THE UNIVERSITY OF CHICAGO

MACHINE LEARNING ON MEDICAL IMAGING FOR BREAST CANCER RISK ASSESSMENT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MEDICAL PHYSICS

BY

KAYLA RAE ROBINSON

CHICAGO, ILLINOIS

JUNE 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Several individuals had a profound impact on the work presented in this dissertation, and my journey through graduate school. My advisor and mentor, Dr. Giger, demonstrated unconditional support and encouragement as my research skills developed. From her, I learned how to be not only a good scientist, but a good writer and science communicator.

The members of my thesis committee contributed immeasurably to the rigor of this dissertation. I thank Dr. Hui Li, Dr. Ingrid Reiser, Dr. Sam Armato, and Dr. David Schacht for their great discussions and conversations on the goals and methods of this work. They held this work to the highest scientific standard and provided thoughtful and engaging feedback throughout the inception and the actualization of the project.

Each member of the Giger Lab contributed in a special way to this project and to my experiences as a graduate student. Dr. Hui Li taught me about research and experience through our conversations, and his constant and honest encouragement and advice. He and Li Lan contributed to my enjoyment of working on this project at the University of Chicago. Additionally, members including Karen Drukker, Heather Whitney, John Papaioannou and Sasha Edwards helped provide valuable technical assistance and feedback throughout my PhD. Each helped improve my presentation and communication skills through their active engagement in providing feedback as I prepared for presentations throughout my graduate school experience.

Each member of the Committee on Medical Physics at the University of Chicago supported me in some way. Specifically, I would like to thank Ruth Magaña for her warm guidance and assistance. I would like to thank Julie Hlavaty for her help in enabling my travel to academic conferences throughout my time at the University of Chicago. I would like to thank Hoang Ngo for his eagerness to help guide me through my final year as a graduate student.

This dissertation would not have been possible without the support of the following grants: NIBIB of the NIH grant number T32 EB002103, the NCI of the NIH grant number NIH QIN U01 CA 195564, U01 189240, F31 CA 228247, the Paul C. Hodges Research Award, and the Lawrence H. Lanzl Graduate Fellowship Award.

Lastly and most importantly, I dedicate this dissertation to my mother, father, sister and husband. In particular, my father's cancer diagnosis during the final year of my PhD brought a personal significance to this, and all cancer research efforts. Observing, through his experiences, the impact that diagnostic imaging and treatment technologies reaffirmed my commitment to making quality contributions to the field of cancer research.

ABSTRACT

Breast cancer was the most frequently diagnosed cancer in women in 2018, and this trend is expected to continue in years to come. Imaging-based cancer screening, including full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT), plays a significant role in the early detection and diagnosis of breast cancer, as well as in cancer risk assessment. Breast cancer risk assessment is important to breast cancer screening protocols as it aims to identify women at an elevated risk of breast cancer who may benefit from specialized screening. Family history and genetics have been shown to play a large role in risk assessment and are typically key factors used to identify candidates for specialized screening. However, there is growing interest in the potential of computer-aided image evaluation, or radiomics, to provide additional information for identifying these populations.

Radiomic features are quantitative image descriptors of human-defined features such as coarseness or contrast. While their utility has been widely demonstrated on homogeneous image sets acquired under consistent imaging conditions, inconsistent imaging conditions may introduce feature variations. Parameters that can cause differences in radiomic feature values on mammography may include x-ray beam quality, detector material, and post-processing algorithms. However, consideration is not typically given to radiomic feature robustness in radiomics studies. Additionally, there does not exist consensus in the field on methodology and metrics for characterizing feature robustness. To fill this need, this dissertation proposes novel metrics for characterizing radiomic feature robustness on image sets in which the population underwent imaging on mammography units from two different vendors.

Having proposed metrics for characterizing robustness, this dissertation then incorporates these metrics into a two-stage feature selection scheme. The goal of this scheme is to identify a set of robust, non-redundant, and descriptive features specific to a given clinical task. Stage one involves hierarchical clustering and robustness metric comparison to identify robust and non-redundant features as feature candidates in an unsupervised manner. Stage two involves classification evaluation of the feature candidates identified in stage one to select features that are also descriptive in the clinical task at hand. This method was demonstrated on paired data acquired on equipment from two different vendors, however this method could be used in cases without paired data by using robustness metrics that compare populations, such as the Kolmogorov-Smirnov test statistic. This two-stage method was demonstrated on FFDM in the task of classifying the risk of breast cancer among patients with no abnormalities detected on mammographic screening exams.

