of images. The last column gives the rank of each region in terms of testing AUC score. The rank provides a quick way to compare the effectiveness of the various regions within each feature class and choice of classifier.

Two-tailed pairwise t-tests were used to determine if the difference in classification performance between texton and oriented tissue structure features could be explained by chance alone. Thus the total accuracy scores of texton features using the ensemble k nearest neighbor classifier (Table 5.2 (a)) were compared with the total accuracy scores of the oriented tissue structure features using ensemble k nearest neighbor classifier (Table 5.3 (a)). The results indicate that there is a significant difference between the performance ($p = 0.043$, $n = 6$) and the mean differences is 0.07. Similarly, the total accuracies found using the SVM (Table 5.2 (b) and Table 5.3 (b)) were found to be significantly different ($p = 0.018$, $n = 6$) with the mean differences of 0.10 as were the testing AUC scores for the Fisher classifier (Table 5.2 (c) and Table 5.3 (c)) ($p = 0.016$, $n = 6$) with the mean differences of 0.23.

5.5 Conclusion and Discussion

In this chapter, two types of texture measures were used to predict the risk of breast cancer; one based on textons derived from intensities in $N \times N$ neighborhoods ($N = 3$) and one based on oriented tissue structure features. For each type of texture feature, three methods were used to assess the quality of risk prediction: accuracy obtained using an ensemble k-nearest neighbor classifier, accuracy obtained using a SVM classifier and the AUC score obtained using the Fisher classifier.

Texton measures outperformed oriented tissue structure features in all three classifiers. According to pairwise t-tests matching regions, the difference was statistically significant at the $p = .05$ level in each case (Section 5.4).

The observation that texture based on intensities in $N \times N$ neighborhoods outperforms texture based on oriented tissue structure is perhaps surprising since there is no known connection between biological properties of breast tissue and the intensity distribution of pixels in 3×3 neighborhoods. A typical breast cancer cell has a diameter of 13μm to 15μm (Sastre-Garau et al. [2004]), well below the spatial Nyquist frequency of the image data which has a spatial resolution of approximately 50μm per pixel. Thus the observed textures do not represent properties of individual cells, for example. The texture features based on oriented tissue structure were chosen with the expectation that variation in this structure would be more likely associated with properties of breast tissue and cancer than the apparently arbitrary features based on $N \times N$ neighborhoods. The opposite is indicated by the results found here. Although the author is not aware of any biological explanation for this observation, other studies have found that classification based on $N \times N$ neighborhoods provides classification equal to, or superior to, classification based on texture features extracted from filter banks that, a priori, seem to be better suited to the task (Varma and Zisserman [2003]).

Leaving aside the biological interpretation, features based on $N \times N$ neighborhoods did provide better prediction of breast cancer and hence these will be the focus of further discussion. Generally, results from the three classifiers (parts (a) - (c) in Table 5.2) indicate that region 1, the region just behind the nipple, provides more information regarding breast cancer risk than any other single region of the breast. This observation is consistent with previous research (Li et al. [2004]) in classifying two risk groups based on BRCA1/2 gene risk factor. Interestingly, region Ω_3 at the back of the breast (closest to the chest wall) also presents good prediction performance. However, region Ω_2, situated between Ω_1 and Ω_3 provided relatively poor prediction. The author is not aware of previous observations along this line and this was not observed in Li et al. [2004].

On the other hand, classification results from oriented tissue structure features

indicate that region Ω_1 is consistently poor while region Ω_4 is consistently high compared to other oriented tissue feature results. The full breast region Ω_6 ranks among the top three regions in all oriented tissue structure feature classification. However, the classification scores from oriented tissue structure are so weak generally, that none of the regions demonstrate a serious contribution to breast cancer risk assessment.

Comparisons between the study in Li et al. [2004] and this study must be made judiciously. The study in Li et al. [2004] focused on distinguishing mammograms of BRCA1 / BRCA2 gene-mutation carriers (high risk) from non carriers (low risk). Here, high risk and low risk are defined as having cancer found in the current screening round or not. Both criteria measure breast cancer risk indirectly and neither can be viewed as a gold standard for breast cancer risk (no such gold standard exists). Since the two studies use different criteria for breast cancer risk, identical results cannot be expected. In addition, the features measured in this study were obtained from locally normalized images (background intensity independent texture features). Direct comparison can be made with Keller et al. (Keller et al. [2012], where the unaffected breast in cancer cases were used as the high risk group, and randomly selected breasts from cases with no cancer detected were used, as the low risk group as was done here. They found that absolute measures of density resulted in AUC scores of $0.65 - 0.67$, percent density resulted in an AUC score of 0.57 and shape-location features resulted in an AUC score of $0.56 - 0.65$. A combined area-volumetric model for density resulted in an AUC score of 0.70. In the study presented here, the testing AUC score for region Ω_6 (Table 5.1 (c)) was 0.763 and so compares favourably with the study by Keller et al. This indicates the density and texture independently contribute at similar levels to breast cancer risk assessment.

The purpose of this study was to determine which region of the breast is best suited for predicting cancer risk based on texture alone. Here, texture alone means texture independent of local tissue density. As a result, density as a determinant factor for breast cancer risk is not used in this study.

The shape of local regions in this study are more naturally associated with the shape of breast and the size of the local regions scale with the size of the particular breast. This seems more reasonable than choosing regions of fixed size and rectangular shape. Nevertheless, the results in this paper largely agree with Li et al. (Li et al. [2004]) in that the best local region is the one just behind the nipple. The regions used in Li et al. (Li et al. [2004]) were smaller and results were not compared with the full breast. The regions used here are the largest possible that roughly coincide with natural regions of the breast. Even if smaller regions could, in principle, provide better risk assessment, the actual improvement would be mitigated by the problem locating these small regions consistently and automatically

within each breast. Due to natural changes in the breast, differences in acquisition parameters, positioning during acquisition and the folding of soft tissue during compression, identifying small regions, even in the same breast in consecutive visits, is very difficult (Ma et al. [2010]). For these reasons and the fact that Li et al. (Li et al. [2004]) found that window size had essentially no effect, exploring a variety regions of different sizes has not been pursued.

In this study, the best prediction of breast cancer risk was obtained by considering the full breast. Thus despite confirming that the single best local region is the region just behind the nipple, the signs of breast cancer risk are not diluted by considering other regions, but instead are enhanced. This observation was not expected given that region Ω_5, obtained by combining regions Ω_1, Ω_2, and Ω_3 did not yield very good prediction results and region Ω_4, comprising the region of the breast lying outside the union of regions Ω_1, Ω_2 and Ω_3, also did not yield very good prediction results.

The two surprising observations in this chapter are that features from 3×3 pixel patches outperform features from oriented tissue structures and that textures measured over the full breast outperform individual regions - may have a common explanation, namely the global nature of textons. Textons represent clusters in feature space from vectors representing the entire image (or region) while the oriented tissue structure features are strictly local. In addition, textons computed over the entire breast are not expected to be the same as the aggregate of textons computed over individual regions. If T_i denotes the set of textons found by clustering the feature vectors from region S_i, $i = 1, 2$, and T denotes the set of textons found by clustering the feature vectors from region $S = S_1 \cup S_2$, then $T_1 \cup T_2$ need not have any elements in common with T. The elements of T are more likely to encompass the variation of appearances in S than the elements of $T_1 \cup T_2$ which may duplicate several main patterns.

Chapter 6

Higher-order Textons

In this chapter, a method is proposed for extending the notion of textons discussed thus far to higher-order textons. Higher-order textons are a novel extension of textons developed by the author in the course of this study.

Section 6.1 introduces the background for developing higher-order textons. Section 6.2 presents the theory for generating higher-order textons and shows the potential of higher-order textons for the classification of texture images with a toy example. Different implementations of higher-order textons are compared in Section 6.3. Texture features calculated from both first-order and higher-order textons are found to perform better than texture features computed from first-order textons alone. Section 6.4 proposes a modified method for generating higher-order textons, which is more reasonable theoretically.

Higher-order textons presented in this chapter will be used for the final temporal breast cancer risk assessment study in Chapter 8.

6.1 Introduction

In order to incorporate the spatial distribution of the texton map, a natural extension is to study the spatial co-occurrence of textons over the image. Schmid (Schmid [2001]) computed "generic descriptors" (textons) based on a "Gabor-like" filter bank and considered spatial frequency clusters. This second-order texton analysis (though not referred to as such) was found to improve image retrieval. However, Varma and Zisserman (Varma and Zisserman [2005]) found that orientation co-occurrence statistics did not improve texture classification.

By Taylor's theorem, higher order polynomial approximations require higher-order derivatives, which in the discrete setting of image analysis, translates to larger neighborhoods to allow numerical estimates of the required derivatives. Theoretically, there is no limit as to how well the surface may be approximated by consid-

ering ever higher-order derivatives and hence ever larger neighborhoods. However, such computations are not practical. First, numerical estimates of high-order derivatives are notoriously unstable. Second, a neighborhood of diameter N results in a feature space of the order of N^2 and so becomes cumbersome for large N. Third, the assumption that the image intensity surface is well approximated by a highly differentiable function is often not valid in images where texture is important. (As an aside, projection images such as the mammograms are inherently represented by discontinuous intensity surfaces.) For these reasons, a more practical approach to capture texture beyond simple low order approximations is to consider patterns of these local low order approximations. Second-order textons capture patterns of first-order textons. As an example, the intersection of two ridges would require a fourth-order polynomial approximation and thus requires a single neighborhood of size 5×5. However, the pattern of quadratic approximations on the nine 3×3 neighborhoods within the 5×5 patch also determines this structure as a combination of local quadratic approximations. The labels associated with these local quadratic structures form the second-order feature vector on which the second-order texton is based. The advantages are that only low-order derivatives (which are numerically more stable) are used, low-dimensional feature spaces are considered (9-dimensional instead of 25-dimensional), and the model assumes only a twice differentiable function instead of a four-times differentiable function.

Although the theoretical basis for the method may be explained in terms of differentiable models of the intensity surface, derivatives are not computed explicitly and polynomial models are not constructed. The implementation relies solely on the patterns of local intensity values.

In this chapter, a general notion of higher-order textons is introduced. Second-order textons are textons defined on texton maps and third-order textons are textons defined on second-order texton maps and so on. In general, applying filters to texton maps is meaningless since the values comprising the texton map are labels and so carry no rank information. This is similar to the maximum Gabor response method of generating first-order textons by Gabor filter presented in Section 2.3.3 of Chapter 2. However, higher-order textons do make sense if the process of extracting the features used to construct textons does not involve arithmetic. The $N \times N$ neighborhood intensity features considered by Varma and Zisserman (Section 2.3.2 of Chapter 2) do not require arithmetic, for example. In this method, each pixel in an image is represented by the feature vector of image intensity values in the $N \times N$ neighborhood of the pixel. Because the $N \times N$ neighborhood involves no arithmetic, this version of texton analysis may be applied to the texton map and, iteratively, to higher-order texton maps. Other examples of texture features that do not require arithmetic include features based on gray scale dependence matrices, also known as

co-occurrence matrices (Haralick et al. [1973]) and run length statistics (Galloway [1975]). The restriction against the use of arithmetic only applies to second and higher-order textons. Any method for constructing textons may be used to arrive at the first-order texton map.

6.2 Higher-order Textons

A general framework for higher-order textons is as follows. Let $X^0 = \{X_1^0, X_2^0, \ldots, X_q^0\}$ denote a collection of images or a single image ($q = 1$) and let $p_{i,j}$ denote pixel j in image X_i^0. Let $f^1(i, j)$ denote the feature vector of length L_1 obtained by computing L_1 features associated with pixel $p_{i,j}$. The components of $f^1(i, j)$ may be outputs from linear filters or other descriptors of local phenomena. There is no restriction to the method of feature extraction used in this step. The collection $f^1(i, j)$ over i and j is viewed as a set of points in an L_1-dimensional feature space. A clustering method is applied to the feature space to identify a set of clusters T_1^1, $T_2^1, \ldots, T_{n_1}^1$. These clusters are the first-order textons. For each i, a new image X_i^1 is formed by assigning label $s \in 1, 2, \ldots, n_1$ to pixel $p_{i,j}$ where s is the index of the cluster closest to $f^1(i, j)$ in the feature space using an appropriate norm (usually the Euclidean distance). The images X_i^1 are called the first-order texton maps. First-order textons are the same as textons considered in previous chapters and first-order texton maps are the same as texton maps, etc.

Second-order textons are obtained by constructing local feature vectors $f^2(i, j)$ of length L_2 on $X^1 = \{X_1^1, X_2^1, \ldots, X_q^1\}$. The features comprising the components of $f^2(i, j)$ must not involve arithmetic operations. Except for this key point, the remaining steps are the same. Thus, a clustering algorithm (not necessarily the same one as used for first-order textons) is applied to the L_2-dimensional feature space to form the second-order textons $T_1^2, T_2^2, \ldots, T_{n_2}^2$ and so on (Figure 6.1). The number of textons at each level (texton order) is not necessarily the same. Final representation or classification can be based on the full collection of textons over all levels or a sub-collection.

The following toy example shows that two images (in this case strings) may be indistinguishable by first- and second-order textons but distinguishable by third-order textons. Consider two 1-dimensional binary images X^0 and Y^0, each comprising m entries labeled 1 and the rest labeled 0. Specifically, the distributions of 1s in X^0 is random but in Y^0 all the 1s appear separated by exactly two 0s (except the first and last 1). Thus$Y^0 = (...., 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, ..., 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0 \ldots)$. The feature vector at position i is $f(i) = (f_1(i), f_2(i))$, where $f_1(i) = Y^0(i-1)$ and $f_2(i) = Y^0(i+1)$. For string Y^0, the feature space obtained by plotting $f(i)$ for all i appears in Figure 6.2 (a). Here A is a large value that depends

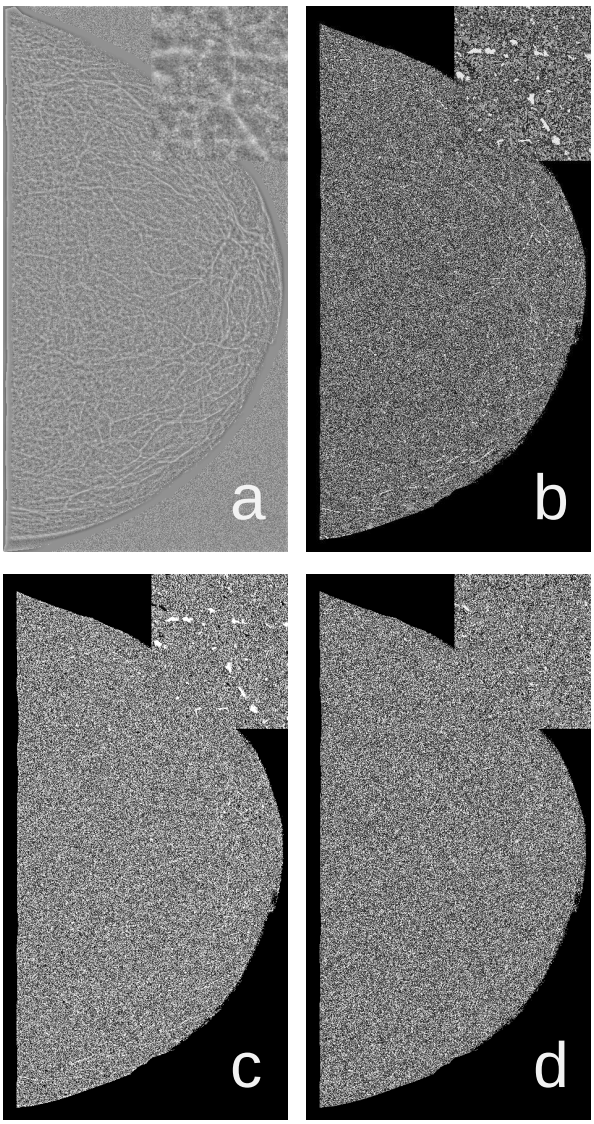

Figure 6.1: Examples of texton maps: (a) the locally normalized image, (b) the first-order texton map, (c) the second-order texton map, (d) the third-order texton map. The first-order texton map for X^1 is the second- order texton map of X^0 and so on. The insets show texture patterns in a patch of size 250×220 from the same location.

(a)

f_2^1		f_1^1	
	1	m	0
	0	A	m
		0	1

(b)

f_2^2		f_1^2		
	2	m	0	0
	1	0	0	m
	0	B	m	0
		0	1	2

(c)

f_2^3		f_1^3			
	3	0	0	m	0
	2	m	0	0	0
	1	0	0	0	m
	0	C	m	0	0
		0	1	2	3

(d)

f_2^3		f_1^3			
	3	0	0	m	0
	2	2	m-2	0	0
	1	0	0	0	m
	0	D	2	0	0
		0	1	2	3

Figure 6.2: First-, second- and third-order feature spaces for the toy example in section 6.2: (a) the feature space for both X^0 and Y^0, (b) the feature space for both X^1 and Y^1, (c) the feature space for X^2 and (d) the feature space for Y^2. A, B, C, D and m are constants that depend only on the length of the strings and not the patterns of 1s and 0s.

on the length of the string and it is assumed that $A \gg m$. Since the resolution in the feature space is low, clustering is not quite meaningful, but a reasonable analog is to accept three clusters $T_1^1 = (0, 0)$, $T_2^1 = (1, 0)$ and $T_3^1 = (0, 1)$. The clusters T_j^1, $j = 1, 2, 3$ are first-order textons. With this choice of clusters, the texton map for Y^0 is given by $Y^1 = (..., 0, 0, 0, 1, 0, 2, 1, 0, 2, 1, 0, 2, 1, ..., 2, 1, 0, 2, 0, 0, 0, \ldots)$. If the same process is applied to Y^1, the feature space becomes Figure 6.2 (b). Here f_1^2 and f_2^2 are used to indicate second-order features. The number B is large and again depends on the length of the string. A natural set of clusters is $T_1^2 = (0, 0)$, $T_2^2 = (1, 0)$, $T_3^2 = (0, 2)$, $T_4^2 = (2, 1)$. These clusters are the second-order textons.

If all the 1s in X^0 are sufficiently separated (as expected since m is small compared to the length) then the first-order and second-order feature spaces of X^0 are identical to the first- and second-order feature spaces of Y^0. However, if the process is applied once more, then the third-order feature spaces for X^0 (Figure 6.2 (c)) and Y^0 (Figure 6.2 (d)) are different. Hence third-order textons distinguish X^0 and Y^0 while lower orders do not.

Computing higher-order textons requires no new techniques. First-order textons are exactly conventional textons. Computing conventional textons requires a choice of filter bank and a choice of clustering algorithm but is otherwise straightforward. This process has been described previously (Chapters 2 - 5). Higher-order textons are computed using exactly the same steps as first-order textons except the labeled texton image (the texton map) from the previous order texton is used in place of the original image. The only restriction is that any filter bank may be used for first-order textons, but higher-order textons are restricted to methods in which the feature

vectors are constructed without arithmetic.

6.3 Implementations of Higher-order Textons

As the focus of the thesis is on texture analysis independent of density in breast cancer risk assessment, local normalization (Chapter 3) will be used to separate texture from density. The purpose of this section is to determine if texture features calculated from higher-order textons have the potential to improve breast cancer risk assessment. Two experiments were conducted on the use of higher-order textons in estimating breast cancer risk: the first using textons based on $N \times N$ neighborhoods, and the second using textons based on Gabor filters.

6.3.1 Data Set

The data set constructed from the DDSM database as described in Section 5.1 was used in the experiments described in this chapter (the first risk criteria in Table 2.1). The whole breast region is used for risk assessment since the whole breast region generally achieved the best performance than other local regions in two risk group classification (Chapter 5).

6.3.2 Textons based on $N \times N$ Neighborhoods

In the first experiment, textons of orders one, two and three were computed based on pixel values in $N \times N$ neighborhoods from normalized images. The procedure of first-order texton generation presented below is the same as the one described in Section 5.3.1 for each region Ω_n but with more details. Each pixel p was replaced by a vector of length $N^2 - 1$ comprising the $N^2 - 1$ normalized intensity values in its $N \times N$ neighborhood (excluding the central pixel p). Then the image array was sub-sampled by $5 \times 5 \longrightarrow 1$ so that every breast tissue patch of 25 pixels of the background intensity independent image was represented by a single vector of length $N^2 - 1$.

In order to avoid missing any fundamental textures from different parenchymal patterns during the construction of the texton dictionary, a separate 5-texton sub-dictionary was constructed for each training BI-RADS class. To do this, separate K-means clustering with $K = 5$ was applied to the $N^2 - 1$-dimensional feature space generated by the training images in each of the four BI-RADS classes. The four 5-texton sub-dictionaries were combined into a single dictionary of 20 textons. The value $K = 5$ in the K-means clustering steps was selected since this was the smallest value of K resulting in a mono-modal distribution for each texton feature over the

set of training images.

With the 20 textons, the first-order texton maps were generated for the 320 mammograms. Pixels outside the template regions were assigned the special texton index 0 and so the first-order texton map comprised 21 texton indexes. Second-order texton dictionaries were generated similarly using the first-order texton maps as input and using the labels in $N \times N$ neighborhoods to construct feature vectors of length $N^2 - 1$. The same clustering process was used and pixels in the first-order texton map were replaced by the index of the second-order textons to arrive at second-order texton maps. Finally, third-order texton maps were constructed in the same way but with the second-order texton maps as input. Consequently, each image was represented by the distribution of the 60 textons (20 for each texton order).

Three classifiers, an ensemble version of k-nearest neighbor classifier, a SVM with linear kernel, and a Fisher classifier were trained and tested for estimating risk based on first-, second- and third-order texton features and combinations thereof. The texton label 0 was omitted from these calculations.

This entire process was repeated for $N = 3$, 5 and 7.

6.3.3 Textons Based on Gabor Filters

In the second experiment, first-order textons were computed from the responses of Gabor filters which are commonly used in texton construction (the Gabor filter bank method presented in Section 2.3.3).

Every pixel p was replaced by a vector of length 10 comprising the 10 Gabor filter responses of pixel p. Similar to the steps for generating textons based on $N \times N$ neighborhoods (Section 6.3.2), the resulting array of vectors was sub-sampled by $5 \times 5 \rightarrow 1$ so that every breast tissue patch of 25 pixels was represented by a single vector of length 10.