Given recent advances in computer algorithms and machine learning methodologies, this dissertation then presents work investigating the use of deep learning in medical image analysis. Transfer learning was investigated for evaluation of FFDM and DBT images. Transfer learning involves extraction of unintuitive features using a pre-trained convolutional neural network. These features are then used for classification in a conventional classifier, such as a support vector machine. In this dissertation, we applied transfer learning to the task of characterizing lesions as malignant or benign. Performance was compared between DBT and FFDM images as network inputs. This comparison is clinically relevant given the growing adoption of DBT at medical centers throughout the country.

An additional area where machine learning may add value to medical image analysis is in the evaluation of temporal sequences of screening mammograms. It is standard clinical procedure to

consider prior mammograms in the evaluation of current mammograms when that data is available. Thus, this motivated the incorporation of antecedent images into computer analysis. Long short-term memory (LSTM) networks were investigated for use in evaluating temporal sets of mammograms for classifying future lesions as malignant or benign. LSTM networks were chosen to use in temporal image sequence analysis as they have performed well in previous natural language processing and medical imaging studies. By incorporating temporal sequences of images, patterns over time may potentially lead to improvements in classification performance.

This dissertation presents the following results. First, metrics descriptive of feature equivalence, variability and correlation are proposed and used to characterize the robustness of a set of features on FFDM images acquired of an asymptomatic population of women over mammography unit vendor. This investigation found that box counting fractal dimension, Minkowski fractal dimension, and power law beta features were relatively robust across vendors. These features demonstrated high correlation over the two units examined, similar feature intensities, and small variability. Given the proposed robustness metrics, a two-stage feature selection method was proposed and used in the task of predicting the risk of breast cancer in patients with no detectable mammographic abnormalities. A monotonically decreasing trend was observed in classification performance as feature robustness restrictions were loosened. Additionally, comparison between the proposed two-stage method and a feature harmonization method (ComBat) demonstrated significant differences in performance on intra-vendor comparisons. Thus, these results suggest value in considering robustness during feature selection when using heterogeneous data. Investigations into transfer learning for lesion characterization found that, for mass and architectural distortion lesions, classification performance using a key

slice from DBT was higher than classification performance using a FFDM image of the same lesion. We failed to demonstrate a significant difference in comparing classification performance between FFDM images and synthesized 2D images derived from the DBT imaging data. Thus, this provides evidence in support of the efficacy of computer-aided diagnosis on DBT images. Additionally, investigation into temporal mammography sequence analysis using deep learning revealed higher performance of LSTM networks over conventional evaluation of a single time point in the task of classing future lesions as malignant or benign. This suggests that there is potential value in considering antecedent images in a deep learning framework when performing computerized image analysis for classifying breast cancer lesions.

This work is clinically significant as medical imaging for breast cancer screening is a common practice across the world, and an improvement in the analysis of screening images has the potential to impact a large number of people. This work presents a method for the incorporation of robustness considerations into radiomics studies to improve their generalizability of these studies. Additionally, the methods presented in this dissertation may improve our ability to anticipate the future incidence of breast cancer using medical images that are already collected as part of standard practice.

Keywords: breast cancer, computer-aided diagnosis, risk assessment, mammography, tomosynthesis, deep learning, convolutional neural networks, recurrent neural networks, long short-term memory networks.

CHAPTER 1

1. Introduction

Breast cancer is the most common cancer in women, and the second leading cause of death in women. The American Cancer Society projects 878,980 new cancer cases and 286,010 cancer deaths in women in the United States in 2019.[1] Breast cancer is expected to make up 266,120 of these new cancer diagnoses and is expected to claim the lives of 40,920 women.[1] Thus, breast cancer is projected to make up about 30% of new cancer cases among women and 14% of cancer deaths among women, second only to lung and bronchus cancer.[1]

While improvements in breast cancer treatment are being continuously developed, early detection through routine screening has been responsible for substantial reduction of breast cancer-related mortality.[2] Mammography first emerged as an accepted technology in the 1960s, and it was primarily used to aid in the diagnosis of complex cases.[3] After technological improvements, mammography was adopted for screening, thus has contributing in part to the reduction in breast cancer mortality.[4,5]