In the same way as presented in Section 6.3.2, 20 first-order textons were generated in total from the 10-dimensional feature space of training images of all four BI-RADS classes. Then first-order texton maps were generated from these first-order textons.

Applying Gabor filters again to texton maps is meaningless since it incorporates arithmetic operations and so second- and third-order textons were computed using 3×3 neighborhood representations as described in Section 6.3.2. The 160 training images were used to train the three classifiers and the resulting classifiers were tested on the 160 testing images.

Table 6.1: Risk classification performance for $N \times N$ neighborhood experiments with different sizes of $N \times N$ neighborhoods: (a) total accuracies of ensemble k-nearest neighbor classifier, (b) total accuracies for the SVM classifier, (c) testing AUC scores for the Fisher classifier.

(a)

texton order	3×3	5×5	7×7
1st	0.663	0.644	0.625
2nd	0.526	0.613	0.550
3rd	0.526	0.500	0.588
1st & 2nd	0.675	0.638	0.613
1st & 3rd	0.576	0.657	0.606
2nd & 3rd	0.507	0.563	0.588
1st & 2nd & 3rd	0.663	0.644	0.638

(b)

texton order	3×3	5×5	7×7
1st	0.713	0.656	0.650
2nd	0.594	0.544	0.519
3rd	0.563	0.500	0.538
1st & 2nd	0.744	0.631	0.613
1st & 3rd	0.688	0.650	0.588
2nd & 3rd	0.588	0.544	0.581
1st & 2nd & 3rd	0.732	0.650	0.600

(c)

texton order	3×3	5×5	7×7
1st	0.760	0.648	0.634
2nd	0.636	0.558	0.563
3rd	0.573	0.532	0.536
1st & 2nd	0.775	0.637	0.582
1st & 3rd	0.728	0.677	0.611
2nd & 3rd	0.597	0.570	0.595
1st & 2nd & 3rd	0.723	0.679	0.590

6.3.4 Results

With three orders of textons, there are seven texton order combinations (1st, 2nd, 3rd, 1st & 2nd, 1st & 3rd, 2nd & 3rd and 1st & 2nd & 3rd). The ensemble k-nearest neighbor classifier achieved the highest performance score with $N = 3$ for three of the seven order combinations and twice each with $N = 5$ and $N = 7$ (Table 6.1). For both the SVM and Fisher classifiers, the highest performance was achieved with $N = 3$ for every texton order combination (Table 6.1). In addition, the highest performance over all the texton order combinations was achieved with $N = 3$ and the combination of 1st- and 2nd-order textons for every classifier (Table 6.1). This result, in combination with the fact that larger N results in increased computational load, motivated the use of $N = 3$ in computing higher-order textons in the $N \times N$ neighborhood experiment and Gabor filter experiment. Thus further analysis was focused on the $N = 3$ results.

The two experiments both used features obtained from 3×3 neighborhoods to construct second- and third-order textons. The experiments differed in the method used to construct first-order textons. For ease of exposition, the method described in Section 6.3.2 in which first-order textons were based on 3×3 neighborhoods will be referred to as the 3×3 method and the method described in Section 6.3.3 will be referred to as the Gabor filter method.

The classification results from the SVM classifier with the linear kernel were recorded as the proportion of correct assignments of the testing images for high or low risk groups separately (Tables 6.2 (b) and 6.3 (b)). For the Fisher classifier, AUC score was computed for both the training and the testing data (Tables 6.2 (c) and 6.3 (c)). For the ensemble k-nearest neighbor classifier, classification results were recorded as the proportion of testing images correctly assigned as high or low risk groups separately (Tables 6.2 (a) and 6.3 (a)). The ensemble method yielded a value of $k = 7$ for the 3×3 experiment (Section 6.3.2). The ensemble method requires two additional parameters; the number of features and the number of learners. These were 18 and 5 respectively. For the Gabor filter experiment (Section 6.3.3), the parameter values were $k = 25$, 19 for the number of features and 45 for the number of learners. Details on how these parameters are determined have appeared in Section 2.5.1 of Chapter 2. The total accuracy in parts (a) and (b) of Tables 6.2 and 6.3 is the total proportion of correct assignments.

In order to decide if higher-order textons provide a statistically significant improvement over first-order textons alone, 5-fold cross validation was performed using the Fisher classifier and the AUC score as performance measure. For the 3×3 method, the best performance was achieved by the combination of first- and second-order textons albeit the improvement was not statistically significant (Table 6.4). For the Gabor filter method, the best performance was achieved by the combination of all three texton orders and this improvement was significant at the $p = .05$ level (Table 6.5). For the Gabor filter method, the combination of first- and second-order textons, the combination of first- and third-order textons, and third-order textons on their own all performed significantly better at the $p = .05$ level than first-order textons alone (Table 6.5).

In order to confirm that the classification performance of 3×3 method was not due to chance alone, the Fisher classifier was applied to randomly generated data matching the experimental data in terms of sample size and numbers of features. 5000 training and testing trials were run for each analogous texton group to determine the mean and std (standard deviation) of the AUC score for such random data. For random training data comprising 20 features (analogous to first-, second- and third-order textons separately), the mean AUC score was $0.701 \pm .033$. For random training data comprising 40 features (analogous to combining two of first-, second- or third-order textons), the mean AUC score was 0.792 ± 0.030 and for training data comprising 60 features (analogous to combining textons of all orders) the mean AUC score was 0.862 ± 0.028. For random testing data, the mean AUC was 0.536 ± 0.028 regardless of the numbers of features.

Table 6.2: Classification performance for higher-order textons using the 3×3 method with different classifiers: ensemble k-nearest neighbor classier, SVM classifier and Fisher classifier. The rank for (a) and (b) is based on the total accuracy. The rank for (c) is based on the testing AUC score.

(a) 3×3 method with ensemble k-nearest neighbor classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.713	0.613	0.663	2
2nd	0.463	0.588	0.526	4
3rd	0.463	0.588	0.526	4
1st & 2nd	0.650	0.700	0.675	1
1st & 3rd	0.563	0.588	0.576	3
2nd & 3rd	0.563	0.450	0.507	5
1st & 2nd & 3rd	0.663	0.663	0.663	2

(b) 3×3 method with SVM classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.850	0.575	0.713	3
2nd	0.650	0.538	0.594	5
3rd	0.638	0.488	0.563	7
1st & 2nd	0.850	0.638	0.744	1
1st & 3rd	0.838	0.513	0.688	4
2nd & 3rd	0.638	0.538	0.588	6
1st & 2nd & 3rd	0.875	0.588	0.732	2

(c) 3×3 method with Fisher classifier

texton order	training AUC	testing AUC	rank
1st	0.890	0.763	2
2nd	0.852	0.636	5
3rd	0.711	0.573	7
1st & 2nd	0.928	0.775	1
1st & 3rd	0.904	0.728	3
2nd & 3rd	0.762	0.597	6
1st & 2nd & 3rd	0.953	0.723	4

Table 6.3: Classification performance for higher-order textons using the Gabor filter method with different classifiers: ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier. The rank for (a) and (b) is based on the total accuracy. The rank for (c) is based on the testing AUC score.

(a) Gabor filter method with ensemble k-nearest neighbor classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.350	0.725	0.538	6
2nd	0.413	0.713	0.563	4
3rd	0.525	0.675	0.600	2
1st & 2nd	0.380	0.788	0.584	3
1st & 3rd	0.338	0.788	0.563	4
2nd & 3rd	0.500	0.738	0.619	1
1st & 2nd & 3rd	0.350	0.800	0.575	5

(b) Gabor filter method with SVM classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.675	0.588	0.632	2
2nd	0.713	0.500	0.607	3
3rd	0.588	0.550	0.569	5
1st & 2nd	0.688	0.588	0.638	1
1st & 3rd	0.663	0.500	0.582	4
2nd & 3rd	0.650	0.475	0.563	6
1st & 2nd & 3rd	0.650	0.513	0.582	4

(c) Gabor filter method with Fisher classifier

texton order	training AUC	testing AUC	rank
1st	0.874	0.639	2
2nd	0.783	0.539	6
3rd	0.754	0.607	4
1st & 2nd	0.928	0.656	1
1st & 3rd	0.945	0.591	5
2nd & 3rd	0.511	0.470	7
1st & 2nd & 3rd	0.957	0.633	3

Table 6.4: 5-fold cross validation results for the $N \times N$ method with $N = 3$ for all texton orders. "texton order" refers to the texton order or combination of texton orders, "mean" is the mean of the AUC scores from the cross validation, "std" is the standard deviation of the AUC scores from the cross validation, and p-value is the probability that the mean is different from the mean of the first-order texton (on its own) by chance alone.

texton order	1st	2nd	3rd	1st & 2nd	1st & 3rd	2nd & 3rd	1st & 2nd & 3rd
mean	0.764	0.643	0.577	0.772	0.727	0.602	0.729
std	0.064	0.071	0.110	0.057	0.091	0.120	0.087
p-value	-	0.002	0.001	0.595	0.080	0.009	0.189

Table 6.5: 5-fold cross validation results for the Gabor filter method. The rows have the same meaning as in Table 6.4.

texton order	1st	2nd	3rd	1st & 2nd	1st & 3rd	2nd & 3rd	1st & 2nd & 3rd
mean	0.486	0.525	0.618	0.630	0.595	0.472	0.646
std	0.087	0.071	0.070	0.086	0.076	0.034	0.047
p-value	-	0.481	0.019	0.049	0.019	0.746	0.015

6.3.5 Conclusion and Discussion

As presented in Section 5.3.1, the literature reports that 3×3 neighborhoods have been found to be the best choice for applying first-order textons compared to other $N \times N$ neighborhoods. The results found here (Tables 6.1, 6.2 and 6.3 in this chapter and Tables B.3 and B.4 in Section B.2 of Appendix B) are in keeping with these trends. $N = 3$ produces results as good or better than larger values of N in nearly all situations and reduces run time.

For the data set considered in this study, the SVM classifier provides greater total accuracy than the ensemble k-nearest neighbor classifier for all texton order combination for the $N \times N$ method (Table 6.2 (a) and (b)) and for five out of seven texton order combination for the Gabor filter method (Table 6.3 (a) and (b)). Accordingly, greater store should be placed on the SVM results. The Fisher classifier cannot be compared directly to SVM or ensemble k-nearest neighbor classifier since the performance criterion (AUC score) is not the same.

In comparing Table 6.2 (c) to the random data results (Section 6.3.4 Paragraph 5), all orders of textons on their own performed better than could be expected by chance alone although this was only marginally true for third-order textons. Thus higher-order textons presented do have positive prediction value for estimating risk of breast cancer.

Ensemble k-nearest neighbor classification results of testing images (Table 6.2 (a)), SVM classification results of testing images (Table 6.2 (b)) and AUC scores of Fisher classifier (Table 6.2 (c)) indicate that first-order textons contributed more to the classification task than any other order of textons taken on its own. The toy example (Section 6.2) shows that this is not automatically true, but for real images, this is not surprising since first-order textons interrogate the image most directly.

For both the 3×3 method and the Gabor filter method, the combination of textons of orders 1 and 2 outperformed textons of order 1 alone (Tables 6.2 and 6.3). In all these cases, except the case of k-nearest neighbors applied to the Gabor filter method, the combination of textons of orders 1 and 2 outperformed all other combinations of texton orders. For k-nearest neighbors applied to the Gabor filter method the combination texton orders 2 and 3 outperformed all others.

The cross validation results (Tables 6.4 and 6.5) support the finding that combinations of texton orders provide higher performance than first-order textons alone. Although the improvement was not statistically significant for the 3×3 method (Table 6.4), a significant improvement was found for the Gabor filter method for all but two higher-order textons or texton combinations (Table 6.5). These results indicate that higher-order textons have the potential to improve on conventional textons.

The results found here can not be compared to studies in which other criteria for risk were used, such as Wolfe's patterns (Magnin et al. [1986], Caldwell et al. [1990], Tahoces et al. [1995]), SCC categories (Byng et al. [1996, 1997]), BI-RADS classes (Petroudi et al. [2003], Gong and Petroudi [2006]), carriers of BRCA1/2 gene-mutation (Huo et al. [2000, 2002], Li et al. [2010]) or ER receptor status (Karemore et al. [2012]). Keller et al. (Keller et al. [2012]) used the same criterion for risk as this study. Several measures of the distribution of breast density from area-volumetric model were used to estimate risk resulting in AUC scores in the range of 0.65 to 0.70. Thus the 5-fold cross validated scores of 0.772, 0.727 and 0.720 obtained by the combinations of first- and second-order textons, first- and third-order textons and first-, second- and third-order textons, respectively, compare favorably (Table 6.4).

6.4 Label-Independent Higher-order Texton Generation using $N \times N$ Neighborhoods

In the experiments reported in the previous section (Section 6.3), second- and third-order textons were generated from feature vectors consisting of texton labels based on first- and second-order textons, respectively. A key point was made that the computation of feature vectors for higher-order textons must not involve arithmetic of texton labels. The use of arithmetic was avoided by using the $N \times N$ neighborhood method for generating higher-order textons. Even so, the method described in Section 6.3 depends on the choice of labeling at the clustering stage. Since the clustering step does use arithmetic, this introduces a theoretically unsatisfactory aspect to higher-order textons. Experiments were conducted to ascertain the practical consequences of this label dependence. Results indicate that the choice of labels has little effect on actual risk assessment (Table 6.6). Nevertheless, a method for generating higher-order textons that is truly independent of the choice of labels in the clustering step is desirable.

Table 6.6: Classification AUC scores for second-order textons calculated from several relabeled first-order texton maps. The second row shows the AUC scores of the original first-order texton maps. The third row shows the AUC scores when relabeling texton 7 and 8 by 21 and 22. The fourth row shows the AUC scores when relabeling texton 7 and 8 by 25 and 29. The fifth row shows the AUC scores when relabeling texton 7 and 8 by 25 and 32.

tests of relabeling	training AUC scores	testing AUC scores
original texton map (7, 8)	0.852	0.636
7, 8 $\rightarrow$ 21, 22	0.856	0.678
7, 8 $\rightarrow$ 25, 29	0.817	0.638
7, 8 $\rightarrow$ 25, 32	0.843	0.638

In this section, a variation on the method for generating textons is introduced that does not depend on the label assignments in the clustering step. For the convenience of exposition, the higher-order textons discussed in Sections 6.1 - 6.3 will be referred to as the label-dependent higher-order textons and the textons described in this section will be called label-independent higher-order textons.

Consider the first-order texton map generated by any of the methods discussed thus far. Instead of using the texton labels in a neighborhood as components of a feature vector, the labels in the neighborhood are used to construct a local histogram of texton labels. If there are 20 first order textons, the local histograms will comprise 20 bins. The feature vector associated with a pixel is the histogram of first-order texton occurrences in the $N \times N$ neighborhood of the pixel. Thus the feature vector is the vector of length 20 comprising the 20 histogram values. In the clustering step, the arithmetic operations necessary for determining cluster centers (the second-order textons) are performed on the histogram values and not on the labels themselves. Similarly, third-order textons are generated from second-order texton maps, and so on. Examples of label-independent higher-order texton maps are shown in Figure 6.3.

In label-independent higher-order texton generation, regardless of the labeling choice, the feature vector values are the same because they are the occurrence of textons. This means the values of feature vectors are actual number with rank meaning rather than a labeled number with hypothetical meaning. So label-independent higher-order texton generation is more satisfying theoretically. By using 3×3 neighborhoods for generating first-, second- and third-order textons in this way, the risk classification performance of different order texton features and the combination of them are shown in Table 6.7.

According to the results for label-independent higher-order texton generation in Table 6.7, it seems the inclusion of higher-order textons did not improve the risk performance over using first-order textons alone. However, the performance of higher-

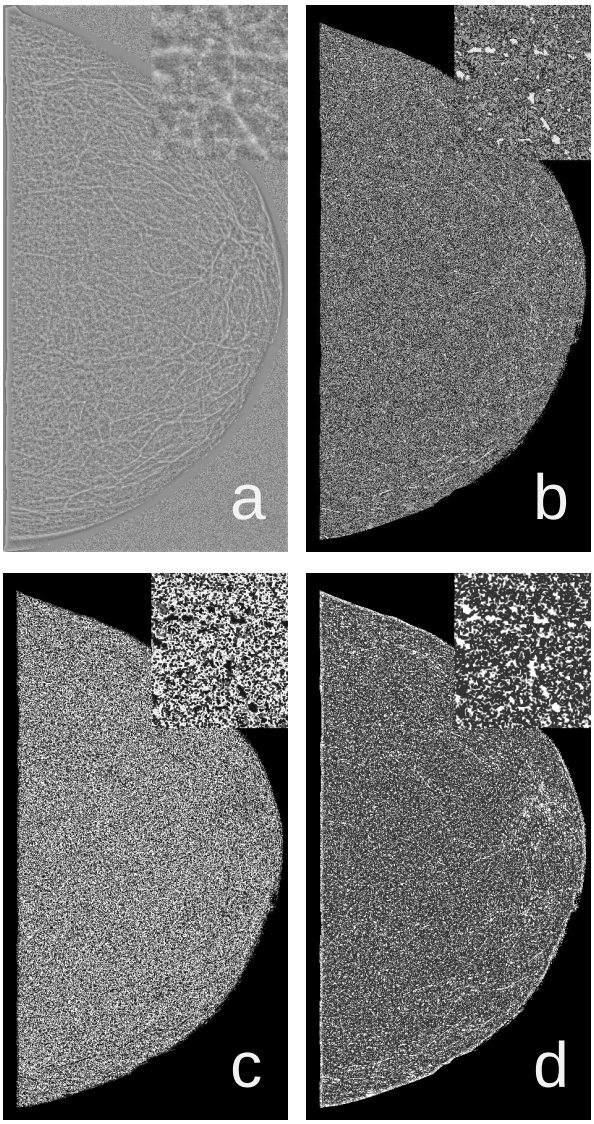

Figure 6.3: Examples of texton maps for label-independent higher-order texton generation: (a) the locally normalized image, (b) the first-order texton map, (c) the second-order texton map, (d) the third-order texton map. The inset shows texture patterns in a patch of size 250 $\times$ 220 from the same location.

Table 6.7: Classification performance for label-independent higher-order textons using the 3×3 method with different classifiers: ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier. The rank for (a) and (b) is based on the total accuracy. The rank for (c) is based on the testing AUC score.

(a) Ensemble k-nearest neighbor classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.713	0.613	0.663	1
2nd	0.638	0.625	0.631	3
3rd	0.550	0.663	0.588	4
1st & 2nd	0.650	0.638	0.643	2
1st & 3rd	0.613	0.650	0.631	3
2nd & 3rd	0.575	0.688	0.631	3
1st & 2nd & 3rd	0.625	0.638	0.631	3

(b) SVM classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.850	0.575	0.713	1
2nd	0.675	0.613	0.644	6
3rd	0.425	0.775	0.600	7
1st & 2nd	0.800	0.538	0.669	3
1st & 3rd	0.825	0.500	0.663	4
2nd & 3rd	0.638	0.663	0.651	5
1st & 2nd & 3rd	0.775	0.613	0.694	2

(c) Fisher classifier

texton order	training AUC	testing AUC	rank
1st	0.892	0.760	1
2nd	0.866	0.615	6
3rd	0.819	0.570	7
1st & 2nd	0.910	0.740	2
1st & 3rd	0.878	0.726	3
2nd & 3rd	0.913	0.626	5
1st & 2nd & 3rd	0.954	0.703	4

order textons together with first-order textons is not significantly lower than that of using first-order textons alone. In addition, the above performance was computed without feature selection and cross validation. To some extent, the inclusion of higher-order textons without feature selection did not weaken the general performance too much and the performance reduction might due to the use of redundant features. In addition, label-independent higher-order textons are theoretically more suitable and the focus is on measuring texton map from another point of view. As a result, the issue as to which of these two methods for computing higher-order textons is taken to be unresolved. In the next chapter, Chapter 7, both methods are used to compare the contribution of texture and density to risk assessment.

Chapter 7

Texture versus Density

Density is generally regarded as the most important mammographic risk factor for cancer (Section 1.3). Here the contributions of texture and density to risk assessment are compared (the third risk criteria in Table 2.1).

Texture features calculated from higher-order textons generated from the two methods in Chapter 6 and a single density feature calculated from h_{int} images (Section 7.2) are compared in estimating breast cancer risk. In Chapter 5, the comparison between the performance of texton features and density features proposed in the study by Keller et al. (Keller et al. [2012]) showed that the contribution of texture and density to risk assessment are similar. However, this comparison is not conclusive since the two studies used different mammogram image data sets. In addition, the study by Keller et al. (Keller et al. [2012]) used feature selection and leave-one-out cross validation, neither of which were used in the texture study in Chapter 5. In that chapter, the work by Keller was used only to see if the performance of texture was reasonable in comparison to density. In this chapter, density and texture are compared using the same data set, feature selection and proper validation. Hence the conclusions regarding texture and density reached in Chapter 5 are reassessed in this chapter using higher-order textons, a common data set and proper validation.

Results indicate that the contribution to breast cancer risk assessment of texture alone is at least as important as density alone. Combining texture and density features does not always perform better than texture or density alone, but depends on the way texture and density features are combined and the number of texture or density features used. In addition, texture features calculated from label-independent higher-order textons seem to perform better than texture features computed from label-dependent higher-order textons if sequential feature selection and 5-fold cross validation are used.

The structure of this chapter is as follows: Section 7.1 describes risk classification carried out with texture features alone, Section 7.2 describes risk classification carried out with the density feature alone, Section 7.3 describes risk classification

carried out with combined texture and density features, Section 7.4 presents the risk classification results of the previous three sections and Section 7.6 presents conclusions and discussions arising from these results.

7.1 Risk Classification with Texture Features

Two processes for calculating higher-order textons, label-dependent and label-independent, were described in Sections 6.3.2 and 6.4 of Chapter 6, respectively. Regardless of which of these two higher-order texton generation methods are used, there is a total of 60 texture features. (The 20 first-order texton features are the same for each of higher-order texton generation method.) To determine the performance of texture alone, two methods of feature selection are tested; sequential backward feature selection and exhaustive search feature selection. Sequential backward feature selection was applied to the training set of 80 low and 80 high risk images to reduce the 60 texton features to between 1 and 6 features. Exhaustive search feature selection was used to select between 1 and 4 features using the set of training images. The final number of features was restricted to 4 in the exhaustive search method due to computational load. With 4 features chosen from 60 there are $\begin{pmatrix} 60 \\ 4 \end{pmatrix} = 487,635$ combinations to test but with 6 features there are 50,063,860 combinations. The AUC score was used as the classification criterion. Five-fold cross validation on the testing set of 80 low and 80 high risk images was used to arrive at an estimate of classification performance on unseen data. Thus, for sequential backward feature selection, six prediction classification scores were found: one for the best single feature, one for the best two features, and so on up to the best combination of 6 features. Similarly, for exhaustive search feature selection, four prediction classification scores were found. The reason for choosing 5-fold cross validation is that the image data set used (the same as in Chapters 5 and 6) is reasonably big and so $k = 5$ follows the recommendation in the literature (Section 2.6).

7.2 Risk Classification with a Density Feature

In order to compare texture to density, a reliable measure of breast density is needed. Here, density scores are obtained by imitating the methods introduced by Highnam et al. (Highnam et al. [2010]) based on the h_{int} model. In this model, scattering of X-rays through two classes of breast tissue, "interesting" and fat tissue, is differentiated. The fat component is removed leaving behind an image that represents just the interesting tissue in the breast. A simple normalized sum is then used to calculate the density estimate. The details of the h_{int} calculation are somewhat opaque so what

follows is an empirical reconstruction of the algorithm.

Figure 7.1 shows a schematic arrangement during the process of taking a mammogram. The breast is compressed between two paddles meaning that, over the majority of the breast, the X-rays pass through a constant depth of tissue. Fatty tissue is assumed to absorb uniformly while the scatter due to the parenchymal breast tissue is modeled as a radially symmetric function, shown in Figure 7.2.

To take account of the non-linear nature of the imaging process, the first step is to normalize the local variation in the image. This is achieved using the local normalization of Section 3.2 except the local mean is restored to the image.

The next step is to focus the image by removing the effect of the scatter. The normalized image is filtered using the scatter function and the result subtracted from the normalized image. Near the breast boundary, the scatter in the breast can exceed the local intensity, producing negative values which are a measure of the proportion of fatty tissue in the breast. By setting all these negative values to zero we are left with a representation of the "interesting" tissue in the breast (hence the designation h_{int}).

This processing leaves the dynamic range of the image somewhat compressed and the final step is to perform histogram equalization to restore that range to one similar to the original to produce the h_{int} image. The density feature is then simply the sum of the non-zero elements of the h_{int} image divided by the number of non-zero elements.

In the implementation of this density feature in this section, 5-fold cross validation was used on the 80 high and 80 low risk testing image set to estimate the classification performance according to the AUC score.

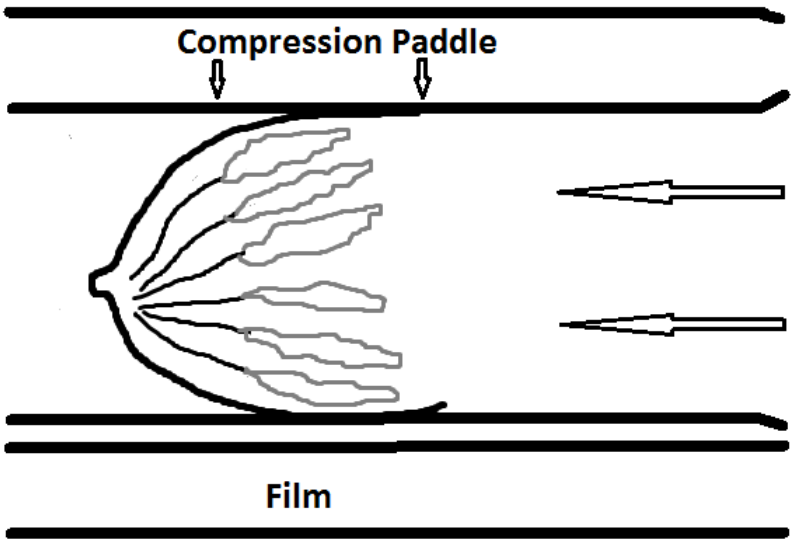

Figure 7.1: Schematic of the physical layout of a mammographic X-ray machine.

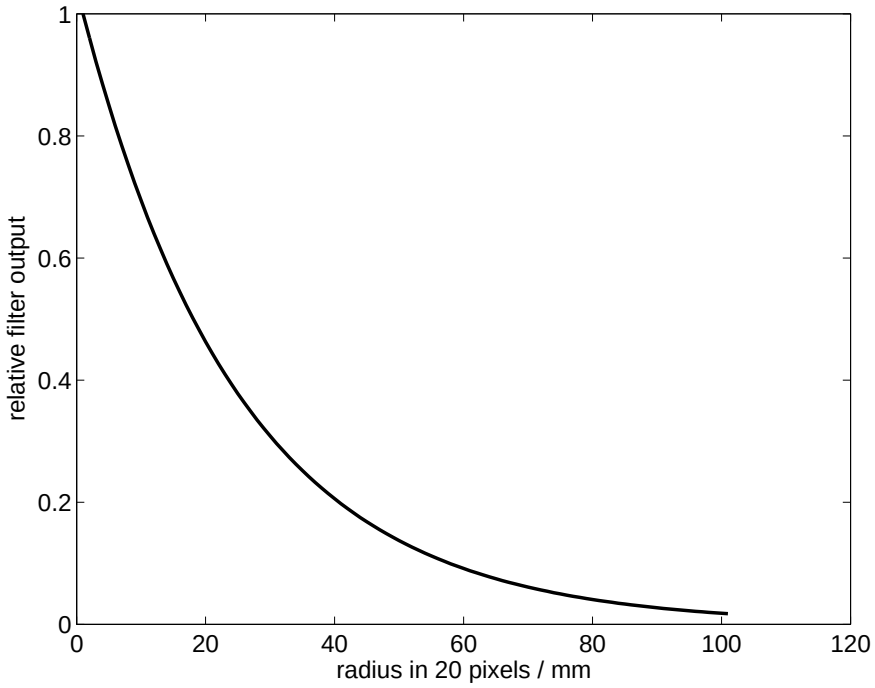

Figure 7.2: Radial projection of the scattering filter used in the density feature calculation.

7.3 Risk Classification with Combined Texture and Density Features

Three methods for combining texture and density features are considered: the augmented feature set method, the reselected feature set method, and the recalculated feature set method.

7.3.1 The Augmented Feature Set Method

In the augmented feature set method, each of the first five optimal texton feature sets found with sequential backward feature selection (Section 7.1) was augmented by the single density feature (Section 7.2), resulting in five feature sets of sizes 2, 3, 4, 5, 6. The performance of these feature sets were evaluated using 5-fold cross validation on the 160 testing images. Similarly, each of the first three optimal texton feature sets found with exhaustive search feature selection (Section 7.1) was also augmented by the single density feature resulting in three feature sets of sizes 2, 3 and 4. Again, the performance of these feature sets were evaluated using 5-fold cross validation on testing images.

7.3.2 The Reselected Feature Set Method

In the reselected feature set method, the density feature was added initially to the pool of the original 60 texture features. Sequential backward feature selection was then applied to this augmented feature set of size 61 to obtain new optimal feature sets of size 1, 2, ..., 6 based on the training set. Next, 5-fold cross validation was used on the testing set to estimate the performance of these six feature sets on unseen data. Similarly, exhaustive search feature selection was applied to this augmented feature set of size 61 to obtain new optimal feature sets of size 1, 2, 3, and 4. Again, 5-fold cross validation was used on the testing set to estimate the performance of these four feature sets on unseen data.

7.3.3 The Recalculated Feature Set Method

In the recalculated feature set method, texture features were calculated from h_{int} images introduced in Section 7.2 instead of the normalized images used elsewhere in the thesis. The h_{int} image is different from the normalized image, in that density information is explicitly retained. Here, label-dependent and label-independent higher-order texton method (Sections 6.2 and 6.4) was applied to h_{int} images respectively. Thus a total of 60 texton density features was available. Sequential backward feature selection was used on these 60 higher-order texton density features to reduce the number of features to between 1 and 6 based on training images. The selected optimal texton density features were used to classify testing images and the classification performance was estimated by 5-fold cross validation. Similarly, exhaustive search feature selection was used to select a number of features between 1 and 4 based on training images. Again, selected optimal texton density features were used for risk classification on testing images and the performance was estimated by 5-fold cross validation.

7.4 Results for Sequential Feature Selection

Risk classification performances for optimal texture features, the density feature, and the combination of texture and density features with the augmented feature set method, the reselected feature set method and the recalculated feature set method obtained from sequential feature selection are presented in Table 7.1 and Table 7.2 (Detailed results are in Tables B.5 to B.12 in Section B.3 of Appendix B).

Risk classification based on density alone is better than risk classification based on any single texture feature (Table 7.1 and Table 7.2) but lies within one standard deviation (SD) of the single texture feature result. Density alone does not perform as well as risk classification based on the best combination of four or more texture

Table 7.1: Classification AUC scores for the label-dependent higher-order texton method for texture alone, density alone and the combination of texture and density according to the augmented method, the reselected method and the recalculated method. Values are the 5-fold cross validation averages ± SD. n denotes the number of optimal features which were obtained from sequential backward feature selection. The index set for the texture features comprising the optimal set of n features is shown underneath. Indices $1-20$ are first-order textons, $21-40$ are second-order textons, and $41-60$ are third-order textons. The label d is used to denote the single density feature.

n	texture alone	augmented	reselected	recalculated	density
1	0.723±0.052	-	0.723±0.052	0.588±0.083	0.740±0.032
	{15}	-	{15}	{12}	{d}
2	0.719±0.078	0.724±0.057	0.719±0.078	0.674±0.069	-
	{11,15}	{15,d}	{11,15}	{12,49}	-
3	0.770±0.059	0.725±0.074	0.770±0.059	0.664±0.071	-
	{9,11,15}	{11,15,d}	{9,11,15}	{12,49,54}	-
4	0.771±0.056	0.759±0.059	0.771±0.056	0.725±0.058	-
	{9,11,15,	{9,11,15,	{9,11,15,	{12,49,54,	-
	49}	d}	49}	59}	-
5	0.765±0.057	0.761±0.051	0.765±0.057	0.717±0.053	-
	{9,11,15,	{9,11,15,	{9,11,15,	{12,30,49,	-
	49,56}	49,d}	49,56}	54,59}	-
6	0.749±0.058	0.773±0.054	0.749±0.058	0.734±0.060	-
	{9,11,15,	{9,11,15,	{9,11,15,	{12,30,34,	-
	49,50,56}	49,56,d}	49,50,56}	49,54,59}	-

features, but lies within one SD of all optimal combinations of six or fewer texture features (Table 7.1 and Table 7.2). Augmenting the best texture feature set with the density feature, reselecting the best features with density included in the pool of features, or recalculating the best features dependent of density from h_{int} images does not always improve the classification.

7.5 Results for Exhaustive Search Feature Selection

Risk classification performances for optimal texture features, the density feature and the combination of texture and density features with the augmented feature set method, the reselected feature set method and the recalculated feature set method obtained from exhaustive search feature selection are presented in Table 7.3 and Table 7.4 (Detailed results are in Tables B.13 to B.20 in Section B.3 of Appendix B).

Table 7.2: Classification AUC scores for the label-independent higher-order texton method for texture alone, density alone and the combination of texture and density according to the augmented method, the reselected method and the recalculated method. All the values have the same meaning as explained in Table 7.1.

n	texture alone	augmented	reselected	recalculated	density
1	0.723±0.052	-	0.723±0.052	0.476±0.077	0.740±0.032
	{15}	-	{15}	{35}	{d}
2	0.722±0.084	0.724±0.057	0.719±0.080	0.587±0.121	-
	{15,60}	{15,d}	{15,45}	{10,35}	-
3	0.718±0.084	0.727±0.083	0.717±0.083	0.658±0.129	-
	{12,15,60}	{15,60,d}	{12,15,45}	{10,17,35}	-
4	0.751±0.094	0.723±0.086	0.711±0.082	0.688±0.105	-
	{12,15,25,	{12,15,60,	{12,15,34,	{12,17,30,	-
	60}	d}	45}	35}	-
5	0.756±0.091	0.746±0.093	0.719±0.081	0.679±0.101	-
	{4,12,15,	{12,15,25,	{12,15,34,	{10,17,30,	-
	25,60}	60,d}	40,45}	35,43}	-
6	0.775±0.083	0.754±0.091	0.715±0.081	0.684±0.094	-
	{4,12,15,	{4,12,15,	{12,15,34,	{10,17,23,	-
	25,27,60}	25,60,d}	40,43,45}	30,35,43}	-

Risk classification based on density alone performs better than risk classification based on any single texture feature (Table 7.3 and Table 7.4) but lies within one SD of the single texture feature result. This is also true for risk classification based on more than one texture features. Augmenting the best texture feature set with the density feature, reselecting the best features with density included in the pool of features or recalculating the best features dependent of density from h_{int} images does not always improve the classification.

7.6 Conclusion and Discussion

Conclusion and discussion for Chapter 7 were made according to classification performance for testing images. Anywhere in this thesis, classification performance alone indicates classification performance for testing images. Globally, exhaustive search feature selection finds the maximum performance for training images (Appendix B.4). However, in this chapter, training results for exhaustive search feature selection are lower than sequential feature selection (Appendix B.3). This is because feature selection was optimized over the whole training set. On the full training data, sequential feature selection was not better than exhaustive search (Appendix B.4).

Table 7.3: Classification AUC scores for the label-dependent higher-order texton method for texture alone, density alone and the combination of texture and density according to the augmented method, the reselected method and the recalculated method. Values are the 5-fold cross validation averages ± SD. n denotes the number of optimal features which were obtained by exhaustive search feature selection. The index set for the texture features comprising the optimal set of n features is shown underneath. Indices $1-20$ are first-order textons, $21-40$ are second-order textons, and $41-60$ are third-order textons. The label d is used to denote the single density feature.

n	texture alone	augmented	reselected	recalculated	density
1	0.723±0.052	-	0.723±0.052	0.588±0.083	0.740±0.032
	{15}	-	{15}	{12}	{d}
2	0.723±0.055	0.724±0.057	0.723±0.055	0.553±0.066	-
	{13,15}	{15,d}	{13,15}	{12,30}	-
3	0.715±0.091	0.719±0.064	0.715±0.091	0.664±0.071	-
	{13,15,33}	{13,15,d}	{13,15,33}	{12,49,54}	-
4	0.729±0.079	0.718±0.089	0.729±0.079	0.725±0.058	-
	{13,15,33,	{13,15,33,	{13,15,33,	{12,49,54,	-
	37}	d}	37}	59}	-

Table 7.4: Classification AUC scores for the label-independent higher-order texton method for texture alone, density alone and the combination of texture and density according to the augmented method, the reselected method and the recalculated method. All the values have the same meaning as explained in Table 7.3.

n	texture alone	augmented	reselected	recalculated	density
1	0.723±0.052	-	0.723±0.052	0.588±0.083	0.740±0.032
	{15}	-	{15}	{12}	{d}
2	0.725±0.077	0.724±0.057	0.721±0.056	0.587±0.121	-
	{15,27}	{15,d}	{15,27}	{10,35}	-
3	0.718±0.067	0.723±0.055	0.734±0.0937	0.627±0.070	-
	{15,16,45}	{15,27,d}	{15,16,45}	{19,35,52}	-
4	0.716±0.070	0.729±0.088	0.706±0.058	0.598±0.070	-
	{9,15,27,	{15,16,45,	{9,15,27,	{26,35,36,	-
	41}	d}	41}	37}	-

The performance values shown were the average of K-fold cross validation results. For each fold, the features selected based on the full training data were used but since these are features were not necessarily optimal for any single fold, sequential feature selection may (and did) outperform exhaustive search.

Training AUC scores for the whole training data using exhaust search feature selection are always higher than that of using sequential feature selection (Appendix B.4). While testing AUC scores for the whole testing data using exhaust search feature selection are generally lower than that of using sequential feature selection. This indicates that comparing with sequential feature selection, exhaust feature selection has bigger possibility of overtraining.

Testing AUC scores for the whole testing data with hold-out validation (training on the 160 training images and testing on the 160 testing images) are generally lower than that of using 5-fold cross validation (training on 4 fold testing images and testing on the remaining 1 fold testing images). This might because straightforward cross validation involves more variability comparing with 5-fold cross validation. But k-fold cross validation tends to have higher bias (Hastie et al. [2008]).

Comparing the AUC scores of texture alone obtained with sequential feature selection (second columns of Tables 7.1 and 7.2) and exhaustive search feature selection (second columns of Tables 7.3 and 7.4), the best AUC scores obtained with sequential feature selection are better than those obtained with exhaustive search feature selection for both methods of generating higher-order textons. In consequence, the following conclusion will be drawn based on the results gained from sequential feature selection.

From the results in the second columns of Table 7.1 and Table 7.2, texture features calculated from the label-independent higher-order texton method perform slightly better than label-dependent higher-order texton method. This conclusion is opposite to the one made in Chapter 6 without feature selection and cross validation. The label-independent higher-order texton method is also more satisfying compared to label-dependent higher-order texton method because it completely avoids arithmetic operations on label values. Therefore, in the next chapter, only the label-independent higher-order texton method will be used for carrying out temporal risk assessment.

For the results in the second and third columns of Table 7.1 and Table 7.2, augmenting an optimal set of n texture features with density may reasonably be compared to the same set of features without density (previous row) or with the set of $n+1$ optimal texture features (same row). In both cases, no clear improvement is indicated by including density.

From the results in the second and fourth columns of Table 7.1 and Table 7.2, including density in the original pool and reselecting optimal features did not result

in density being selected in any of the six optimal combinations. These findings suggest that there may be substantial overlap of the information regarding risk between density and texture.

Results in the second and fifth column of Table 7.1 and Table 7.2, indicate that combining texture and density together before feature extraction resulted in lower AUC scores than texture alone, density alone or the former two methods of combining texture and density. This may be because that density and texture are not complementary in terms of risk.

The density feature used here is an adaptation of methods for computing volumetric density available from the literature. Performance based on this density feature alone compares favorably with results using density reported in the literature (Keller et al. [2012], AUC $= 0.70$) where the same criterion of risk was used as in this study. Thus the measure of density used here may be seen as representing the state of the art.

Risk classification based on four or more texture features performed at least as well as density alone (Tables 7.1 and 7.2). This indicates the role of texture alone in risk assessment is as important or more important than that of density alone. Combining density and texture does not improve risk assessment substantially over either density or texture alone. As the literature indicates that reliable volumetric estimates of density are problematic, texture may offer a preferable alternative.

Since higher-order textons (features with index > 20) appear in all combinations of n features for $n \geq 2$, this extension of basic textons is verified to be worthwhile. However, the full scope of possible texture features has not been explored and so further improvement may be possible.

Chapter 8

Temporal Risk Assessment

True risk assessment requires that an estimate is made of who will contract cancer at a future date. This chapter presents experiments on using textons to assess the risk of developing mammographically apparent breast cancer (the third risk criteria in Table 2.1) in two and four years (Section 8.3).

The overall processing strategies for extracting texture features with textons in this chapter are based on the results of the previous chapters. Thus local normalization (Section 3.2) is used prior to feature extraction; the 3×3 neighborhood method is used to construct feature vectors; only full CC view images are used; only the label-independent higher-order texton method for generating higher-order textons is used; classification is based on the Fisher method; and classification performance is measured according to the AUC score.

In order to develop a method for predicting future cancer based on mammograms, images from screening visits are needed from dates prior to the appearance of cancer. Since the DDSM data set used in the studies reported in previous chapters does not include temporal sequences of images, the experiments reported in this chapter use the BSSA data set (Section 1.5) instead. However, the BSSA data set is smaller than the DDSM data set and, more importantly, does not include BI-RADS class assignments. Since textons computed in previous chapters were based on distinct BI-RADS classes, some details of implementation of textons were re-evaluated in preparation for the temporal study.

Details of the structure of the BSSA data set for the temporal study are introduced in Section 8.1. Section 8.2 presents four preliminary studies to adjust the texton methods developed on the DDSM data set to the BSSA data set. From Chapter 6, first-order textons make the most significant contribution to classify high risk and low risk. Using first-order textons only, current breast cancer risk assessment with digital mammograms obtained satisfactory AUC scores (the third row of Table 6.7 (c) in Chapter 6). As a result, for the convenience of conducting preliminary experiments, the AUC score obtained from just first-order texton features will be used

as the criterion to choose the final experiment strategy for carrying out the temporal risk study. The first study examines the possibility of applying textons derived from the DDSM data set without change to the BSSA data set (Section 8.2.1). The results indicate that DDSM textons do not perform as well on the BSSA data set as on the DDSM data set. In the second study, new textons are trained on the BSSA data set (Section 8.2.2). Although performance based on BSSA specific textons was better than that of DDSM based textons, the level of performance was limited by the fact that separate textons could not be computed from different BI-RADS classes. In the third study, BI-RADS classes were assigned informally by the author (referred to as in-house BI-RADS classes) and new textons were trained using these in-house BI-RADS assignments (Section 8.2.3). Because the BSSA data set is of limited size (100 normal cases and 100 cancer cases), both breasts from cancer cases and normal cases were used in these studies. Thus breasts destined to develop cancer and the unaffected breast from the same woman were both included as high risk examples. However, this attempt failed to improve the performance. In anticipation that the actual risk of cancer for these two breasts might not be the same, a fourth study was conducted in which textons were trained separately on breasts destined to develop cancer and the contralateral breast (Section 8.2.4). In addition, in-house BI-RADS classes are used to insure wide mammographic appearance of training images. The strategy described in Section 8.2.4 is adopted to carry out the final temporal study on risk assessment in Section 8.3.

8.1 Data Set

In the final temporal study (Section 8.3), evidence of risk is tracked separately in the breast with cancer, referred to as the ipsilateral breast, and the breast without cancer, referred to as the contralateral breast. The BSSA data set comprises 900 CC images; for each of the three time periods (current, two year previous and four year previous), there are 100 images in each of the three experimental groups (ipsilateral high risk group, contralateral high risk group, and the low risk group) (Table 8.1).