Breast cancer diagnosis in the United States is composed of three stages. First, screening techniques including mammography and physical examination are employed to detect abnormalities. Second, the abnormality is diagnosed through biopsy and further imaging such as ultrasound and magnetic resonance imaging (MRI). Third, if found to be malignant, the lesion will be characterized by subtype and will be staged based on tumor size and extent of invasion.[3]

Randomized in various countries have evaluated the impact of screening mammography on breast cancer mortality among women.[6–13] The majority of these studies reported a relative risk below one which suggests that mortality was reduced by the introduction of screening

mammography (Table 1.1). Furthermore, meta-analysis has confirmed the benefit of screening mammography in women over age 50.[14,15]

Table 1.1. Results of randomized clinical trials establishing the efficacy of screening mammography. While more recent studies exist, the studies listed here played a role in instituting mammography for breast cancer screening.

Trial	Entry Years	Number of Women	Relative Risk (95% CI)
HIP Trial[6]	1963-69	60,696	0.77 (0.61-0.97)
Malmö[7]	1976-86	41,478	0.81 (0.62-1.07)
Two-County[8]	1977-85	133,065	0.70 (0.58-0.85)
Stockholm[9]	1981-85	59,176	0.71 (0.4-1.2)
Gothenburg[10]	1982-88	49,553	0.86 (0.54-1.37)
Edinburgh[11]	1978-85	54,671	0.82 (0.61-1.11)
Canada NBSS I[12]	1980-87	50,430	1.10 (0.78-1.54)
Canada NBSS II[13]	1980-87	39,405	0.87 (0.62-1.52)

While mammograms provide quantitative representations of breast tissue, interpretation of these images is typically performed qualitatively by radiologists. Radiologists use their training and experience to visually identify signs of malignancy, including microcalcifications, architectural distortions, asymmetrical densities, masses, and densities that have developed or changed in size or appearance since the previous exam. Computer aided-detection (CADe) systems have been designed to complement the qualitative impressions of radiologists.

While mammographic screening has many benefits, there are also concerns about its efficacy. Namely, concerns exist about the risk of carcinogenesis arising from repeated exposure to ionizing radiation. While an excess of breast cancers has been observed in populations of women exposed to high doses of radiation, research suggests that the risk of carcinogenesis from the low dose associated with screening mammography is negligible relative to the benefits of

mammography.[16] Therefore, radiation dose is not typically a deterrent of the use of screening mammography.

While mammography has improved early detection of breast cancer, its effectiveness is limited by its 2D nature. This concern is of particular importance for women with dense breasts. Digital breast tomosynthesis (DBT) was developed to address this limitation. DBT uses projection images acquired at various angles around the breast to construct a pseudo 3D reconstruction of the breast. DBT has been shown to reduce missed cancers[17–19] and reduce the screening recall rate.[20,21] As a result, its adoption has been growing throughout the United States.

While FFDM and DBT are appropriate screening methods for women at an average risk of breast cancer, different imaging methods exist for women at an elevated risk of breast cancer. For example, MRI and ultrasound can be used to provide complementary information to mammographic images to aid in lesion detection and characterization.[22] However, these imaging methods are more expensive and resource-intensive than mammography. Thus, national guidelines advise limiting MRI to high-risk groups.[23] Current research is investigating the potential use of MRI for use in population screening programs.[24]

1.1. Breast Imaging Modalities

1.1.1. Mammography

Typical screening mammograms consist of at least two x-ray images of each breast. One is taken from the mediolateral oblique (MLO) view, and one is taken from the craniocaudal (CC) view. Diagnostic mammography may also include specialized views and magnifications to image suspicious findings with higher resolution.

Mammography produces 2D projection images of the breast using low-energy x-rays. Low energies (20-30 keV) are used to enhance tissue discrimination through the exploitation of the photoelectric effect. The probability of photon interaction through the photoelectric effect is proportional to the cube of the atomic number of a material, thus absorption differences are accentuated at low energies where the photoelectric effect is the dominant mode of photon interaction. The particular anode material used may vary across vendors, causing differences in x-ray spectra. A common unit geometry is illustrated in Figure 1.1.