For each experimental group, the 100 available images were divided into two sets of 50 images each; one set of the 50 was used for training and the remaining set was reserved for testing. The 50 cases for each group were selected so as to represent wide mammographic appearance. This was done by informally assigning images to BI-RADS classes by the author (referred to as in-house BI-RADS classes) and selecting approximately half the images from each in-house BI-RADS class for training. The BSSA data set contains fewer images than the DDSM data set and it is not possible to divide the data set so that every group has equal representation from all the BI-RADS classes. This limited the representation of mammographic

Table 8.1: Illustration of the structure of the BSSA data set. n is the number of images in each experimental group for every time period.

		2005, 2006	2003, 2004	2001, 2002
cancer cases	breast with cancer	current ipsilateral high risk $n = 100$	2 year previous ipsilateral high risk $n = 100$	4 year previous ipsilateral high risk $n = 100$
	breast without cancer	current contralateral high risk $n = 100$	2 year previous contralateral high risk $n = 100$	4 year previous contralateral high risk $n = 100$
normal cases	random breast	current low risk $n = 100$	2 year previous low risk $n = 100$	4 year previous low risk $n = 100$

appearance in the training images.

The preprocessing steps used for the BSSA data set are the same as in Chapters 4, 5, 6 and 7.

8.2 Preliminary Experiments

As the texture analysis method in this thesis was developed from the DDSM data set, images from the current high risk group and images from the current low risk group (Table 8.1) were used to conduct preliminary experiments to determine the most suitable experiment strategy for carrying out the final temporal risk assessment study. Thus a similar criterion of the surrogate of true risk is used in these experiments as was used in Chapters 5, 6 and 7, the only difference being the use of the BSSA data set instead of DDSM and, in some cases, using both the ipsilateral and contralateral images as described below. In addition, since the performance of first-order texton features is found to be better than higher-order texton features, only texture features calculated from first-order textons will be used to decide the final strategy for the temporal risk assessment experiment.

8.2.1 DDSM Textons Applied to BSSA Data

Satisfactory risk classification performance was achieved from texton features computed from images from the DDSM database (Chapters 5, 6 and 7), thus the first-order texton dictionary learnt from DDSM data set was used to calculate texture features for the current contralateral high risk and current low risk groups from the

BSSA data set (Table 8.1). Details of this preliminary experiment are described below.

Local normalization with $r = 22$, the 3×3 neighborhood, and sub-sampling with factor $5 \times 5 \rightarrow 1$ were applied to every image based on the explanation reported in Chapter 6.

Next, the first-order texton dictionary learnt from the DDSM data set was used to construct a texton map for each image. Pixels outside the breast boundary were found according to the image template and labeled with 0. Final texture features were the texton frequencies from 1 to 20 excluding the texton index 0.

In the final classification, the Fisher classifier trained by training images of the DDSM data set was used to classify the 100 current contralateral high and 100 current low risk images. Final classification performance was validated by 5-fold cross validation.

The classification performance for training was AUC $= 0.892$ and AUC $= 0.516$ for testing. AUC scores with 5-fold cross validation are shown in Table 8.2. From these AUC scores, it is obvious that the textons learnt from the DDSM data set are not suitable for the BSSA data set. The reason for this might be that the subsampling population used for mammogram collection is very different between these two data sets, or because the technique and scanners used to digitize these two databases are very different.

8.2.2 BSSA Textons without BI-RADS Assignments

The experiment in Section 8.2.1 showed that textons trained on the DDSM data set are not useful for determining risk for images in the BSSA data set. Accordingly, a new set of textons was computed based on the BSSA data. Since the BSSA data do not include BI-RADS scores, separate textons could not be derived for the individual four BI-RADS groups. Instead, K-means clustering with $K = 20$ was used to simply determine 20 textons for the entire set of training images. Otherwise, the data set, the preprocessing steps, methods for generating feature vectors, etc., were all as described in Section 8.2.1. AUC scores obtained from 5-fold cross validation was used to determine the performance (Table 8.3).

From the results above, even through the classification performance was improved to some extent, the performance was low compared to results in previous chapters (Chapters 5 - 7) and reliable conclusions cannot be drawn from these results. Hence simply computing 20 textons over all the images is not a suitable way to determine risk.

Table 8.2: Testing AUC scores for DDSM textons applied to BSSA data using 5-fold cross validation.

5-fold cross validation	training AUC	testing AUC
fold 1	0.751	0.459
fold 2	0.729	0.676
fold 3	0.751	0.558
fold 4	0.751	0.626
fold 5	0.740	0.646
average	0.744	0.593

Table 8.3: Testing AUC scores for BSSA textons without BI-RADS assignments using 5-fold cross validation.

5-fold cross validation	training AUC	testing AUC
fold 1	0.597	0.64
fold 2	0.727	0.395
fold 3	0.708	0.57
fold 4	0.642	0.765
fold 5	0.668	0.665
average	0.668	0.607

8.2.3 BSSA Textons with BI-RADS Assignments

From the results of Section 8.2.2, even through textons were learnt from the BSSA data set, the final risk classification was only slightly better than the results in Section 8.2.1. The difference between texton generation using the DDSM data set (Chapters 5, 6 and 7) and the BSSA data set is the availability of BI-RADS scores. For the DDSM data set, professional BI-RADS scores are available for every image while there is no professionally assigned BI-RADS scores for the BSSA data set. This might be the reason why the textons learnt from the BSSA data set did not improve risk classification performance. In addition, the BSSA data set is not as big as the DDSM data set. To overcome these problems, the author subjectively assigned unofficial BI-RADS scores (in-house BI-RADS scores) based on mammographic appearance to the images in the BSSA data set. In order to increase the number of images, the 100 current ipsilateral high risk and 100 current contralateral high risk groups (Table 8.1) were merged into a single high risk group. The 200 current low risk group consisting of both CC view breast images of the 100 normal cases was taken as the low risk group in this preliminary experiment.

These 400 images were divided into three groups - training, validation and testing. This is illustrated by Table 8.4.

As in Section 8.2.1, local normalization, the 3×3 neighborhood method and sub-

Table 8.4: Illustration of 100 current ipsilateral high risk images, 100 current contralateral high risk images and 100 current low risk images from the BSSA data set plus 100 contralateral low risk images from the original mammogram data set obtained from BreastScreen SA used in Section 8.2.3. "Tr" denotes the training group, "V" denotes the validation group and "Te" denotes the testing group.

	In-house BI-RADS classes	number of low risk breast images		number of high risk breast images	
Tr	I	right	3	ipsilateral	2
		left	2	contralateral	1
	II	right	7	ipsilateral	13
		left	4	contralateral	4
	III	right	24	ipsilateral	30
		left	14	contralateral	10
	IV	right	16	ipsilateral	15
		left	10	contralateral	5
V	I	right	3	ipsilateral	2
		left	1	contralateral	1
	II	right	6	ipsilateral	8
		left	1	contralateral	4
	III	right	24	ipsilateral	20
		left	5	contralateral	10
	IV	right	17	ipsilateral	10
		left	3	contralateral	5
Te	I	right	0	ipsilateral	0
		left	4	contralateral	2
	II	right	0	ipsilateral	0
		left	8	contralateral	13
	III	right	0	ipsilateral	0
		left	29	contralateral	30
	IV	right	0	ipsilateral	0
		left	19	contralateral	15

Table 8.5: Testing AUC scores for BSSA textons with BI-RADS assignments using 3-fold cross validation.

3-fold cross validation	training AUC	testing AUC
fold 1	0.61	0.714
fold 2	0.784	0.454
fold 3	0.772	0.500
average	0.722	0.556

sampling by $5 \times 5 \rightarrow 1$ were applied to each image. Then all training images of each in-house BI-RADS class from low and high risk groups were aggregated together and used to generate 5 textons, resulting in a total of 20 textons. The process for generating first-order texton is the same as for the DDSM data set used in Chapters 5, 6 and 7. The texton map for each image was constructed as described previously (Section 8.2.1) and the final set of texture features for each image was the set of texton frequencies from 1 to 20.

Exhaustive search feature selection (Section 2.8.1) was applied to evaluation images with the Fisher classifier trained by training images to select optimal features from the total of 20 features. The optimal features selected by this process (features with indices 3, 5, 6, 10, 19, 20) were used to classify images in the validation group. Final performance was represented by the average of 3-fold cross validation AUC scores. Classification results are shown in Table 8.5.

The results shown in Table 8.5 are not better than the results of the previous two preliminary experiments (Section 8.2.1 and Section 8.2.2). The application of BI-RADS scores, expanding BSSA data set and using fewer folds in cross validation did not improve risk assessment. There are a number of possible reasons why better results were not obtained. First, the signature of risk of developing cancer in the ipsilateral breast (the breast with known cancer) is likely to be different from that in the contralateral breast. Hence mixing these groups may have distorted the texture signal associated with risk. Second, the in-house BI-RADS classes may not have been accurate enough to clarify the characteristics of texture of different mammographic appearance. Third, given the small size of the data, separating the images into separate training, evaluation and validation groups reduced the number of images in each group to unrepresentative low numbers and dictated the use of 3-fold cross validation instead of the more widely accepted 5-fold cross validation.

8.2.4 Separating Ipsilateral and Contralateral Breasts

A final preliminary experiment was conducted to take into account the results and discussions in Section 8.2.3. Thus, in this experiment, risk of breast cancer was estimated separately for ipsilateral and contralateral high risk images. Instead of

Table 8.6: Testing AUC scores for separating ipsilateral and contralateral breasts without 5-fold cross validation.

ipsilateral high risk vs low risk		contralateral high risk vs low risk	
training AUC	testing AUC	training AUC	testing AUC
0.956	0.704	0.905	0.573

constructing separate textons based on different in-house BI-RADS classes, separate textons were constructed based on different risk classes (high and low). The in-house BI-RADS assignments were used to enforce diversity of mammographic appearance in the training and testing groups as described in Section 8.1, but the distribution of in-house BI-RADS images was not the same for each group. Compared with Section 8.2.3, the evaluation stage was removed allowing the available images to be divided among training and testing sets only. This increased the number of images in the testing set sufficiently to allow 5-fold cross validation.

As in Section 8.2.1, local normalization, the 3×3 neighborhood and sub-sampling by factor $5 \times 5 \rightarrow 1$ were applied to each image of the 100 current contralateral high risk images and 100 current low risk images. Next, textons were generated separately from high and low risk training images. Feature vectors from all high risk training images composed of four in-house BI-RADS class images were aggregated into a single array and 10 textons were generated from this array by applying K-means clustering with $K = 10$. Similarly, 10 textons were generated from the array of low risk feature vector collection with K-means clustering. Thus, the final texton dictionary consisted of 20 textons.

Texton maps were constructed in the same process as in the above three preliminary experiments. Final texture features were the texton frequencies from 1 to 20 excluding the outside breast region, texton 0.

In the last classification step, the Fisher classifier was trained on the training images and then used to classify testing images into high or low risk groups. Final performance was validated by 5-fold cross validation using AUC score as the performance criterion.

The whole process was repeated on 100 current ipsilateral high risk images and 100 current low risk images. Results for separating ipsilateral and contralateral breasts are shown in Tables 8.6, 8.7 and 8.8.

In comparison with the results of the previous three preliminary experiments in Sections 8.2.1, 8.2.2 and 8.2.3, results of this section are much more encouraging. Thus, in next section, the final experiment on temporal risk assessment will be carried out with the strategy proposed in this section.

Table 8.7: Testing AUC scores for ipsilateral high risk vs low risk for separating ipsilateral and contralateral breasts using 5-fold cross validation.

5-fold cross validation	training AUC	testing AUC
fold 1	0.881	0.771
fold 2	0.870	0.819
fold 3	0.882	0.747
fold 4	0.878	0.836
fold 5	0.859	0.819
average	0.874	0.798

Table 8.8: Testing AUC scores for contralateral high risk vs low risk for separating ipsilateral and contralateral breasts using 5-fold cross validation.

5-fold cross validation	training AUC	testing AUC
fold 1	0.575	0.658
fold 2	0.807	0.745
fold 3	0.807	0.698
fold 4	0.815	0.661
fold 5	0.808	0.681
average	0.762	0.689

8.3 Final Experiment on Temporal Risk Assessment

The degree to which image texture anomalies correlate to breast cancer risk is not known. In addition, little is known about changes in texture prior to cancers becoming discernible mammographically. Neither is it clear if changes in texture are restricted to the breast in which cancer will eventually appear or if the changes are bilateral.

In this section, a temporal study is presented on estimating risk of breast cancer based on texture independent of breast density. Estimates of risk are obtained separately for ipsilateral and contralateral breasts.

The experiment on temporal risk assessment will be carried out according to the strategy proposed in Section 8.2.4. From the conclusions in Chapter 7, texture features will be calculated from the label-independent higher-order textons. The final performance of temporal mammogram images will be validated with sequential feature selection, exhaustive search feature selection and 5-fold cross validation.

8.3.1 Methods

Twenty first-order textons, first-order texton maps and first-order texton representations were obtained with the same process described above in the last preliminary experiment in Section 8.2.4.

The algorithm for higher-order texton generation by the label-independent method described previously in Section 6.4 was used to generate second- and third-order textons for texture feature extraction. Details are described below:

Feature vectors from 50 training first-order texton maps of the current ipsilateral high risk group and 50 training first-order texton maps of the current low risk group were used to determine a set of 10 second-order textons each. The resulting 20 second-order textons were used to generate second-order texton maps and the histogram of second-order textons. Similarly, 20 third-order textons, the third-order texton map and the third-order texton histogram for each image were computed in the same way as their second-order analogs but were based on the second-order texton maps. Together, each image in the current ipsilateral high risk or low risk group was represented by a combined texton histogram of 60 features for classification.

To determine the best feature combination for classifying current ipsilateral high risk and low risk mammograms, sequential backward feature selection was applied with the AUC score on training images as the optimization criterion. The maximum number of features was set at eight for sequential backward feature selection and four for exhaustive search feature selection. The feature set identified by this process was used to estimate classifier performance on unseen data by using 5-fold cross validation on the 100 testing images from the two groups. The average AUC score from the cross validation will be referred to as the current ipsilateral AUC score.

The steps described above were repeated for the 200 two year previous ipsilateral high risk and two year previous low risk group images and again for the 200 images from the four year previous period to obtain the previous two year ipsilateral AUC score and the previous four year ipsilateral AUC score.

Finally, the entire process was repeated with the contralateral high risk and low risk group images in current, two and four year previous periods.

All the parameters used in these procedures including the choice of 3×3 neighborhood instead of larger ones, the sub-sampling factor and the choice of $K = 10$ in K-means clustering were based on the experiment in Section 8.2.4.

8.3.2 Results for Sequential Feature Selection

Temporal risk assessment results obtained with sequential feature selection are shown in Figure 8.1. From Figure 8.1, the AUC scores for classifying ipsilateral high risk and low risk breast images were 0.749 for the current period, 0.674 for the two year previous period and 0.601 for the four year previous period. (Detailed results for distinguishing ipsilateral high risk from low risk breast images are shown in Tables B.37, B.39 and B.41 in Section B.5 of Appendix B.) Although there was a decrease in classification performance as a function of time prior to the detection of cancer,

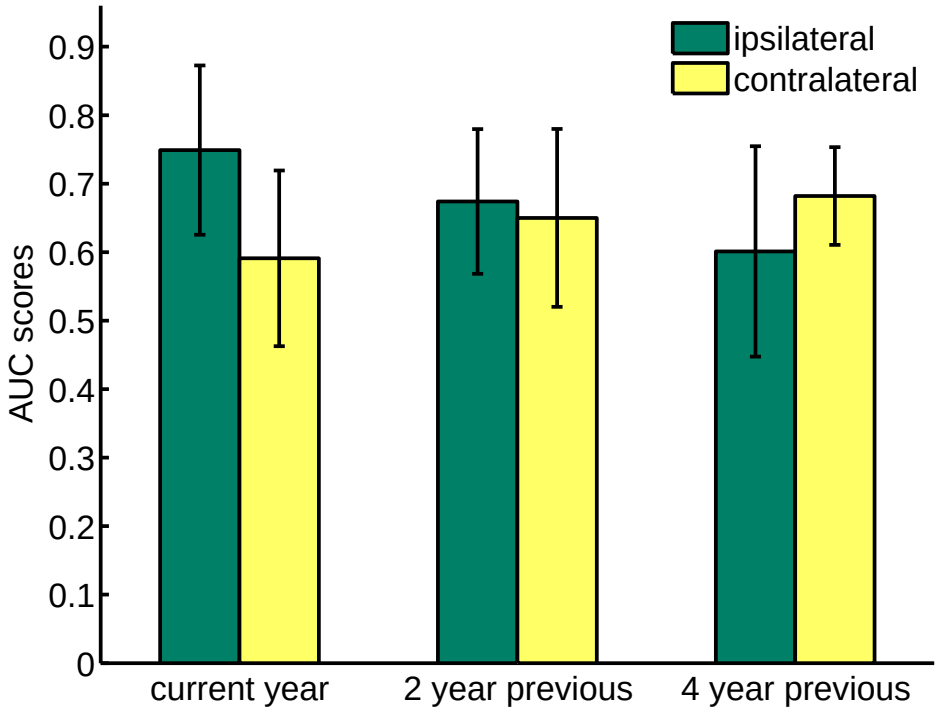

Figure 8.1: Temporal AUC scores for risk classification with label-independent higher-order textons and sequential backward feature selection. The three dark green bars are the AUC scores for the classification of ipsilateral high risk vs low risk in the current year, previous two years and previous four years; the three yellow bars are the AUC scores for the classification of contralateral high risk vs low risk in the current year, previous two years and previous four years. The error bars show one SD.

there was no significant difference between two year previous and current images ($p = 0.206$, $n = 5$) or between four year previous and current images ($p = 0.071$, $n = 5$). The prediction AUC scores for classifying contralateral high risk and low risk images were 0.591 for the current period, 0.650 for the two year previous period and 0.682 for the four year previous period. (Detailed results for classifying contralateral high risk and low risk breast images are shown in Tables B.38, B.40 and B.42 in Section B.5 of Appendix B.) Again there was no significant difference between two year previous and current images ($p = 0.536$, $n = 5$) or between four year previous and current images ($p = 0.286$, $n = 5$) although there was a slight increase in classification performance as a function of time prior to the detection of cancer.

In addition, there were no significant differences of the classification performance between ipsilateral high risk vs low risk and contralateral high risk vs low risk AUC scores for current ($p = 0.073$, $n = 5$), two year previous ($p = 0.795$, $n = 5$) and four year previous ($p = 0.114$, $n = 5$) periods.

8.3.3 Results for Exhaustive Search Feature Selection

Temporal risk assessment results obtained with exhaustive search feature selection are shown in Figure 8.2. From Figure 8.2, the AUC scores for classifying ipsilateral high risk and low risk breast images were 0.686 for the current period, 0.685 for the two year previous period and 0.627 for the four year previous period. (Detailed results for classifying ipsilateral high risk and low risk breast images are shown in Tables B.47, B.49 and B.51 in Section B.5 of Appendix B.) Although there was a decrease in classification performance as a function of time prior to the detection of cancer, there was no significant difference between two year previous and current images ($p = 0.147$, $n = 5$) or between four year previous and current images ($p = 0.140$, $n = 5$). The prediction AUC scores for classifying contralateral high risk and low risk images were 0.610 for the current period, 0.650 for the two year previous period and 0.659 for the four year previous period. (Detailed results for separating contralateral high risk from low risk breast images are shown in Tables B.48, B.50 and B.52 in Section B.5 of Appendix B.) Again there was no significant difference between two year previous and current images ($p = 0.720$, $n = 5$) or between four year previous and current images ($p = 0.591$, $n = 5$) although there was a slight increase in classification performance as a function of time prior to the detection of cancer.

In addition, there were no significant differences of classification performance between ipsilateral high risk vs low risk and contralateral high risk vs low risk AUC scores for current ($p = 0.487$, $n = 5$), two year previous ($p = 0.202$, $n = 5$) and four

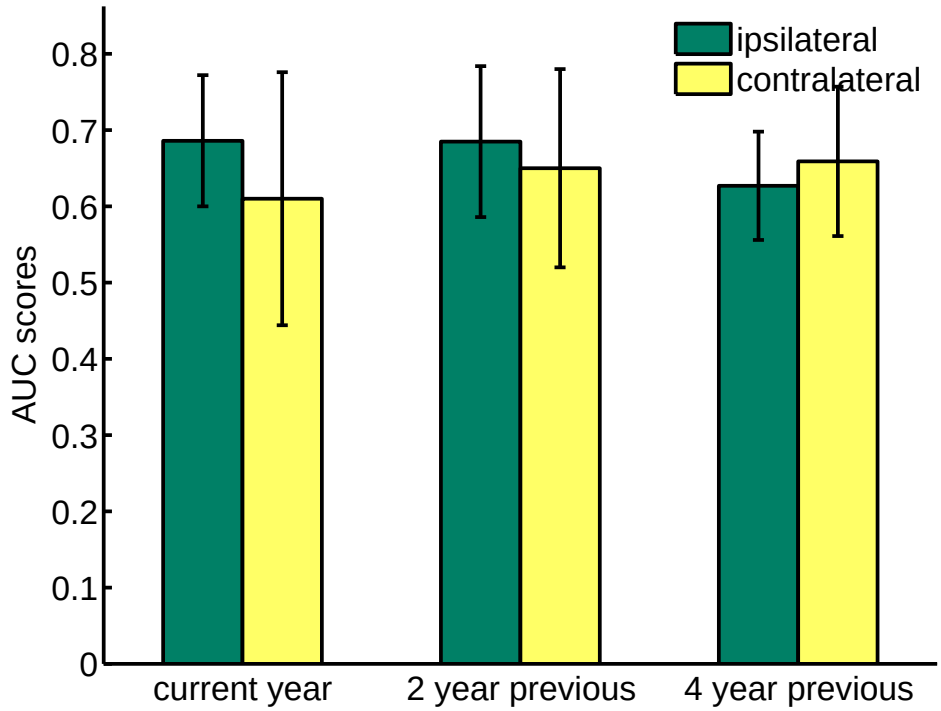

Figure 8.2: Temporal AUC scores for risk classification with label-independent higher-order textons and exhaustive search feature selection. Details of the representation are as in Figure 8.1.

year previous ($p = 0.364$, $n = 5$) periods.

8.3.4 Conclusion and Discussion

Although the author of thesis is not a certified radiologist, she has gained substantial experience in viewing mammograms. The word "informal" or "in-house" is used to indicate that these assignments were done by someone who does not have formal training. BI-RADS scores are not hard to assign in many cases. In difficult cases, the variation between trained radiologists and even for the same radiologist viewing the images at different times are such that an intelligent amateur can approach similar performance fairly quickly.