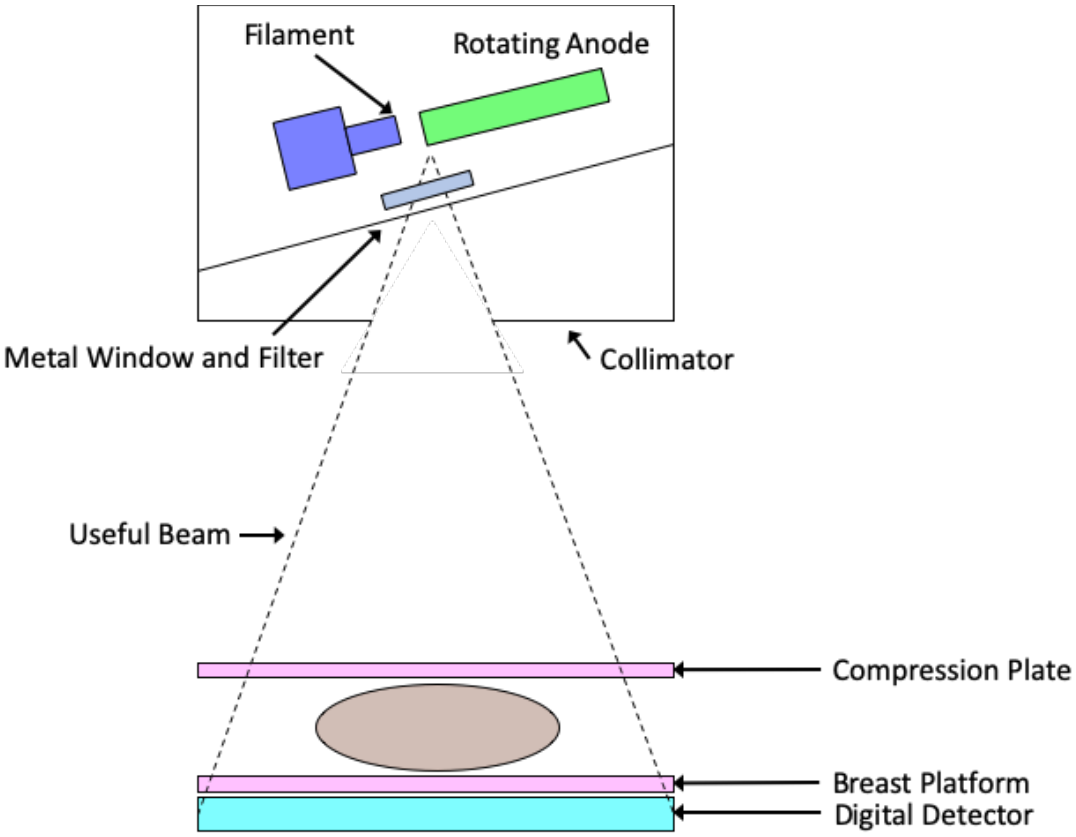


Figure 1.1. Diagram of a typical mammographic x-ray tube and breast support.[25] X-rays are produced by a rotating anode and are passed through a filter and collimator.

After passing through tissue, transmitted x-rays are collected on a digital detector. The detector's receptors can be either direct or indirect. Direct receptors have a photoconductor coupled to the active matrix array. When a photon interacts with the photoconductor, charge is collected and stored in capacitors until it is read out by the electronics of the unit. Indirect receptors function differently in that they use an intensifying screen as opposed to a photoconductor. Charge is collected by an array of photodetectors and is stored until the detector is read out by the unit circuitry.[25]

After image acquisition, post-processing algorithms are applied to enhance certain features of the mammogram to facilitate review by a radiologist. One of the key objectives of these processing algorithms is to manipulate fine differences in image contrast to help differentiate between tissue types. Commonly, image processing algorithms include techniques such as manual intensity windowing, histogram-based intensity windowing, mixture-model intensity windowing, contrast-limited adaptive histogram equalization, unsharp masking, peripheral equalization and Trex processing.[26,27] Each processing algorithm affects pixel intensities in different ways, making quantitative comparisons between differently processed images difficult. Mammography vendors implement different processing algorithms causing quantitative image differences across their images.

1.1.2. Tomosynthesis

DBT has emerged as a promising modality to improve breast cancer screening sensitivity and accuracy.[17] DBT yields pseudo-3D images by rotating an x-ray tube in a partial arc around the breast while acquiring projection images. This principle is illustrated in Figure 1.2.

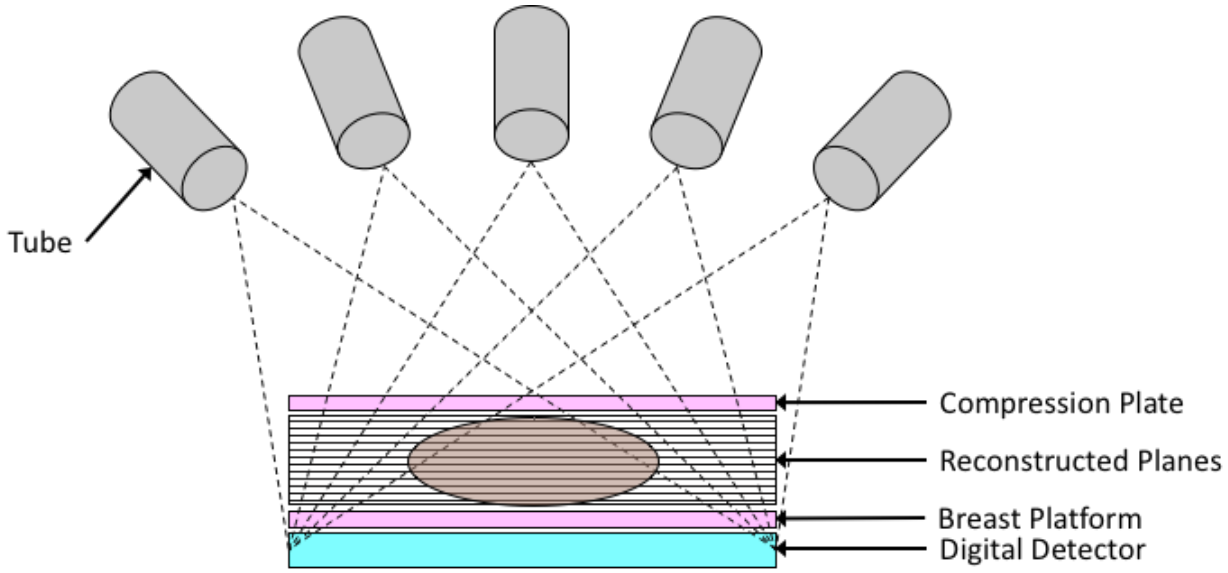


Figure 1.2. Diagram illustrating the mechanism of digital breast tomosynthesis acquisition. The x-ray tube is moved in a partial arc around the compressed breast, and projection images are either acquired continuously or in a step-and-shoot manner.

Each projection image is acquired in a similar manner to those of conventional full-field digital mammography (FFDM) but with a reduced radiation dose. The scanning angle, number of projections and other parameters can vary across manufacturers. Imaging parameters for three Food and Drug Administration-approved units are listed in Table 1.2.

Table 1.2. Image acquisition parameters for digital breast tomosynthesis units.[28]

Manufacturer	Hologic	GE Healthcare	Siemens Healthcare
Model	Selenia Dimensions	SenoClaire	Mammomat Inspiration
Scanning angle (°)	15	25	50
Projections	15	9	25
Scanning time (s)	4	7	24
Tube motion	Continuous	Step and shoot	Continuous
Reconstruction	Filtered backprojection	Iterative	Filtered backprojection

A growing number of studies have shown that tomosynthesis significantly reduces screening recall rates and increases cancer detection rates.[29–32] By providing volume data as opposed to single projection images, DBT gives a clearer visualization of regions of interest and reduces the impact of overlaying tissue. DBT is expected to be particularly useful for women with dense breasts, for whom overlaying parenchymal tissue may obscure breast lesions.[33] However, the human observer studies that have been used to suggest the efficacy of DBT involve inherently qualitative measurements and subjective interpretations. The objectivity of computer vision methods may therefore help inform image interpretation.

Radiologists often review DBT images using a cine loop in which reconstructed planes of the breast are displayed like frames of a video. While the reconstructed planes are not 3D, they portray some depth information that is useful in the detection and diagnosis of breast lesions. In order to maintain the benefits of 2D breast imaging, synthesized 2D image are also typically constructed from the projection images. However, rather than mimicking FFDMs, the synthesized 2D images are processed to emphasize suspicious features and improve their conspicuity to radiologists.[34]

While the physics principles of tomosynthesis data acquisition are similar to those of FFDM, the data are processed differently and have different characteristics. These varying characteristics make the review of DBT images by a radiologist different, often requiring additional training. Therefore, DBT and FFDM are both informative for detecting suspicious lesions, and they provide data in different formats which may need to be handled differently in quantitative analyses.