The BI-RADS scores assigned by the author (in-house BI-RADS scores) are used in a way that has extremely little influence on the final conclusion of the thesis.

In-house BI-RADS scores were used in Section 8.2.3 as surrogates for breast cancer risk in the same way as expert assigned BI-RADS scores were use in Chapter 4. However, this direction of inquiry was abandoned for a variety of reasons (last paragraph of Section 8.2.3) including our reluctance to rely on the in-house BI-RADS scores in a crucial role.

In-house BI-RADS scores were also used in Sections 8.2.3 and all parts of Section 8.3 which reports a temporal analysis of risk. In these sections, the in-house BI-RADS scores were not used as a surrogate for risk. The surrogates for risk considered here were the fact that the women was found to have cancer at the time of screening, the fact that the woman was found to have cancer 2 years later or the fact that the woman was found to have cancer four years later (plus additional versions that distinguish between cancer found in one breast or the other). In-house BI-RADS scores were only used to spread the diversity of mammographic appearance somewhat equally between training and testing images. Thus within the high risk group (determined by the presence of cancer at screening) the in-house BI-RADS class I images were divided between the training and testing sets randomly and as equally in number as possible. This was repeated for the other three in-house BI-RADS classes and similarly for the low risk group. Thus the in-house BI-RADS scores do not contribute the surrogate for risk in these sections. If there had been substantial error in the assignment of BI-RADS scores, this would have served to introduce more variation between the training and testing sets resulting in poorer prediction error. Since the results obtained match, or exceed, previously published results (where comparable), any negative effects of BI-RADS assignments must have been very small.

Finally, if a researcher wishes to reproduce the results with the same set of images, the informal assignments of BI-RADS classes can be made available.

Similar to the discussion made in Section 7.6, globally, exhaustive search feature selection could find the maximum performance for training images. However, in this chapter, training results for exhaustive search feature selection are lower than sequential feature selection (Appendix B.5). This is because feature selection was optimized over the whole training set. On the full training date, sequential feature selection was not better than exhaustive search (Appendix B.6). The performance values shown were the average of K-fold cross validation results (Appendix B.5). For each fold, the features selected based on the full training data were used but since these are features were not necessarily optimal for any single fold, sequential feature selection may (and did) outperform exhaustive search.

Because the focus was on testing the "in principle" information content relevant to breast cancer risk, separate textons and optimal feature sets were found for each time period (current, two year previous and four year previous). This is not a practical result for assessing risk clinically since a separate test would be required for estimating risk for different times in the future. Interpreting results would be difficult if, for example, a woman was found to have high risk of developing cancer in two years time, but low risk of developing cancer in four years time.

A more practical result would be a single test to estimate risk. This was con-

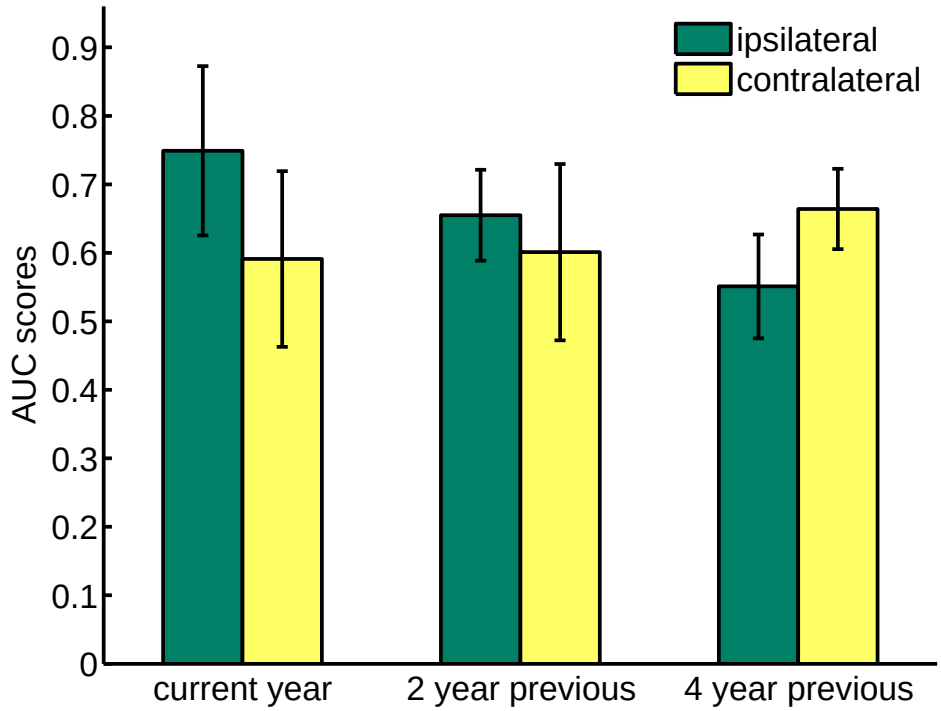

Figure 8.3: Temporal AUC scores for risk classification with label-independent higher-order textons learnt from the current year period and sequential backward feature selection. Details of the representation are as in Figure 8.1.

sidered by running a minor variation of the experiments described in Section 8.3.1. In this variation, the higher-order textons learnt from the label-independent method and optimal features found using the current images were computed and tested on the images in the two year previous period and the four year previous period. (Detailed results for classifying high risk and low risk breast images with sequential feature selection are shown in Tables B.43, B.44, B.45 and B.46 in Section B.5 of Appendix B; detailed results for distinguishing high risk from low risk breast images with exhaustive search feature selection are shown in Tables B.53, B.54, B.55 and B.56 in Section B.5 of Appendix B.) The results obtained with sequential feature selection are not substantially different from the original experiment (($p = 0.056$, $n = 6$), Figure 8.3). Similarly, the results obtained with exhaustive search feature selection are not substantially different from the original experiment (($p = 0.144$, $n = 6$), Figure 8.4). Accordingly, the methods described here could, in principle, contribute to a clinically useful scheme for estimating breast cancer risk. The following discussion and conclusion will be made based on results presented in Figures 8.1 and 8.2.

Temporal AUC scores obtained with sequential feature selection in Figure 8.1 and temporal AUC scores obtained with exhaustive search feature selection in Figure 8.2 are not significantly different ($p = 0.727$, $n = 6$, mean difference is 0.005). In consequence, the following discussion and conclusion will be made according to AUC scores in Figure 8.1 obtained from sequential feature selection since they

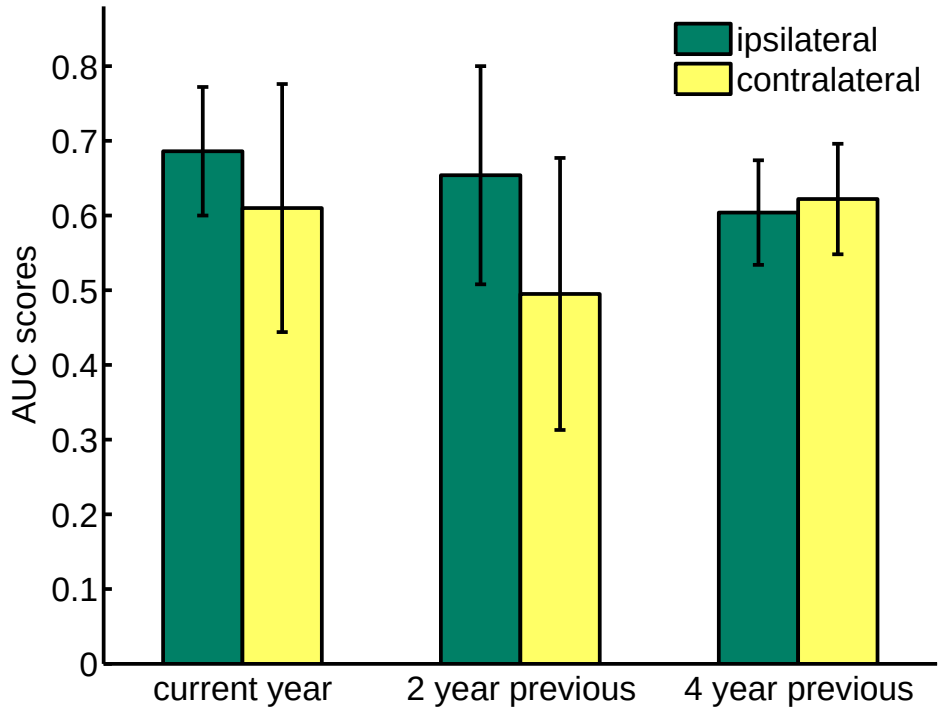

Figure 8.4: Temporal AUC scores for risk classification with label-independent higher-order textons learnt from the current year period and exhaustive search feature selection. Details of the representation are as in Figure 8.1.

are slightly better than those in Figure 8.2 obtained from exhaustive search feature selection.

From Figure 8.1, there is an increasing trend of the mean risk classification performance for classifying ipsilateral high risk and low risk images in the three time periods from the four year previous to the current year. For the breast destined to have cancer in the future, the risk classification AUC score gets better as time goes on. This indicates a positive breast cancer risk prediction ability of texture analysis based on higher-order textons. Meanwhile, a decreasing trend of the mean risk classification performance for classifying contralateral high risk and low risk images is presented from the four year previous period to the two year previous period and to the current year period. It seems that the breast contralateral to the breast destined to have cancer is not associated with increasing risk over time before the period that cancer is detected. In addition, the risk of this breast is lower than the contralateral breast destined to develop cancer (even through not significantly lower).

The AUC scores for the different groups and times (Figure 8.1) indicate some trend, but none of the differences are statistically significant. Hence texture information relevant to breast cancer risk seems not to be restricted to, or significantly stronger in, the breast destined to develop cancer. Similarly, texture information relevant to breast cancer risk is present at least four years previous to the emergence of mammographically apparent signs of cancer at levels not significantly different

from those at the time cancer is detected.

The results reported here use only texture information and were not combined with other mammographic information such as density or clinical information to provide a comprehensive estimate of breast cancer risk. Thus, it is not yet clear if texture provides information that is complementary to other measures or if it largely reproduces existing measures. This will be the objective for a future study.

This work used only digitized film images. This was necessary since the temporal structure of the experiment required that images from as long ago as 2001 were needed, well before full-field digital mammogram images were available in South Australia. This study indicates that the temporal results reported here could represent conservative estimates of risk classification if the method is applied to full-field digital mammograms.

Chapter 9

Final Remarks

The contributions of this thesis are in two areas, image analysis and breast cancer risk. The contribution to image analysis is the introduction of a general protocol for computing higher-order textons (Chapter 6). Standard textons, viewed here as first-order textons, allow a global characterization of local patterns. Second-order textons allow global characterization of patterns of patterns and third-order textons may be viewed as patterns of patterns of patterns, and so forth. The original notion of textons introduced by Julesz in 1981 centered on human perception of patterns. In particular, Julesz asked if humans were able to perceive differences in iso-second-order texture, meaning differences in images having the same first- and second-order statistics. In this thesis, local background and local variance are removed from images prior to analysis (Chapter 3). So in one sense, the images considered are all iso-second-order. However, this is quite a narrow view of Julesz's notion of iso-second-order since local mean and variance play only a small role in pattern recognition by humans and machines alike. The idea of higher-order textons formalized in this thesis (Chapter 6) is closer to the spirit of Julesz's notion. An obvious difference is that Julesz was concerned with human perception and this thesis is about machine recognition of patterns. Nevertheless, the example in Section 6.3.4 demonstrates that higher-order textons are able to quantify differences between iso-second-order images, where iso-second-order is taken to mean that images have the same first- and second-order texton histograms. By simple extension of the example, $n+1$-order textons can separate iso-nth-order images.

The main objective of this thesis is to study the contribution of texture in screening mammograms for assessing breast cancer risk. Higher-order textons emerged as an important part of this effort. Given the success of first-order textons reported in the literature (Varma and Zisserman [2005, 2003]) coupled with the improved performance reported here in the case of breast cancer risk assessment, higher-order textons may well play a significant role in texture analysis applied to other areas such as computer vision, object detection, tracking and image understanding. Even

though higher-order textons obtained satisfactory performance in risk classification, for further improvement of the performance, methods for computing other texture measures or combining texture measures with other risk factors (density etc.) should be explored. The method for generating higher-order textons presented in this thesis are not the only ones. Other methods should be explored for generating higher-order textons although methods should be restricted to the ones that do not involve arithmetic of texton labels as stipulated in Chapter 6.

The contribution to breast cancer risk assessment is estimating risk based on texture alone. The two main findings are that texture alone provides estimates of risk as good or better than density alone and that positive predictive value of risk based on texture is possible at least four year prior to the onset of cancer. In addition, a number of observations were made that may be important on their own. First, the signal of risk seems to be stronger in the breast destined to develop cancer than in the other breast. Second, the texture analysis methods considered in this thesis perform better on CC view images than MLO view images. Third, texture measured over the whole breast outperforms texture measured on local regions of the breast only.

Each of these observations have possibly important ramifications and are seeds for future work. First, whether increased risk is a characteristic of the women or of each breast separately is not known. This thesis does not settle this issue entirely, but findings in Chapter 8 indicate that while the texture signal is stronger in the breast destined to develop cancer, there is some signal in the unaffected contralateral breast too. This suggests that studies linking texture and biological changes in the breast might lead to better insight into conditions that increase risk. These results also suggest that bilateral differences in texture may be of benefit to computer-aided breast cancer detection. Second, the fact that risk assessment based on texture from CC view images outperforms risk assessment from MLO views may also be important to computer-aided breast cancer detection. The results of the thesis suggest that algorithms for computer-aided breast cancer detection might improve if tuned separately for MLO and CC views. Third, the fact that textures measured over the full breast perform better than texture measured on local regions complements the first observation above in that the texture signal is present over the full breast and does not seem to be restricted to the location of the future cancer. Future temporal studies on comparing texture at specific locations of cancer in mammograms could shed more light on these issues. For the time being, this observation also encourages implementation of risk assessment based on texture as complicate devices for consistently identifying the region of focused texture information are not necessary. Implementing texture based risk assessment based on textons and higher-order textons is fairly straightforward.

Appendix A

Feature Indexing

Table A.1: DDSM data set feature indexing for texture features calculated from higher-order textons generated with the label-dependent and label-independent methods. Table entries are indices to features described in Chapters 6 and 7.

	BI-RADS I	BI-RADS II	BI-RADS III	BI-RADS IV
1st-order	1, 2, 3, 4, 5	6, 7, 8, 9, 10	11, 12, 13, 14, 15	16, 17, 18, 19, 20
2nd-order	21, 22, 23, 24, 25	26, 27, 28, 29, 30	31, 32, 33, 34, 35	36, 37, 38, 39, 40
3rd-order	41, 42, 43, 44, 45	46, 47, 48, 49, 50	51, 52, 53, 54, 55	56, 57, 58, 59, 60
density	d			

Table A.2: BSSA data set feature indexing for texture features calculated from higher-order textons generated with the label-independent method. Table entries are indices to features described in Chapter 8.

	low risk	high risk
1st-order	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	11, 12, 13, 14, 15, 16, 17, 18, 19, 20
2nd-order	21, 22, 23, 24, 25, 26, 27, 28, 29, 30	31, 32, 33, 34, 35, 36, 37, 38, 39, 40
3rd-order	41, 42, 43, 44, 45, 46, 47, 48, 49, 50	51, 52, 53, 54, 55, 56, 57, 58, 59, 60

Appendix B

Supplementary Experimental Results

B.1 Supplementary Results for Region Dependent Risk Assessment

Tables B.1 and B.2 show the classification results for texton features generated from 5×5 and 7×7 neighborhoods, respectively (described in Section 5.3.1). Part (a) of these two tables show the testing accuracies of classifying high risk images from low risk images using an ensemble k-nearest neighbor classifiers. For the ensemble k-nearest neighbor classifier with 5×5 neighborhoods, the values of three parameters are $k = 5$ for the number of nearest neighbors, 18 for the number of predictors and 3 for the number of learners. For ensemble k-nearest neighbor classifier with 7×7 neighborhoods, the values of three parameters are $k = 7$ for the number of nearest neighbors, 9 for the number of predictors and 5 for the number of learners. Details of how these parameters were chosen are discussed in Section 2.5.1 of Chapter 2. Part (b) of these two tables show the testing accuracies of distinguishing high risk images from low risk images using a SVM classifier with linear kernel. As the number of testing images is the same for the high- and the low-risk class, the total accuracy listed in parts (a) and (b) of these two tables is simply the average of the individual accuracies for each region. Part (c) of these two tables show the classification AUC scores for separating high risk images from low risk images using a Fisher classifier (the AUC score for both training and testing sets are recorded). The "rank" column in each part orders the performance for each region from "1", the highest, to "6", the lowest so that the relative importance of the texture information for risk assessment in each region can be easily assessed.

Table B.1: Classification performance for the 5×5 neighborhood method with different classifiers; ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier.

(a) 5×5 neighborhoods with ensemble k-nearest neighbor

region	low risk	high risk	total accuracy	rank
Ω_1	0.725	0.588	0.656	2
Ω_2	0.563	0.500	0.531	6
Ω_3	0.563	0.513	0.538	5
Ω_4	0.613	0.513	0.563	3
Ω_5	0.588	0.525	0.556	4
Ω_6	0.725	0.600	0.663	1

(b) 5×5 neighborhoods with SVM classifier

region	low risk	high risk	total accuracy	rank
Ω_1	0.588	0.588	0.588	4
Ω_2	0.488	0.513	0.500	6
Ω_3	0.538	0.600	0.569	5
Ω_4	0.425	0.763	0.594	3
Ω_5	0.638	0.625	0.631	2
Ω_6	0.688	0.625	0.656	1

(c) 5×5 neighborhoods with Fisher classifier

region	training AUC	testing AUC	rank
Ω_1	0.630	0.597	4
Ω_2	0.679	0.517	6
Ω_3	0.730	0.523	5
Ω_4	0.802	0.612	3
Ω_5	0.831	0.625	2
Ω_6	0.903	0.648	1

Table B.2: Classification performance for the 7×7 neighborhood method with different classifiers; ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier.

(a) 7×7 neighborhoods with ensemble k-nearest neighbor

region	low risk	high risk	total accuracy	rank
Ω_1	0.725	0.538	0.631	2
Ω_2	0.575	0.525	0.550	6
Ω_3	0.700	0.463	0.581	4
Ω_4	0.588	0.588	0.588	3
Ω_5	0.588	0.538	0.563	5
Ω_6	0.713	0.588	0.650	1

(b) 7×7 neighborhoods with SVM classifier

region	low risk	high risk	total accuracy	rank
Ω_1	0.625	0.450	0.538	4
Ω_2	0.563	0.325	0.444	6
Ω_3	0.600	0.550	0.575	3
Ω_4	0.550	0.713	0.631	2
Ω_5	0.575	0.363	0.469	5
Ω_6	0.613	0.688	0.650	1

(c) 7×7 neighborhoods with Fisher classifier

region	training AUC	testing AUC	rank
Ω_1	0.756	0.567	3
Ω_2	0.695	0.423	6
Ω_3	0.714	0.560	4
Ω_4	0.808	0.661	1
Ω_5	0.709	0.425	5
Ω_6	0.868	0.634	2

B.2 Supplementary Results for Higher-order Textons

Tables B.3 and B.4 show the classification results for higher-order texton features generated from 5×5 and 7×7 neighborhoods, respectively (described in Section 6.3.5). Part (a) of these two tables show the testing accuracies for classifying high risk images from low risk images using an ensemble k-nearest neighbor classifier. For the ensemble k-nearest neighbor classifier with 5×5 neighborhoods, the values of three parameters are $k = 5$ for the number of nearest neighbors, 12 for the number of predictors and 10 for the number of learners. For the ensemble k-nearest neighbor classifier with 7×7 neighborhoods, the values of three parameters are $k = 8$ for the number of nearest neighbors, 16 for the number of predictors and 8 for the number of learners. Details of how these parameters were chosen are discussed in Section 2.5.1 of Chapter 2. Part (b) of these two tables show the testing accuracies for distinguishing high risk images from low risk images using a SVM classifier with linear kernel. As the number of testing images is the same for both high- and low-risk class, the total accuracy listed in parts (a) and (b) of these two tables is simply the average of the individual accuracies of each feature combination. Part (c) of these two tables show the classification AUC scores for separating high risk images from low risk images using a Fisher classifier (the AUC score for both training and testing sets are recorded). The "rank" column in each part orders the performance for each feature combination from "1", the highest, to "7", the lowest so that the efficiency of the texture information for risk assessment in each feature combination can be easily assessed.

Table B.3: Classification performance for higher-order textons using the 5×5 method with different classifiers; ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier. The rank for (a) and (b) is based on the total accuracy. The rank for (c) is based on the testing AUC score.

(a) 5×5 method with ensemble k-nearest neighbor classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.713	0.575	0.644	2
2nd	0.513	0.713	0.613	4
3rd	0.350	0.650	0.500	6
1st & 2nd	0.663	0.613	0.638	3
1st & 3rd	0.675	0.638	0.657	1
2nd & 3rd	0.400	0.725	0.563	5
1st & 2nd & 3rd	0.535	0.75	0.644	2

(b) 5×5 method with SVM classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.688	0.625	0.656	1
2nd	0.675	0.413	0.544	4
3rd	0.550	0.450	0.500	5
1st & 2nd	0.738	0.525	0.631	3
1st & 3rd	0.725	0.575	0.650	2
2nd & 3rd	0.625	0.463	0.544	4
1st & 2nd & 3rd	0.750	0.550	0.650	2

(c) 5×5 method with Fisher classifier

texton order	training AUC	testing AUC	rank
1st	0.903	0.648	3
2nd	0.765	0.558	6
3rd	0.654	0.532	7
1st & 2nd	0.824	0.637	4
1st & 3rd	0.893	0.677	2
2nd & 3rd	0.827	0.570	5
1st & 2nd & 3rd	0.953	0.679	1

Table B.4: Classification performance for higher-order textons using the 7×7 method with different classifiers; ensemble k-nearest neighbor classifier, SVM classifier and Fisher classifier. The rank for (a) and (b) is based on the total accuracy. The rank for (c) is based on the testing AUC score.

(a) 7×7 method with ensemble k-nearest neighbor classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.688	0.563	0.625	2
2nd	0.475	0.625	0.550	6
3rd	0.550	0.625	0.588	5
1st & 2nd	0.625	0.600	0.613	3
1st & 3rd	0.650	0.563	0.606	4
2nd & 3rd	0.550	0.625	0.588	5
1st & 2nd & 3rd	0.585	0.690	0.638	1

(b) 7×7 method with SVM classifier

texton order	low risk	high risk	total accuracy	rank
1st	0.613	0.688	0.650	1
2nd	0.563	0.475	0.519	7
3rd	0.588	0.488	0.538	6
1st & 2nd	0.675	0.550	0.613	2
1st & 3rd	0.588	0.588	0.588	4
2nd & 3rd	0.600	0.563	0.581	5
1st & 2nd & 3rd	0.650	0.550	0.600	3

(c) 7×7 method with Fisher classifier

texton order	training AUC	testing AUC	rank
1st	0.868	0.634	1
2nd	0.758	0.563	6
3rd	0.824	0.536	7
1st & 2nd	0.915	0.582	5
1st & 3rd	0.941	0.611	2
2nd & 3rd	0.762	0.595	3
1st & 2nd & 3rd	0.952	0.590	4

B.3 Detailed Results for Risk Classification of Texture vs Density - Part I

Feature indexing for Tables B.5 to B.20 in this section is described in Table A.1 of Appendix A. These tables (Table B.5 to Table B.20) show detailed results for the experiment of texture versus density described in Section 7.4. Results reported here were obtained by selecting features using the whole 160 training image data set and classification performance for the 160 testing image data set was calculated with 5-fold cross validation.

Table B.5: Detailed risk classification AUC scores for 60 texton features calculated from label-dependent higher-order textons using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.7234±0.052	{15}
2	0.721±0.018	0.7186±0.078	{11, 15}
3	0.768±0.011	0.7700±0.059	{9, 11, 15}
4	0.769±0.013	0.7714±0.056	{9, 11, 15, 49}
5	0.768±0.013	0.7652±0.057	{9, 11, 15, 49 56}
6	0.769±0.010	0.7492±0.058	{9, 11, 15, 49, 50, 56}

Table B.6: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.718±0.021	0.722±0.084	{15, 60}
3	0.718±0.022	0.718±0.084	{12, 15, 60}
4	0.761±0.022	0.751±0.094	{12, 15, 25, 60}
5	0.768±0.022	0.756±0.091	{4, 12, 15, 25, 60}
6	0.805±0.020	0.775±0.083	{4, 12, 15, 25, 27, 60}

Table B.7: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the augmented feature set method using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.722±0.017	0.724±0.057	{15, 61}
3	0.726±0.024	0.725±0.074	{15, 30, 61}
4	0.775±0.016	0.759±0.059	{15, 30, 42, 61}
5	0.772±0.011	0.762±0.051	{15, 28, 30, 42, 61}
6	0.781±0.013	0.773±0.054	{2, 15, 28, 30, 42, 61}

Table B.8: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the augmented feature set method using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.722±0.017	0.724±0.057	{15, 61}
3	0.727±0.025	0.727±0.083	{15, 60, 61}
4	0.726±0.028	0.723±0.086	{12, 15, 60, 61}
5	0.766±0.026	0.746±0.093	{12, 15, 25, 60, 61}
6	0.7698±0.022	0.754±0.091	{4, 12, 15, 35, 60, 61}

Table B.9: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the reselected feature set method using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.721±0.018	0.719±0.078	{11, 15}
3	0.769±0.011	0.770±0.059	{9, 11, 15}
4	0.769±0.013	0.771±0.056	{9, 11, 15, 49}
5	0.768±0.013	0.765±0.057	{9, 11, 15, 49, 56}
6	0.769±0.010	0.749±0.058	{9, 11, 15, 49, 50, 56}

Table B.10: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the reselected feature set method using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.721±0.020	0.719±0.080	{15, 45}
3	0.720±0.021	0.717±0.083	{12, 15, 45}
4	0.714±0.021	0.711±0.082	{12, 15, 34, 45}
5	0.750±0.022	0.719±0.081	{12, 15, 34, 40, 45}
6	0.754±0.023	0.715±0.081	{12, 15, 34, 40, 43, 45}

Table B.11: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-dependent higher-order textons through the recalculated feature set method using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.586±0.018	0.588±0.083	{31}
2	0.682±0.017	0.674±0.069	{13, 31}
3	0.680±0.020	0.664±0.071	{11, 13, 31}
4	0.746±0.015	0.725±0.058	{11, 13, 25, 31}
5	0.748±0.016	0.717±0.053	{11, 13, 25, 31, 42}
6	0.756±0.019	0.734±0.060	{11, 13, 16, 25, 31, 42}

Table B.12: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-independent higher-order textons through the recalculated feature set method using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.595±0.030	0.476±0.077	{35}
2	0.605±0.028	0.587±0.121	{10, 35}
3	0.695±0.026	0.658±0.129	{10, 17, 35}
4	0.724±0.022	0.688±0.105	{10, 17, 30, 35}
5	0.729±0.024	0.679±0.101	{10, 17, 30, 35, 43}
6	0.731±0.022	0.684±0.094	{10, 17, 23, 30, 35, 43}

Table B.13: Detailed risk classification AUC scores for 60 texton features calculated from label-dependent higher-order textons using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.722±0.016	0.723±0.055	{13, 15}
3	0.716±0.021	0.715±0.091	{13, 15, 33}
4	0.741±0.020	0.729±0.079	{13, 15, 33, 37}

Table B.14: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.721±0.019	0.725±0.077	{15, 27}
3	0.719±0.018	0.718±0.067	{15, 16, 45}
4	0.734±0.010	0.716±0.070	{9, 15, 27, 41}

Table B.15: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the augmented feature set method using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.722±0.017	0.724±0.057	{15, d}
3	0.7262±0.022	0.719±0.064	{13, 15, d}
4	0.725±0.027	0.718±0.089	{13, 15, 33, d}

Table B.16: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the augmented feature set method using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.722±0.017	0.724±0.057	{15, d}
3	0.724±0.018	0.723±0.055	{15, 27, d}
4	0.737±0.027	0.729±0.088	{15, 16, 45, d}

Table B.17: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the reselected feature set method using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.722±0.016	0.723±0.055	{13, 15}
3	0.716±0.021	0.715±0.091	{13, 15, 33}
4	0.741±0.020	0.729±0.079	{13, 15, 33, 37}

Table B.18: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the reselected feature set method using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.720±0.012	0.723±0.052	{15}
2	0.721±0.014	0.721±0.056	{15, 27}
3	0.735±0.022	0.734±0.093	{15, 16, 45}
4	0.727±0.014	0.706±0.058	{9, 15, 27, 41}

Table B.19: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-dependent higher-order textons through 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.586±0.018	0.588±0.083	{12}
2	0.592±0.018	0.553±0.066	{12, 30}
3	0.680±0.020	0.664±0.071	{12, 49, 54}
4	0.746±0.015	0.725±0.058	{12, 49, 54, 59}

Table B.20: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-independent higher-order textons through 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.586±0.018	0.588±0.083	{12}
2	0.605±0.028	0.587±0.121	{10, 35}
3	0.667±0.017	0.627±0.070	{19, 35, 52}
4	0.649±0.019	0.598±0.070	{16, 35, 36, 37}

B.4 Detailed Results for Risk Classification of Texture vs Density - Part II

Feature indexing for Tables B.21 to B.36 in this section is described in Table A.1 of Appendix A. These tables (Table B.21 to Table B.36) show detailed results for the experiment of texture versus density described in Section 7.4. Results reported here were obtained by selecting features using the whole 160 training image data set and classification performance for the 160 testing image data set was calculated with hold-out validation (training on the 160 training images and testing on the 160 testing images).

Table B.21: Detailed risk classification AUC scores for 60 texton features calculated from label-dependent higher-order textons using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.79	0.720	{15}
2	0.857	0.708	{11, 15}
3	0.872	0.722	{9, 11, 15}
4	0.874	0.717	{9, 11, 15, 49}
5	0.876	0.720	{9, 11, 15, 49 56}
6	0.884	0.719	{9, 11, 15, 49, 50, 56}

Table B.22: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.850	0.715	{15, 60}
3	0.855	0.721	{12, 15, 60}
4	0.886	0.744	{12, 15, 25, 60}
5	0.890	0.740	{4, 12, 15, 25, 60}
6	0.891	0.718	{4, 12, 15, 25, 27, 60}

Table B.23: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the augmented feature set method using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.793	0.722	{15, 61}
3	0.860	0.711	{15, 30, 61}
4	0.868	0.723	{15, 30, 42, 61}
5	0.879	0.718	{15, 28, 30, 42, 61}
6	0.880	0.723	{2, 15, 28, 30, 42, 61}

Table B.24: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the augmented feature set method using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.793	0.722	{15, 61}
3	0.852	0.722	{15, 60, 61}
4	0.864	0.723	{12, 15, 60, 61}
5	0.883	0.746	{12, 15, 25, 60, 61}
6	0.885	0.743	{4, 12, 15, 25, 60, 61}

Table B.25: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the reselected feature set method using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.857	0.708	{11, 15}
3	0.872	0.722	{9, 11, 15}
4	0.874	0.717	{9, 11, 15, 49}
5	0.876	0.720	{9, 11, 15, 49, 56}
6	0.884	0.773	{9, 11, 15, 49, 50, 56}

Table B.26: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the reselected feature set method using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.866	0.709	{15, 45}
3	0.882	0.711	{12, 15, 45}
4	0.886	0.699	{12, 15, 34, 45}
5	0.893	0.678	{12, 15, 34, 40, 45}
6	0.897	0.684	{12, 15, 34, 40, 43, 45}

Table B.27: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-dependent higher-order textons through the recalculated feature set method using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.672	0.591	{31}
2	0.717	0.572	{13, 31}
3	0.740	0.631	{11, 13, 31}
4	0.750	0.620	{11, 13, 25, 31}
5	0.763	0.673	{11, 13, 25, 31, 42}
6	0.772	0.659	{11, 13, 16, 25, 31, 42}

Table B.28: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-independent higher-order textons through the recalculated feature set method using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.623	0.487	{35}
2	0.759	0.603	{10, 35}
3	0.771	0.679	{10, 17, 35}
4	0.772	0.695	{10, 17, 30, 35}
5	0.784	0.676	{10, 17, 30, 35, 43}
6	0.803	0.669	{10, 17, 23, 30, 35, 43}

Table B.29: Detailed risk classification AUC scores for 60 texton features calculated from label-dependent higher-order textons using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.864	0.703	{13, 15}
3	0.8801	0.705	{13, 15, 33}
4	0.886	0.696	{13, 15, 33, 37}

Table B.30: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.874	0.695	{15, 27}
3	0.883	0.711	{15, 16, 45}
4	0.892	0.700	{9, 15, 27, 41}

Table B.31: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the augmented feature set method using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.793	0.722	{15, d}
3	0.864	0.703	{13, 15, d}
4	0.875	0.709	{13, 15, 33, d}

Table B.32: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the augmented feature set method using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
2	0.793	0.722	{15, d}
3	0.876	0.691	{15, 27, d}
4	0.880	0.717	{15, 16, 45, d}

Table B.33: Detailed risk classification AUC scores for combining 60 texton features calculated from label-dependent higher-order textons and one density feature through the reselected feature set method using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.864	0.703	{13, 15}
3	0.880	0.705	{13, 15, 33}
4	0.886	0.696	{13, 15, 33, 37}

Table B.34: Detailed risk classification AUC scores for combining 60 texton features calculated from label-independent higher-order textons and one density feature through the reselected feature set method using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.790	0.720	{15}
2	0.874	0.695	{15, 27}
3	0.883	0.711	{15, 16, 45}
4	0.892	0.700	{9, 15, 27, 41}

Table B.35: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-dependent higher-order textons through hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.708	0.584	{12}
2	0.738	0.583	{12, 30}
3	0.765	0.631	{12, 49, 54}
4	0.788	0.669	{12, 49, 54, 59}

Table B.36: Detailed risk classification AUC scores for combined 60 texton density features calculated from label-independent higher-order textons through hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.708	0.584	{12}
2	0.759	0.603	{10, 35}
3	0.790	0.652	{19, 35, 52}
4	0.772	0.584	{16, 35, 36, 37}

B.5 Detailed Results for Temporal Breast Cancer Risk Assessment - Part I

Feature indexing for Tables B.37 to B.46 in this section is described in Table A.2 of Appendix A. These tables (Table B.37 to Table B.46) show detailed results for the experiment of temporal breast cancer risk assessment described in Section 8.3.2 and Section 8.3.3. Results reported here were obtained by selecting features using the whole 100 training image data set and classification performance for the 100 testing image data set was calculated with 5-fold cross validation.

Table B.37: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying current ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.738±0.031	0.749±0.124	{14, 19}
3	0.734±0.034	0.743±0.120	{14, 19, 54}
4	0.735±0.030	0.735±0.123	{14, 19, 35, 54}
5	0.735±0.024	0.713±0.114	{14, 19, 35, 54, 57}
6	0.738±0.028	0.714±0.128	{14, 19, 23, 35, 54, 57}
7	0.738±0.029	0.696±0.142	{14, 19, 23, 29, 35, 54, 57}
8	0.743±0.031	0.654±0.140	{14, 19, 23, 25, 29, 35, 54, 57}

Table B.38: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying current contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.580±0.035	0.562±0.132	{24, 60}
3	0.627±0.027	0.591±0.128	{24, 54, 60}
4	0.625±0.038	0.578±0.133	{24, 53, 54, 60}
5	0.632±0.031	0.513±0.104	{23, 24, 53, 54, 60}
6	0.634±0.032	0.495±0.104	{23, 24, 35, 53, 54, 60}
7	0.665±0.020	0.465±0.094	{23, 24, 33, 35, 53, 54, 60}
8	0.667±0.013	0.461±0.087	{3, 23, 24, 33, 35, 53, 54, 60}

Table B.39: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying previous two year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.566±0.035	0.524±0.154	{30, 53}
3	0.663±0.021	0.602±0.098	{30, 53, 58}
4	0.728±0.026	0.674±0.106	{28, 30, 53, 58}
5	0.739±0.021	0.637±0.116	{19, 28, 30, 53, 58}
6	0.738±0.018	0.624±0.112	{19, 28, 30, 53, 58, 60}
7	0.746±0.011	0.618±0.010	{19, 28, 30, 40, 53, 58, 60}
8	0.761±0.009	0.562±0.177	{19, 26, 28, 30, 40, 53, 58, 60}

Table B.40: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying previous two year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.687±0.035	0.650±0.130	{9, 15}
3	0.693±0.044	0.616±0.123	{9, 15, 36}
4	0.693±0.044	0.622±0.125	{9, 15, 36, 47}
5	0.691±0.046	0.584±0.141	{9, 15, 35, 36, 47}
6	0.692±0.047	0.559±0.136	{9, 15, 35, 36, 47, 49}
7	0.701±0.043	0.544±0.068	{3, 9, 15, 35, 36, 47, 49}
8	0.7134±0.047	0.552±0.079	{3, 4, 9, 15, 35, 36, 47, 49}

Table B.41: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying previous four year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.6522±0.038	0.606±0.174	{40, 44}
3	0.651±0.040	0.595±0.172	{40, 44, 49}
4	0.650±0.038	0.550±0.158	{11, 40, 44, 49}
5	0.660±0.045	0.532±0.142	{11, 40, 44, 49, 56}
6	0.682±0.038	0.515±0.133	{11, 40, 44, 49, 56, 59}
7	0.765±0.039	0.594±0.174	{11, 40, 43, 44, 49, 56, 59}
8	0.766±0.044	0.601±0.154	{11, 19, 40, 43, 44, 49, 56, 59}

Table B.42: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying previous four year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.582±0.028	0.503±0.085	{48, 51}
3	0.614±0.027	0.522±0.070	{38, 48, 51}
4	0.722±0.019	0.648±0.084	{24, 38, 48, 51}
5	0.754±0.017	0.682±0.071	{24, 28, 38, 48, 51}
6	0.763±0.025	0.654±0.105	{23, 24, 28, 38, 48, 51}
7	0.779±0.022	0.660±0.079	{23, 24, 28, 38, 46, 48, 51}
8	0.784±0.022	0.662±0.094	{23, 24, 28, 29, 38, 46, 48, 51}

Table B.43: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year ipsilateral high and low risk images using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.552±0.027	0.551±0.115	{17}
2	0.679±0.040	0.653±0.165	{17, 31}
3	0.700±0.030	0.633±0.146	{17, 31, 54}
4	0.698±0.018	0.621±0.094	{14, 17, 31, 54}
5	0.696±0.015	0.631±0.093	{14, 17, 26, 31, 54}
6	0.697±0.013	0.611±0.078	{14, 17, 26, 31, 49, 54}
7	0.711±0.021	0.623±0.063	{14, 15, 17, 26, 31, 49, 54}
8	0.726±0.021	0.655±0.066	{14, 15, 17, 23, 26, 31, 49, 54}

Table B.44: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year contralateral high and low risk images using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.566±0.038	0.562±0.193	{36}
2	0.564±0.049	0.451±0.133	{36, 37}
3	0.586±0.027	0.469±0.139	{32, 36, 37}
4	0.589±0.032	0.452±0.140	{32, 36, 37, 58}
5	0.601±0.028	0.402±0.100	{32, 34, 36, 37, 58}
6	0.708±0.049	0.552±0.180	{32, 34, 36, 37, 54, 58}
7	0.720±0.039	0.563±0.157	{16, 32, 34, 36, 37, 54, 58}
8	0.736±0.037	0.601±0.129	{16, 32, 34, 36, 37, 42, 54, 58}

Table B.45: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year ipsilateral high and low risk images using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.527±0.032	0.463±0.143	{20}
2	0.608±0.018	0.551±0.076	{20, 22}
3	0.605±0.019	0.532±0.059	{20, 22, 58}
4	0.617±0.024	0.537±0.048	{20, 22, 40, 58}
5	0.623±0.012	0.487±0.078	{20, 22, 28, 40, 58}
6	0.628±0.016	0.453±0.086	{20, 22, 28, 35, 40, 58}
7	0.655±0.013	0.507±0.056	{20, 22, 28, 29, 35, 40, 58}
8	0.691±0.035	0.511±0.107	{20, 22, 28, 29, 32, 35, 40, 58}

Table B.46: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year contralateral high and low risk images using 5-fold cross validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.552±0.028	0.567±0.117	{24}
2	0.551±0.021	0.486±0.079	{22, 24}
3	0.671±0.027	0.650±0.090	{18, 22, 24}
4	0.681±0.023	0.655±0.086	{18, 22, 24, 31}
5	0.705±0.019	0.664±0.059	{18, 22, 24, 31, 35}
6	0.715±0.025	0.611±0.114	{18, 22, 24, 31, 35, 60}
7	0.716±0.024	0.599±0.094	{18, 22, 24, 29, 31, 35, 60}
8	0.720±0.026	0.567±0.086	{5, 18, 22, 24, 29, 31, 35, 60}

Table B.47: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection for classifying current ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.609±0.020	0.610±0.091	{42}
2	0.606±0.034	0.592±0.066	{19, 21}
3	0.693±0.023	0.686±0.086	{2, 19, 37}
4	0.727±0.049	0.683±0.124	{6, 17, 19, 54}

Table B.48: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection for classifying current contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.525±0.031	0.511±0.114	{15}
2	0.580±0.035	0.562±0.132	{24, 60}
3	0.629±0.028	0.592±0.145	{24, 29, 60}
4	0.661±0.030	0.610±0.166	{9, 24, 45, 60}

Table B.49: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection for classifying previous two year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.548±0.061	0.375±0.163	{18}
2	0.628±0.036	0.584±0.145	{1, 43}
3	0.731±0.013	0.685±0.099	{28, 33, 50}
4	0.740±0.005	0.668±0.134	{19, 28, 33, 50}

Table B.50: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection for classifying previous two year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.559±0.035	0.588±0.177	{13}
2	0.687±0.035	0.650±0.130	{9, 15}
3	0.634±0.035	0.535±0.096	{4, 10, 15}
4	0.695±0.044	0.562±0.116	{9, 15, 49, 58}

Table B.51: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection for classifying previous four year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.540±0.030	0.562±0.116	{25}
2	0.614±0.013	0.591±0.077	{26, 44}
3	0.560±0.026	0.394±0.089	{19, 31, 41}
4	0.665±0.023	0.627±0.071	{41, 44, 55, 59}

Table B.52: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and exhaustive search feature selection for classifying previous four year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.544±0.046	0.497±0.174	{41}
2	0.533±0.062	0.434±0.130	{41, 59}
3	0.643±0.025	0.603±0.118	{16, 21, 42}
4	0.702±0.018	0.659±0.098	{16, 21, 28, 52}

Table B.53: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year ipsilateral high and low risk images using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.557±0.037	0.5630±0.141	{25}
2	0.663±0.037	0.617±0.139	{19, 55}
3	0.718±0.036	0.654±0.146	{17, 19, 54}
4	0.667±0.042	0.609±0.113	{14, 17, 51, 58}

Table B.54: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year contralateral high and low risk images using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.531±0.022	0.386±0.089	{18}
2	0.563±0.044	0.495±0.182	{37, 44}
3	0.592±0.037	0.464±0.094	{18, 36, 37}
4	0.644±0.037	0.438±0.087	{18, 24, 37, 57}

Table B.55: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year ipsilateral high and low risk images using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.527±0.032	0.463±0.143	{20}
2	0.618±0.022	0.604±0.07	{21, 51}
3	0.618±0.011	0.552±0.061	{2, 37, 51}
4	0.662±0.013	0.602±0.068	{5, 19, 21, 44}

Table B.56: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year contralateral high and low risk images using 5-fold cross validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.558±0.029	0.558±0.131	{45}
2	0.639±0.034	0.617±0.156	{18, 34}
3	0.656±0.025	0.622±0.074	{14, 18, 52}
4	0.658±0.025	0.614±0.077	{18, 24, 31, 52}

B.6 Detailed Results for Temporal Breast Cancer Risk Assessment - Part II

Feature indexing for Tables B.57 to B.66 in this section is described in Table A.2 of Appendix A. These tables (Table B.57 to Table B.66) show detailed results for the experiment of temporal breast cancer risk assessment by using hold-out validation as an comparison to the experiments described in Section 8.3.2 and Section 8.3.3. Results reported here were obtained by selecting features using the whole 100 training image data set and classification performance for the 100 testing image data set was calculated with hold-out validation (training on the 100 training images and testing on the 100 testing images).