1.1.3 MRI and Ultrasound

While FFDM and DBT are standard screening modalities for women with an average risk of breast cancer, specialized screening and diagnostic imaging exist for high-risk populations. MRI and ultrasound are common choices for follow-up imaging because they provide complementary information to mammography and involve no radiation dose.

In ultrasound breast imaging, high-frequency sound waves are transmitted to the breast tissue by a transducer, and the reflection and refraction behavior of the sound waves are measured. This gives complementary information to that obtained through mammography.[17] For example, while solid and fluid-filled abnormalities may look similar in x-ray imaging, their appearances in ultrasound are quite different. Therefore, ultrasound imaging is useful for some patients in the differentiation of abnormality types.

Breast MRI may be used for screening of populations at an elevated risk of breast cancer. In breast MRI, a contrast agent is injected intravenously, and MRI images are acquired at time intervals following injection. Breast DCE-MRI provides physiological information not given by mammography, however these procedures can be expensive and invasive. Therefore, MRI is typically reserved for women at a high risk of breast cancer for whom mammography may not be sufficient.[35]

1.2 Computer-Aided Detection, Diagnosis and Risk Assessment

While radiologist performance in the task of detecting and distinguishing breast cancer has helped to reduce breast cancer mortality, human interpretation of medical images is not perfect. In some cases, missed cancers may be caused by physical limitations of the imaging modality such as image resolution, signal-to-noise ratio and contrast. However, the majority of missed

cancers have been attributed to interpretation errors by radiologists.[36] Human errors in image interpretation can be caused by both perception errors and cognitive errors. Perception errors occur when an abnormality is not seen in an image, and cognitive errors occur when an abnormality is seen but incorrectly interpreted.[36] Computer analysis methods have been developed to address these errors by assisting human observers in both the interpretation and diagnosis of radiographic abnormalities.

Computer-aided detection and diagnosis (CAD) is an image analysis scheme that automatically or semi-automatically identifies lesion candidates and characterizes their likelihood of malignancy. CAD has become widespread in both the research and clinical arenas for diseases of the chest, breast and other organs.[37–50] By producing information that is complementary to that collected by a radiologist review of an image, CAD stands to improve the overall cancer detection and diagnosis workflow.

1.2.1. Breast Cancer Computer-Aided Detection

CAD has shown promise in clinical tasks concerning the detection of abnormalities in medical images. In computer-aided detection (CADe) for breast disease, radiomic features are quantitatively calculated throughout the breast. These features are used to identify regions that contain abnormalities or are otherwise suspicious and warrant further inspection. The radiologist is then alerted to these regions for further review.[51] In this way, the computer acts as a second reader in assisting the radiologist with his or her interpretation of medical images.

CADe systems have been incorporated into clinical practice for breast cancer screening.[52] Computer assistance is particularly well suited for screening programs due to the large proportion of normal exams relative to abnormal exams. Computers may thus increase the

sensitivity of image interpretation. However, this increase in sensitivity is typically coupled with a reduction in specificity. Therefore, research efforts have focused on maximizing the improvement in sensitivity while minimizing the detriment to specificity. Prospective studies following clinical implementation have demonstrated the effectiveness of CADe systems for cancer detection in breast mammography.[37,38,53]

1.2.2. Breast Cancer Computer-Aided Diagnosis

Computer-aided methods are also used in the diagnosis of cancer through the characterization of suspicious image regions. Computer-aided diagnosis (CADx) involves the differentiation of abnormalities across disease presence, types or states. In this way, CADx performs a type of "virtual biopsy" to assist physician recommendations.[54] By assisting radiologist evaluation of the probability of malignancy, CADx systems aim to improve sensitivity and specificity while also reducing intra- and inter-observer variability. The development of CADx systems seek to reduce the number of benign lesions sent for biopsy without missing malignant lesions.

1.2.3. Breast Cancer Computer-Aided Risk Assessment

Beyond cancer detection and diagnosis, quantitative radiomic features have also been used to estimate an individual's risk of breast cancer. For example, researchers have used radiomics to predict contralateral cancer, high-risk clinical factors, and future cancer based on antecedent imaging.[55–64]

Typically, personalized risk models are based on characteristics such as age, family history and certain genetic mutations including BRCA1/BRCA2. Developments in CAD suggest that parenchymal texture may also help inform risk.