Table B.57: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and sequential feature selection for classifying current ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.759	0.719	{14, 19}
3	0.901	0.674	{14, 19, 54}
4	0.922	0.666	{14, 19, 35, 54}
5	0.924	0.660	{14, 19, 35, 54, 57}
6	0.946	0.644	{14, 19, 23, 35, 54, 57}
7	0.956	0.641	{14, 19, 23, 29, 35, 54, 57}
8	0.962	0.640	{14, 19, 23, 25, 29, 35, 54, 57}

Table B.58: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and sequential feature selection for classifying current contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.816	0.502	{24, 60}
3	0.859	0.613	{24, 54, 60}
4	0.874	0.507	{24, 53, 54, 60}
5	0.886	0.512	{23, 24, 53, 54, 60}
6	0.896	0.528	{23, 24, 35, 53, 54, 60}
7	0.905	0.527	{23, 24, 33, 35, 53, 54, 60}
8	0.912	0.528	{3, 23, 24, 33, 35, 53, 54, 60}

Table B.59: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and sequential feature selection for classifying previous two year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.647	0.447	{30, 53}
3	0.767	0.553	{30, 53, 58}
4	0.812	0.596	{28, 30, 53, 58}
5	0.883	0.595	{19, 28, 30, 53, 58}
6	0.843	0.612	{19, 28, 30, 53, 58, 60}
7	0.863	0.564	{19, 28, 30, 40, 53, 58, 60}
8	0.874	0.749	{19, 26, 28, 30, 40, 53, 58, 60}

Table B.60: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and sequential feature selection for classifying previous two year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.756	0.648	{9, 15}
3	0.758	0.647	{9, 15, 36}
4	0.800	0.607	{9, 15, 36, 47}
5	0.817	0.608	{9, 15, 35, 36, 47}
6	0.842	0.584	{9, 15, 35, 36, 47, 49}
7	0.846	0.579	{3, 9, 15, 35, 36, 47, 49}
8	0.862	0.594	{3, 4, 9, 15, 35, 36, 47, 49}

Table B.61: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying previous four year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.635	0.638	{40, 44}
3	0.753	0.586	{40, 44, 49}
4	0.774	0.587	{11, 40, 44, 49}
5	0.789	0.569	{11, 40, 44, 49, 56}
6	0.800	0.608	{11, 40, 44, 49, 56, 59}
7	0.813	0.664	{11, 40, 43, 44, 49, 56, 59}
8	0.836	0.696	{11, 19, 40, 43, 44, 49, 56, 59}

Table B.62: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using 5-fold cross validation and sequential feature selection for classifying previous four year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
2	0.708	0.427	{48, 51}
3	0.742	0.449	{38, 48, 51}
4	0.812	0.635	{24, 38, 48, 51}
5	0.829	0.677	{24, 28, 38, 48, 51}
6	0.851	0.701	{23, 24, 28, 38, 48, 51}
7	0.872	0.609	{23, 24, 28, 38, 46, 48, 51}
8	0.887	0.586	{23, 24, 28, 29, 38, 46, 48, 51}

Table B.63: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year ipsilateral high and low risk images using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.643	0.448	{17}
2	0.690	0.588	{17, 31}
3	0.810	0.614	{17, 31, 54}
4	0.836	0.630	{14, 17, 31, 54}
5	0.849	0.629	{14, 17, 26, 31, 54}
6	0.849	0.623	{14, 17, 26, 31, 49, 54}
7	0.858	0.640	{14, 15, 17, 26, 31, 49, 54}
8	0.864	0.617	{14, 15, 17, 23, 26, 31, 49, 54}

Table B.64: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year contralateral high and low risk images using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.655	0.435	{36}
2	0.775	0.485	{36, 37}
3	0.825	0.502	{32, 36, 37}
4	0.839	0.502	{32, 36, 37, 58}
5	0.842	0.502	{32, 34, 36, 37, 58}
6	0.852	0.525	{32, 34, 36, 37, 54, 58}
7	0.893	0.501	{16, 32, 34, 36, 37, 54, 58}
8	0.898	0.514	{16, 32, 34, 36, 37, 42, 54, 58}

Table B.65: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year ipsilateral high and low risk images using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.599	0.519	{20}
2	0.721	0.579	{20, 22}
3	0.754	0.560	{20, 22, 58}
4	0.795	0.579	{20, 22, 40, 58}
5	0.820	0.573	{20, 22, 28, 40, 58}
6	0.834	0.568	{20, 22, 28, 35, 40, 58}
7	0.834	0.577	{20, 22, 28, 29, 35, 40, 58}
8	0.840	0.585	{20, 22, 28, 29, 32, 35, 40, 58}

Table B.66: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year contralateral high and low risk images using hold-out validation and sequential feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.631	0.550	{24}
2	0.725	0.551	{22, 24}
3	0.884	0.680	{18, 22, 24}
4	0.893	0.677	{18, 22, 24, 31}
5	0.915	0.642	{18, 22, 24, 31, 35}
6	0.920	0.640	{18, 22, 24, 31, 35, 60}
7	0.916	0.630	{18, 22, 24, 29, 31, 35, 60}
8	0.926	0.636	{5, 18, 22, 24, 29, 31, 35, 60}

Table B.67: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection for classifying current ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.750	0.389	{42}
2	0.888	0.577	{19, 21}
3	0.922	0.596	{2, 19, 37}
4	0.948	0.635	{6, 17, 19, 54}

Table B.68: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection for classifying current contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.655	0.529	{15}
2	0.816	0.502	{24, 60}
3	0.870	0.506	{24, 29, 60}
4	0.888	0.523	{9, 24, 45, 60}

Table B.69: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection for classifying previous two year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.874	0.563	{18}
2	0.749	0.596	{1, 43}
3	0.786	0.643	{28, 33, 50}
4	0.828	0.637	{19, 28, 33, 50}

Table B.70: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection for classifying previous two year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.643	0.445	{13}
2	0.756	0.648	{9, 15}
3	0.809	0.552	{4, 10, 15}
4	0.839	0.621	{9, 15, 49, 58}

Table B.71: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection for classifying previous four year ipsilateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.600	0.462	{25}
2	0.748	0.589	{26, 44}
3	0.819	0.536	{19, 31, 41}
4	0.854	0.626	{41, 44, 55, 59}

Table B.72: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons using hold-out validation and exhaustive search feature selection for classifying previous four year contralateral high and low risk images.

num of features	training AUC	testing AUC	feature combination
1	0.655	0.552	{41}
2	0.869	0.559	{41, 59}
3	0.916	0.636	{16, 21, 42}
4	0.927	0.652	{16, 21, 28, 52}

Table B.73: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year ipsilateral high and low risk images using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.666	0.445	{25}
2	0.774	0.570	{19, 55}
3	0.814	0.625	{17, 19, 54}
4	0.842	0.552	{14, 17, 51, 58}

Table B.74: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous two year contralateral high and low risk images using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.663	0.476	{18}
2	0.789	0.483	{37, 44}
3	0.850	0.525	{18, 36, 37}
4	0.863	0.559	{18, 24, 37, 57}

Table B.75: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year ipsilateral high and low risk images using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.599	0.519	{20}
2	0.795	0.535	{21, 51}
3	0.813	0.513	{2, 37, 51}
4	0.840	0.583	{5, 19, 21, 44}

Table B.76: Detailed risk classification AUC scores for 60 texton features calculated from label-independent higher-order textons generated from current year mammograms for classifying previous four year contralateral high and low risk images using hold-out validation and exhaustive search feature selection.

num of features	training AUC	testing AUC	feature combination
1	0.654	0.559	{45}
2	0.803	0.631	{18, 34}
3	0.901	0.618	{14, 18, 52}
4	0.911	0.627	{18, 24, 31, 52}

Bibliography

AIHW & AACR. Breast cancer in Australia: an overview 2012. Technical Report Cancer series no. 74. Cat. no. CAN 70., Australia Institute of Health and Welfare & Australia Association of Cancer Registries, Canberra: AIHW, 2012a.

AIHW & AACR. Breast cancer in Australia: an overview. Technical Report Cancer series no. 74. Cat. no. CAN 70., Australia Institute of Health and Welfare & Australia Association of Cancer Registries, Canberra: AIHW, 2012b.

R. Alteri, P. Brandi, and L. Brinton etc. Breast cancer facts & figures 2011-2012. Technical report, American Cancer Society, Atlanta: American Cancer Society, Inc, 2011.

American Association for Cancer Research. Breast cancer screening goes personalized. *Cancer Discovery*, 2(3):200, 2012.

American Cancer Society (ACS). *What are the key statiscs for breast cancer?*, Mar 2012a. http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-key-statistics.

American Cancer Society (ACS). *Breast cancer facts & figures 2011-2012*, Jan 2012b. http://www.cancer.org/Research/CancerFactsFigures/BreastCancerFactsFigures/breast-cancer-facts-and-figures-2011-2012.

American College of Radiology. *American College of Radiology (ACR) Breast Imaging Reporting and Data System Atlas (BI-RADS)*. Reston, third edition, 2003.

E. Amir, D. G. Evans, A. Shenton, F. Lalloo, A. Moran, C. Boggis, M. Wilson, and A. Howell. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *Journal of Medical Generics*, 40 (11):807–814, 2003.

A. R. Backes, D. Casanova, and O. M. Bruno. Color texture analysis based on fractal descriptors. *Pattern Recognition*, 45(5):1984–1992, 2012.

R. Bajcsy. Computer description of textured surfaces. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, number 8 in IJCAI'73, pages 572–579, San Francisco, CA, USA, 1973. Morgan Kaufmann Publishers Inc.

W. E. Barlow, E. White, R. B. Barbash, P. M. Vacek, L. T. Ernstoff, P. A. Carney, J. A. Tice, D. S. M. Buist, B. M. Geller, R. Rosenberg, B. C. Yankaskas, and K. Kerlikowske. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98 (17):1204–1214, 2006.

W. A. Berg, J. D. Blume, and J. B. Cormack. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer: Results of the first-year screen in ACRIN 6666. *JAMA: the Journal of the American Medical Association*, 299(18):2151–2163, 2008.

W. A. Berg, Z. Zhang, and D. Lehrer. Dectection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA: the Journal of the American Medical Association*, 307(13):1394–1404, 2012.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1973.

A. M. J. Bluekens, R. Holland, N. Karssemeijer, M. J. Broeders, and G. J. den Heeten. Comparison of digital screening mammography and screen film mammography in the early detection of clinically relevant cancers: A multicenter study. *Radiology*, 265(3):707–714, 2012.

M. J. Bottema, G. N. Lee, and S. Lu. Automatic image feature extraction for diagnosis and prognosis of breast cancer. In A. Jain, A. Jain, S. Jain, and L. Jain, editors, *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, volume 39 of *Machine Perception and Artificial Intelligence*, chapter 2. World Scientific Publishing Co. Pte. Ltd, 2000.

J. D. Boyce. Program on breast cancer environmental risk factors - ionizing radiation and breast cancer risk. Technical Report Fact Sheet # 52, Sprecher Institute for Comparative Cancer Research, Cornell University, Ithaca, 2004.

N. F. Boyd, B. O'Sullivan, E. Fishell, I. Simor, and G. Cooke. Mammographic patterns and breast cancer risk: Methodologic standards and contradictory results. *Journal of National Cancer Institute*, 72(6):1253–1259, June 1984.

N. F. Boyd, J. W. Bying, R. A. Jong, E. K. Fishell, L. E. Little, A. B. Miller, G. A. Lockwood, D. L. Tritchler, and M. J. Yaffe. Quantitative classification of mammographic densities and breast cancer risk: Results from the Canadian National Breast Screening Study. *Journal of the National Cancer Institute*, 87(9):670–675, 1995.

P. Boyle and B. Levin. *World Cancer Report 2008*. IARC Press, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France, 2008.

J. Brisson, F. Merletti, N. L. Sadowsky, J. A. Twaddle, A. S. Morrison, and P. Cole. Mammographic features of the breast and breast cancer risk. *American Journal of Epidemiology*, 115(3):428–437, Jul 1981.

J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe. The quantitative analysis of mammographic densities. *Physics in Medicine and Biology*, 39(10): 1629, 1994.

J. W. Byng, N. F. Boyd, E. Fishell, R. A Jong, and M. J. Yaffe. Automated analysis of mammographic densities. *Physics in Medicine and Biology*, 41(5):909, 1996.

J. W. Byng, M. J. Yaffe, G. A. Lockwood, L. E. Little, D. L. Tritchler, and N. F. Boyd. Automated analysis of mammographic densities and breast carcinoma risk. *Cancer*, 80(1):66–74, Jul 1997.

C. B. Caldwell, S. J. Stapleton, D. W. Holdsworth, R. A. Jong, W. J. Weiser, G. Cooke, and M. J. Yaffe. Characterization of mammographic parenchymal pattern by fractal dimension. *Physics in Medicine and Biology*, 35(2):235, 1990.

Cancer Research UK. Cancer statistics briefing - cancer in the UK. Technical report, Cancer Research UK, 2012.

Cancer Research UK. Cancer statistics key facts - breast cancer. Technical report, Cancer Research UK, 2013a.

Cancer Research UK. Cancer statistics report - cancer mortality in the UK 2011. Technical report, Cancer Research UK, 2013b.

H-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick. Effects of sample size on classifier design for computer-aided diagnosis. In K. M. Hanson, editor, *Proc. SPIE*, volume 3338 of *Medical Imaging 1998: Image Processing*, pages 845–858, San Diego, CA, Feb 1998.

H-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. *Medical Physics*, 26(12):2654–2668, 1999.

T. Chang and C-C. J. Kuo. Texture analysis and classification with tree-structured wavelet transform. *Image Processing, IEEE Transactions on*, 2(4):429–441, 1993.

C-C. Chen and C-L. Huang. Markov random fields for texture classification. *Pattern Recognition Letters*, 14(11):907–914, Nov 1993.

J. Chen, D. Pee, R. Ayyagari, B. Graubard, C. Shairer, C. Byrne, J. Benichou, and M. H. Gail. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute*, 98(17):1215–1226, 2006.

E. B. Claus, N. Risch, and W. D. Thompson. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer*, 73(3):643–651, Feb 1994.

M. O. Connor, D. Rhodes, and C. Hruska. Molecular breast imaging. *Expert Review of Anticancer Therapy*, 9(8):1073–1080, Aug 2009.

J. P. Costantino, M. H. Gail, D. Pee, S. Anderson, C. K. Redmond, J. Benichou, and H. S. Wieand. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute*, 91(18): 1541–1548, Sep 1999.

G. R. Cross and A. K. Jain. Markov random field texture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-5(1):25–39, 1983.

O. G. Cula and K. J. Dana. 3D texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*, 59(1):33–60, August 2004.

K. J. Dana, B. V. Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics*, 18(1):1–34, Jan 1999.

B. V. Dasarathy and E. B. Holder. Image characterizations based on joint gray level - run length distributions. *Pattern Recognition Letters*, 12(8):497–502, 1991.

B. V. Dasarathy and B. V. Sheela. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708–713, May 1979.

M. Dash and H. Liu. Feature selection for classfiication. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, July 1985.

L. S. Davis, S. A. Johns, and J. K. Aggarwal. Texture analysis using generalized co-occurrence matrices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(3):251–259, 1979.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

C. Desantis, J. Ma, L. Bryan, and A. Jemal. Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 64(1):52–62, Jan 2014.

J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

R. L. Egan and R. C. Mosteller. Breast cancer mammography patterns. *Cancer*, 40 (5):2087–2090, Nov 1977.

D. G. R. Evans and A. Howell. Breast cancer risk-assessment models. *Breast Cancer Research*, 9(5):213–220, Sep 2007. ISSN 1465-5411.

J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin. *Cancer Incidence and Mortality Worldwide*. IARC CancerBase No. 10. IARC Press, International Agency for Research on Cancer, 2010.

H. L. Fred. Drawbacks and limitations of computed tomography. *Texas Heart Institute Journal*, 31(4):345–348, 2004.

K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition (Computer Science and Scientific Computing)*. New York: Academic Press, 1990.

B. D. Gabor and D. Ing. Theory of communication. *Electrial Engineering - Part III: Radio and Communication*, 93(26):429–441, 1946.

M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, Dec 1989.

M. M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics and Image Proceesing*, 4(2):172–179, Jun 1975.

G. Gennaro and C. Maggio. Dose comparison between screen/film and full-field digital mammography. *European Radiology*, 16(11):2559–2566, 2006.

M. L. Giger, N. Karssemeijer, and J. A. Schnabel. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *The Annual Review of Biomedical Engineering*, 15:327–357, May 2013.

Y-C. Gong and S. Petroudi. Texture based mammogram classification and segmentation. In M. Brady S. M. Astley, C. Rose, and R. Zwiggelaar, editors, *Digital Mammography 8th International Workshop, IWDM 2006, Manchester*, volume 4046 of *Lecture Notes in Computer Science*, pages 616–625. Springer Berlin Heidelberg, 2006.

I. T. Gram, E. Funkhouser, and L. Tabár. The Tabár classification of mammographic parenchymal patterns. *European Journal of Radiology*, 24(2):131–136, Feb 1997.

D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. Huntington NY, Krieger, rev edition, 1974.

Y-M. Guo, G-Y. Zhao, and M. Pietikanen. Discriminative features for texture description. *Pattern Recognition*, 45(10):3834–3843, 2012a.

Z-H. Guo, Q. Li, L. Zhang, J. You, W-H. Liu, and J-H. Wang. Texture image classification using complex texton. In D-S. Huang, Y. Gan, P. Gupta, and M. M. Gromiha, editors, *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, volume 6839 of *Lecture Notes in Computer Science*, pages 98–104. Springer Berlin Heidelberg, 2012b.

L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 12(10):993–1001, Oct 1990.

R. M. Haralick, K. Shanmugam, and I. Dinstein. Texture features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer-Verlag, second edition, 2008.

H.R. Hauge, K. Pedersen, S. Hofvind, and H. M. Olerud. Patient doses from screen-film and full-field digital mammography in a population-based screening programme. *Radiation Protection Dosimetry 2012*, 148(1):65–73, 2011.

D-C. He and L. Wang. Texture unit, texture spectrum, and texture analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, 28(4):509–512, 1990.

W-D. He, E. R. E. Denton, and R. Zwiggelaar. Mammographic segmentation based on mammographic parenchymal patterns and spatial moments. In *9th International Conference on Information Technology and Applications in Biomedicine*, pages 1–4, Nov 2009.

W-D. He, R. E. Denton, and R. Zwiggelaar. Mammographic segmentation and risk classification using a novel binary model based Bayes classifier. In A. D. A. Maidment, P. R. Bakic, and S. Gavenonis, editors, *Breast Imaging*, volume 7361 of *Lecture Notes in Computer Science*, pages 40–47. Springer Berlin Heidelberg, Jul 2012.

M. Heath, K. Bowyer, D. l. Kopans, W. P. Kegelmeyer, R. Moore, K. Chang, and S. MunishKumaran. Current status of the digital database for screening mammography. In *the 4th International Workshop on Digital Mammography*, pages 457–460, 1998.

M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer. The digital database for screening mammography. In M. J. Yaffe, editor, *The 5th Intenational Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001.

J. J. Heine, C. G. Scott, T. A. Sellers, K. R. Kathleen, D. J. Serie, F-F. Wu, M. J. Morton, B. A. Schueler, F. J. Couch, J.E.Olson, V. S. Pankratz, and C. M. Vachon. A novel automated mammographic density measure and breast cancer risk. *Journal of the National Cancer Institute*, 104(13):1028–1037, Jul 2012.

M. A. Helvie. Digital mammography imaging: Breast tomosynthesis and advanced applications. *Radiologic Clinics of North America*, 48(5):917–929, 2010.

G. T. Herman. *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag London, 2nd edition, 2009.

R. Highnam, M. Brady, M. J. Yaffe, N. Karssemeijer, and J. Harvey. Robust breast composition measurement - Volpara^{TM}. In J. Martí, A. Oliver, J. Freixenet, and R. Martí, editors, *Digital Mammography*, volume 6136 of *Lecture Notes in Computer Science*, pages 342–349. Springer Berlin Heidelberg, 2010.

T. K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, Aug 1998.

J. M. Hrung, S. S. Sonnad, J. S. Schwartz, and C. P. Langlotz. Accuracy of MR imaging in the work-up of suspicious breast leisons: A diagnostic meta-analysis. *Academic Radiology*, 6(7):387–397, 1999.

T. I. Hsu, A. D. Calway, and R. Wilson. Texture analysis using the multiresolution Fourier transform. In *Proc 8th Scandinavian Conference on Image Analysis*, pages 823–830. IAPR, May 1993.

W-Y. Hu, H. Li, C-Y. Wang, S-M. Gou, and L. Fu. Characterization of collagen fibers by means of texture analysis of second harmonic generation images using orientation-dependent gray level co-occurrence matrix method. *Journal of Biomedical Optics*, 17(2):026007–1–026007–9, 2012.

D. Huawu and D. A. Clausi. Gaussian MRF rotation-invariant features for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(7):951–955, 2004.

Z-M. Huo, M. L. Giger, D. E. Wolverton, W. Zhong, S. Cumming, and O. L. Olopade. Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection. *Medical Physics*, 27(1):4–12, 2000.

Z-M. Huo, M. L. Giger, O. L. Olopade, D. E. Wolverton, B. L. Weber, C. E. Metz, W. Zhong, and S. A. Cummings. Computerized analysis of digitized mammograms of BRCA1 and BRCA2 gene multation carriers. *Radiology*, 225(2):519–526, Nov 2002.

Y. Javed, M. M. Khan, and J. Chanussot. Population density estimation using textons. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 2206–2209, 2012.

B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(290):91–97, May 1981.

B. Julesz and J. R. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *The Bell System Technical Journal*, 62(6):1619–1645, Jul-Aug 1983.

L. M. Kaplan. Extended fractal analysis for texture classification and segmentation. *Image Processing, IEEE Transactions on*, 8(11):1572–1585, 1999.

G. Karemore, B.M. Keller, H. Oh, J. Tchou, M. Nielsen, E. F. Conant, and D. Kontos. Mammographic parenchymal texture analysis for estrogen-receptor subtype specific breast cancer risk estimation. In A. D. A. Maidment, P. R. Bakie, and

S. Gavenonis, editors, *Breast Imaging*, volume 7361 of *Lecture Notes in Computer Science*, pages 596–603. Springer Berlin Heidelberg, Jul 2012.

N. Karssemeijer. Automated classification of parenchymal patterns in mammograms. *Physics in Medicine and Biology*, 43(2):365–378, 1998.

B. M. Keller, E. F. Conant, H. Oh, and D. Kontos. Breast cancer risk prediction via area and volumetric estimates of breast density. In A. D. A. Maidment, P. R. Bakic, and S. Gavenonis, editors, *Breast Imaging*, volume 7361 of *Lecture Notes in Computer Science*, pages 236–243. Springer Berlin Heidelberg, 2012.

J. M. Keller, S. Chen, and R. M. Crownover. Texture description and segmentation through fractal geometry. *Computer Vision, Graphics, and Image Processing*, 45: 150–166, Feb 1989.

K. J. Khouzani and H. S. Soltanian. Rotation-invariant multiresolution texture analysis using radon and wavelet transforms. *Image Processing, IEEE Transactions on*, 14(6):783–795, 2005.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, volume 2, pages 1137–1143, 1995.

D. Kontos, P. R. Bakic, A. K. Carton, A. B. Troxel, E. F. Conant, and A. D. A. Maidment. Parenchymal texture analysis in digital breast tomosynthesis for breast cancer risk estimation: A preliminary study. *Academic Radiology*, 16(3):283–298, March 2009.

D. Kontos, L. C. Ikejimba, Bakic P. R, A. B. Troxel, R. F. Contant, and A. D. Maidment. Analysis of parenchymal texture with digital breast tomosynthesis: Comparison with digital mammography and implications for cancer risk assessment. *Radiology*, 261(1):80–91, Jul 2011.

A. Laine and J. Fan. Texture classification by wavelet packet signatures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(11):1186–1191, 1993.

P. Langley. Selection of relevent features in machine learning. In *proceedings of the AAAI Fall symposium on relevance*, pages 140–144. AAAI Press, 1994.

S. C. Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Education Psychology*, 22(1):45–55, 1931.

K. I. Laws. Texture energy measures. In *Proceedings: Image Understanding Workshop, DARPA, Los Angeles*, pages 47–51, Nov 1979.

G. N. Lee and M. J. Bottema. Significance of classification scores subsequent to feature selection. *Pattern Recognition Letters*, 27(14):1702–1709, June 2006.

T. S. Lee. Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.

T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

J. M. Lewin, C. J. D'Orsi, R. E. Hendrick, P. K. Isaacs, A. Karellas, and G. R. Cutter. Clinical comparion of full-field digital mammography and screen-film mammography for detection of breast cancer. *AJR. American Journal of Roentgenology*, 179(3):671–677, 2002.

H. Li, M. L. Giger, Z-M Huo, O. I. Olopade, L. Lan, B. L. Weber, and L. Bonta. Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: Effect of ROI size and location. *Medical Physics*, 31(3):549–555, 2004.

H. Li, M. L. Giger, O. I. Margolis, and M. R. Chinander. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. *Academic Radiology*, 12(7):863–873, July 2006.

H. Li, M. L. Giger, O. I. Olopade, and L. Lan. Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment. *Academic Radiology*, 14 (5):513–521, May 2007.

H. Li, M. L. Giger, O. I. Olopade, and M. R. Chinander. Power spectral analysis of mammographic parenchymal patterns for breast cancer risk assessment. *Journal of Digital Imaging*, 21(2):145–152, June 2008.

H. Li, M. L. Giger, O. I. Olopade, and L. Lan. Validation of mammographic texture analysis for assessment of breast cancer risk. In J. Martí et al., editor, *Digital Mammography*, volume 6136 of *Lecture Notes in Computer Science*, pages 267–271. Springer Berlin Heidelberg, 2010.

J-M. Li, L. Szekely, L. Eriksson, B. Heddson, A. Sundbom, K. Czene, P. Hall, and K. Humphreys. High-throughput mammographic-density measurement: A tool for risk prediction of breast cancer. *Breast Cancer Research*, 14(4):R114, 2012a. ISSN 1465-5411.

X-Z. Li, S. Williams, and M. J. Bottema. Intensity independent texture analysis in screening mammograms. In A. D. A. Maidment, P. R. Bakic, and S. Gavenonis, editors, *Breast Imaging*, volume 7361 of *Lecture Notes in Computer Science*, pages 474–481. Springer Berlin Heidelberg, 2012b.

X-Z. Li, S. Williams, G. N. Lee, and M. Deng. Computer-aided mammography classification of malignant mass regions and normal regions based on novel texton features. In *Control Automation Robotics Vision (ICARCV), 2012 12th International Conference on*, pages 1431–1436, December 2012c.

X-Z. Li, S. Williams, and M. J. Bottema. Background intensity independent texture features for assessing breast cancer risk in screening mammograms. *Pattern Recognition Letters*, 34(9):1053–1062, July 2013.

X-Z. Li, S. Williams, and M. J. Bottema. Constructing and applying higher-order textons: Estimating breast cancer risk. *Pattern Recognition*, 47(3):1375–1382, Mar 2014a.

X-Z. Li, S. Williams, and M. J. Bottema. Texture and region dependent breast cancer risk assessment from screening mammograms. *Pattern Recoginition Letters*, 36 (15):117–124, Jan 2014b.

X-Z. Li, S. Williams, and M. J. Bottema. Temporal breast cancer risk assessment based on higher-order textons. In Hiroshi Fujita, Takeshi Hara, and Chisako Muramatsu, editors, *Breast Imaging*, volume 8539 of *Lecture Notes in Computer Science*, pages 565–572. Springer International Publishing, 2014c.

Z. Li, G-Z. Liu, Y. Yang, and J-Y. You. Scale- and rotation-invariant local binary pattern using scale-adaptive texton and subuniform-based circular shift. *Image Processing, IEEE Transactions on*, 21(4):2130–2140, 2012d.

Y-H. Liu, M. Muftah, T. Das, L. Bai, and K. Robson. Classification of MR tumor images based on Gabor wavelet analysis. *Journal of Medical and Biological Engeneering*, 32(1):22–28, 2012.

S. Livens, P. Scheunders, G. V. D. Wouwer, and D. V. Dyck. Wavelets for texture analysis, an overview. In *Sixth International Conference on Imaging Processing and Its Applications*, volume 2, pages 581–585, 1997.

H. H. Loh, J. G. Leu, and R. C. Luo. The analysis of natural textures using run length features. *IEEE Transactions on Industral Electronics*, 35(2):323–328, 1988.

O. Losson, A. Porebski, N. Vandenbroucke, and L. Macaire. Color texture analysis using CFA chromatic co-occurrence matrices. *Computer Vision and Image Understanding*, 117(7):747–763, 2013.

N. Lowry, R. Mangoubi, M. Desai, Y. Marzouk, and P. Sammak. Texton-based segmentation and classification of human embryonic stem cell colonies using multistage Bayesian level sets. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 194–197, 2012.

F. Ma, M. Bajger, S. Williams, and M. J. Bottema. Improved detection of cancer in screening mammograms by temporal comparison. In J. Martí, A. Oliver, J. Freixeneti, and R. Martí, editors, *Digital Mammography*, volume 6136 of *Lecture Notes in Computer Science*, pages 752–759. Springer Berlin Heidelberg, 2010.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

I. E. Magnin, F. Cluzeau, C. L. Odet, and A. Bremond. Mammographic texture analysis: An evaluation of risk for developing breast cancer. *Optical Engineering*, 25(6):780–784, Jun 1986.

J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999*, volume 2, pages 918–925, 1999.

J. Malik, S. Belongie, T. Leung, and J-B. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

J. S. Mandelblatt, N. Stout, and A. Trentham-Dietz. To screen or not to screen women in their 40s for breast cancer: Is personalized risk-based screening the answer? *Annals of Internal Medicine*, 155(1):58–60, 2011.

T. Matsuyama, S-I. Miura, and M. Nagao. Structural analysis of natural textures by Fourier transformation. *Computer Vision, Graphics and Image Processing*, 24(3): 347–362, 1983.

V. A. McCormack and I. D. S. Silva. Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. *Cancer Epidemioogy, Biomarkers and Prevention*, 15(6):1159–1169, Jun 2006.

K. McPherson, C. M. Steel, and J. M. Dixon. ABC of breast diseases: Breast cancer-epidemiology, risk factors, and genetics. *British Medical Journal (BMJ)*, 321 (7261):624–628, Sep 2000.

M. E. Mealiffe, R. P. Stokowski, B. K. Rhees, R. L. Prentice, M. Pettinger, and D. A. Hinds. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *Journal of the National Cancer Institute*, 102 (21):1618–1627, 2010.

Medical Advisory Secretariat. Cancer screening with digital mammography for women at average risk for breast cancer, magnetic resonance imaging (MRI) for women at high risk: An evidence-based analysis. *Ontario Health Technology Assessment Series*, 10(3):1–55, March 2010.

G. Medioni and Y. Yasumoto. A note on using the fractal dimension for segmentation. In *IEEE Computer Vision Workshop, Annapolis*, pages 25–30, 1984.

L. Mendell, M. Rosenbloom, and A. Naimark. Are breast patterns a risk index for breast cancer? a reappraisal. *American Journal of Roentgenology*, 128(4):547, 1977.

C. E. Metz. Basic principles of ROC analysis. *Seminars in nuclear medicine*, 8(4): 283–298, Oct 1978.

C. E. Metz, B. A. Herman, and J-H. Shen. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17(9):1033–1053, 1998.

Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, and W. Ding. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71, 1994.

F. Mosteller and J. W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology, Vol. 2*. Addison-Wesley, 1968.

F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

S. W. Myint. Fractal approaches in texture analysis and classification of remotely sensed data: Comparisons with spatial autocorrelation techniques and simple descriptive statistics. *International Journal of Remote Sensing*, 24(9):1925–1947, 2003.

M. Z. Nascimento, A. S. Martins, L. A. Neves, R. P. Ramos, E. L. Flores, and G. A. Carrijo. Classification of masses in mammographic image using wavelet domain features and polynomial classifier. *Expert Systems with Applications*, 40 (15):6213–6221, 2013.

National Breast and Ovarian Cancer Centre (NBOCC). Breast cancer risk factors: A review of the evidence. Technical report, National Breast and Ovarian Cancer Centre, Surry Hills, NSW, 2009.

National Cancer Institute. New guidance for personalized breast cancer screening. Technical Report 16, National Cancer Institute, 2011.

S. Obenauer, K. P. Hermann, and E. Grabbe. Dose reduction in full-filed digital mammography: An anthropomorphic breast phantom study. *The British Journal of Radiology*, 76(907):478–482, 2003.

T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585, 1994.

T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 701–706, 2002.

A. Oliver, J. Freixenet, R. Martí, J. Pont, E. Pérez, E. R. E. Denton, and R. Zwiggelaar. A novel breast tissue density classification methodology. *IEEE Transaction Information Technology In Biomedicine*, 12(1):55–65, Jan 2008.

G. Parmigiani, D. Berry, and O. Aguilar. Determing carrier probabilities for breast cancer- susceptibility genes BRCA1 and BRCA2. *The American Journal of Human Genetics*, 62(1):145–158, Jan 1998.

N. Pashayan, S. W. Duffy, S. Chowdhury, T. Burton, D. E. Neal, D. F. Easton, R. Eeles, and P. Pharoah. Polygenic susceptibility to prostate and breast cancer: Implications for personalized screening. *British Journal of Cancer*, 104(10): 1656–1663, May 2011.

S. Petroudi and M. Brady. Breast density segmentation using texture. In S. M. Astley, M. Brady, C. Rose, and R. Zwiggelaar, editors, *Digital Mammography*, volume 4046 of *Lecture Notes in Computer Science*, pages 609–615. Springer Berlin Heidelberg, 2006.

S. Petroudi and M. Brady. Breast density characterization using texton distributions. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 5004–5007, 2011.

S. Petroudi, T. Kadir, and M. Brady. Automatic classification of mammographic parenchymal patterns: A statistical approach. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 1, pages 798–801, Sep 2003.

D. A. Picca and E. S. Paredes. Calcifications in the breast: A radiologic perspective. *Applied Radiology*, 32(9):30–37, Sep 2003.

E. D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. K. Baum, S. Acharyya, E. F. Conant, L. L. Fajardo, L. Bassett, C. D'Orsi, R. Jong, and M. Rebner. Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine*, 353(17):1773–1783, 2005.

R. Polikar. Ensemble based systems in decision making. *Circuits Systems Magazine, IEEE*, 6(3):21–45, 2006.

R. Quevedo, López-G. Carlos, J. M. Aguilera, and L. Cadoche. Description of food surfaces and microstructural changes using fractal image texture analysis. *Journal of Food Engineering*, 53(4):361–371, 2002.

R. M. Rangayyan. *Biomedical Image Analysis*. the Biomedical Engineering. CRC Pres LLC, 2005.

N. P. Rath, P. Pattnaik, and J. Samantaray. Depth analysis of monocular natural scenes using gray level co-occurrence matrix. In *Intelligent and Advanced Systems (ICIAS), 2012 4th International Conference on*, volume 1, pages 319–323, 2012.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. Technical report, Arizona State University, 2008.

G. Rellier, X. Descombes, F. Falzon, and J. Zerubia. Texture feature analysis using a gauss-Markov model in hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(7):1543–1551, 2004.

I. Resier, S. Lee, and R. M. Nishikawa. On the orientation of mammographic structures. *Medical Physics*, 38(10):5303–5306, Oct 2011.

I. Resier, B. A. Lau, R. M. Nishikawa, and P. R. Bakic. A directional small-scale tissue model for an anthropomorphic breast phantom. In A. D. A. Maidment, P. R. Bakic, and S. Gavenonis, editors, *Breast Imaging*, volume 7361 of *Lecture Notes in Computer Science*, pages 141–148. Springer Berlin Heidelberg, 2012.

M. Sadeghi, T. Lee, D. Mclean, H. Lui, and S. M. Atkins. Global pattern analysis and classification of dermoscopic images using textons. *Proc. SPIE*, 8314:83144X–83144X–6, 2012.

A. F. Saftlas, J. N. Wolfe, R. N. Hoover, L. A. Brinton, C. Schairer, M. Salane, and M. Szklo. Mammographic parenchymal patterns as indicatiors of breast cancer risk. *American Journal of Epidemiology*, 129(3):518–526, 1989.

B. Sahiner, H-P. Chan, N. Petrick, M. A. Halvie, and M. M. Goodsitt. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Medical Physics*, 25(4):516–526, April 1998.

X. Sastre-Garau, P. Genin, A. Rousseau, A. A. Ghuzlan, A. Nicolas, P. Freneaux, C. Rosty, B. Sigal-Zafrani, J. Couturier, J-P. Thiery, H. Magdelenat, and A. Vincent-Salomon. Increased cell size and Akt activation in HER-2/neu-overexpressing invasive ductal carcinoma of the breast. *Histopathology*, 45:142–147, Dec 2004.

G. Schaefer and N. P. Doshi. Multi-dimensional local binary pattern descriptors for improved texture analysis. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2500–2503, 2012.

R. E. Schapire. The strength of weak learn ability. *Machine Learning*, 5(2):197–227, Jun 1990.

C. Schmid. Constructing models for content-based image retrieval. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 39–45, 2001.

J. T. Schousboe, K. Kerlikowske, A. Loh, and S. R. Cummings. Personalizing mammography by breast density and other risk factors for breast cancer: Analysis of health benefits and cost-effectiveness. *Annals of Internal Medicine*, 155(1):10–20, 2011.

J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, January 2009.

P. Skaane and A. Skiennald. Screen-film mammography versus full-field digital mammography with soft-copy reading, randomized trial in a population-based screening program–the Oslo II study. *Radiology*, 232(1):197–204, 2004.

Surveillance Epidemiology and End Results (SEER). *SEER stat fact sheets: breast cancer*, Jan 2012. http://seer.cancer.gov/statfacts/html/breast.html.

J. A. Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14(2):109–121, 1979.

P. G. Tahoces, J. Correa, M. Souto, L. Gómes, and J. J. Vidal. Computer-assisted diagnosis: The classification of mammographic breast parenchymal patterns. *Physics in Medicine and Biology*, 40(1):103, 1995.

X-O. Tang. Texture information in run-length matrices. *Image Processing, IEEE Transactions on*, 7(11):1602–1609, 1998.

P. Taylor, H. Patts, L. Wilkinson, and R. G. Wilson. Impact of CAD with full field digital mammography on workflow and cost. In J. Marti, A. Oliver, J. Freixenet, and R. Marti, editors, *Digital Mammography*, volume 6136 of *Lecture Notes in Computer Science*, pages 1–8. Springer Berlin Heidelberg, 2010.

H. J. Teertstra, C. E. Loo, M. A. van den Bosch, H. van Tinteren, E. J. Rutgers, S. H. Muller, and K. G. Gilhuijs. Breast tomosynthesis in clinical practice: Intial results. *European Radiology*, 20(1):16–24, Jan 2010.

W. Teh and A. R. M. Wilson. The role of ultrasound in breast cancer screening. A consensus statement by the European Group for breast cancer screening. *European Journal of Cancer*, 34(4):449–450, Mar 1998.

D. B. Thomas, J. Whitehead, C. Dorse, B. A. Threatt, F. I. Gilbert, A. J. Present, and T. Carlile. Mammographic calcifications and risk of subsequent breast cancer. *Journal of the National Cancer Institute*, 85(3):230–235, 1993.

J. A. Tice, S. R. Cummings, R. S. Bindman, L. Ichikawa, W. E. Barlow, and K. Kerlikowske. Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Annals of Internal Medicine*, 148(5):337–347, 2008.

J. Tyrer, S. W. Duffy, and J. Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 23(7):1111–1130, April 2004.

M. Unser. Sum and difference histograms for texture classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(1):118–125, 1986.

C. H. G. Van, J. H. Hendriks, R. Holland, N. Karssemeijer, J. D. Otten, H. Staatman, and A. L. Verbeek. Changes in mammographic breast density and concomitant changes in breast cancer risk. *European Journal of Cancer Prevention : The Official Journal of the European Cancer Prevention Organisation (ECP)*, 8(6): 509–515, Dec 1999.

M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages 691–698, Jun 2003.

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

M. Varma, A. Zisserman, and J-M. Geusebroek. Texture classification. Technical report, Visual Geometry Group, Department of Engineering Science, University of Oxford, March 2007.

S. Vinnicombe, S. M. Pinto Pereira, V. A. McCormack, S. Shiel, N. Perry, and I. M. Dos Santos. Full-field digital versus screen-film mammography: Comparison within the UK breast screening program and systematic review of published data. *Radiology*, 251(2):347–358, 2009.

L. Wang and D-C. He. Texture classification using texture spectrum. *Pattern Recognition*, 23(8):905–910, 1990.

X-W. Wang, D. Lederman, J. Tan, X-H. Wand, and B. Zheng. Computerized prediction of risk for developing breast cancer based on bilateral mammographic breast tissue asymmetry. *Medical Engineering and Physics*, 33(8):934–942, 2011.

J. Wei, H-P. Chan, M. A. Helvie, M. A. Roubidoux, B. Sahiner, L. M. Hadjiiski, C. Zhou, S. Paquerault, T. Chenevert, and M. M. Goodsitt. Correlation between mammographic density and volumetric fibroglandular tissue estimated on breast mr images. *Medical Physics*, 31(4):933–942, April 2004.

J. Wei, H-P. Chan, Y-T. Wu, C. Zhou, M. A. Helvie, A. Tsodikov, L. M. Hadjiiski, and B. Sahiner. Association of computerized mammographic parenchymal pattern measure with breast cancer risk: A pilot case-control study. *Radiology*, 260(1): 42–49, 2011.

S. R. Wellings, H. M. Jensen, and R. G. Marcum. An atlas of subgross pathology of the human breast with special reference to possible precancerous lesions. *Journal of the National Cancer Institute*, 55(2):231–273, 1975.

J. Whitehead, T. Carlile, K. J. Kopecky, D. J. Thompson, J. F. Gilbert JR, A. J. Present, B. A. Threatt, P. Krook, and E. Hadaway. The relationship between Wolfe's classification mammograms, accepted breast cancer risk factors, the incidence of breast cancer. *American Journal of Epidemiology*, 122(6):994–1006, 1985.

G. J. Whitman and T. M. Haygood, editors. *Digital Mammography: A Practical Approach*. Cambridge University Press, New York, 2012.

J. N. Wolfe. Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer*, 37(5):2486–2492, May 1976a.

J. N. Wolfe. Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, 126(6):1130–1137, June 1976b.

J. N. Wolfe, A. F. Saftlas, and M. Salane. Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: A case-control study. *AJR American Journal of Roentgenology*, 148(6):1087–1092, Jun 1987.

R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, and J. Manglon et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(21):789–792, 1995.

J. Zhang, H. Zhao, and J-M. Liang. Continuous rotation invariant local descriptors for texton dictionary-based texture classification. *Computer Vision and Image Understanding*, 117(1):56–75, 2013.

B. Zheng, J. H. Sumkin, M. L. Zuley, X-W. Wang, A. H. Klym, and D. Gur. Bilateral mammographic density asymmetry and breast cancer risk: A preliminary assessment. *European Journal of Radiology*, 81(11):3222–3228, 2012.

Y-J. Zheng, Y. Wand, B. M. Keller, E. Conant, J. C. Gee, and D. Kontos. A fully-automated software pipeline for integrating breast density and parenchymal texture analysis for digital mammograms: Parameter optimization in a case-control breast cancer risk assessment study. *Proc. SPIE*, 8670:86701B–86701B–7, 2013.

C. Zhou, H-P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, B. Sahiner, and L. M. Hadjiiski. Computerized image analysis: Estimation of breast density on mammograms. *Medical Physics*, 28(6):1056–1069, June 2001.

S-C. Zhu, C-E. Guo, Y-Z. Wang, and Z-J. Xu. What are textons? *International Journal of Computer Vision*, 62(1-2):121–143, 2005.

H. M. Zorbas. Breast cancer screening. *The Medical Journal of Australia*, 178(12): 651–652, 2003.