Several studies have used quantitative measures of parenchymal texture to evaluate the risk of cancer in asymptomatic females.[55,57,60,63,65–69] These studies use radiomic texture features including fractal dimension,[65] power-law spectral analysis,[60] absolute gray-level, gray-level histogram analysis, neighborhood gray tone difference matrix (NGTDM), and gray-level co-occurrence matrix (GLCM) in their analyses.[70]

Screening recommendations for women at a high risk of breast cancer have been enacted by agencies such as the American Cancer Society. These recommendations raise the potential impact of breast cancer risk assessment because the actionable recommendations may lead to improved early detection of disease in high-risk cohorts as the improved identification of women at a high risk of breast cancer may allow them to take advantage of specialized screening options.[71] Continued research in risk assessment is motivated by the need to best utilize available specialized screening modalities.

1.2. Risk Factors for Breast Cancer

Previously published studies have identified several factors that may elevate an individual's risk of developing breast cancer. Some such factors are related to an individual's demographics and lifestyle, while others are related to the appearance of an individual's breast tissue on mammography.

1.2.1. Demographics

Certain physical and lifestyle characteristics of an individual are demonstrated to have a significant impact on that individual's risk of developing breast cancer. Reproductive factors linked to an elevated risk of developing breast cancer include an early age at menarche, late age

at menopause, and late age at first full-term pregnancy.[72] Furthermore, research has suggested that long-term use of combined estrogen/progestin hormone replacement therapy and long-term use of oral contraceptives also increase the risk of breast cancer.[72]

As a woman's age increases, so does her risk of breast cancer. The incidence of breast cancer is estimated to double with every ten years of age until menopause. After menopause, the incidence still increases, but at a slower rate.[73] Genetics and family history also play a substantial role in determining a woman's risk of developing breast cancer. It is estimated that up to 10% of breast cancer in Western countries can be attributed to genetic predisposition.[73] While scientists anticipate that not all genes linked to breast cancer have been identified, BRCA1 and BRCA2, genes located on the long arms of chromosomes 17 and 13, respectively, have been identified in a large proportion of high-risk families, establishing a link between this genetic mutation and breast cancer risk.[74]

Certain lifestyle characteristics are associated with a reduced risk of breast cancer. For example, physical exercise and a healthy body weight in adolescence were shown to be associated with delayed onset of breast cancer.[74] An epidemiological study reported strong evidence in support of a dose-response relation between alcohol consumption and the risk of breast cancer.[75] There is strong evidence that demographic information and lifestyle decisions are linked to an individual's risk of breast cancer.

1.2.2. Imaging Characteristics

Beyond lifestyle and demographic information, imaging characteristics of the breast parenchyma have also been shown to be indicative of a woman's risk of breast cancer. Both qualitatively

evaluated imaging characteristics, such as density or parenchymal pattern, and quantitative features, such as radiomic texture features, are indicative of risk of breast cancer.

The Breast Imaging and Reporting Data System (BI-RADS) was introduced in the late 1980s to standardize mammography reporting. One component of this reporting system was breast density reporting.[76] Density reporting was included to estimate each woman's breast cancer risk, as breast density has been shown to be a strong predictor.[56,59,62] Qualitatively evaluated texture patterns have also been shown to have predictive value in breast cancer risk assessment.[77]

Wolfe introduced a scheme of qualitatively evaluating the imaging characteristics of the breast not only based on relative amounts of fat, epithelial and connective tissue, but also on the prominence of ducts.[77] Categories based on breast density and parenchymal texture ("Wolfe Patterns") have been used in subsequent research and have shown to be significantly associated with future cancer incidence.[78–80]

1.3. Radiomic Features

When machine learning algorithms and computational power evolved in the 1980s, quantitative assessment of medical images emerged as an active area of research. The process of extracting large amounts of quantitative features from medical images has become known as the field of radiomics. The concept motivating use of radiomics for disease prediction is that radiomic features take distinctive values across disease forms, assisting in addressing various clinical questions.

Following early work in radiomics, a large number of algorithm-based features emerged and have been used in a wide range of clinical tasks. Specifically, this dissertation investigates

first-order histogram features, Fourier transform features, fractal dimension features, gray-level co-occurrence features, power law features, and edge gradient features. Each feature investigated in this dissertation is defined in Chapter 1.4.1-1.4.6.

1.3.1. First-Order Histogram Features

First-order histogram features describe the distribution of individual pixel intensity values in an image without concern for the spatial relationships between pixels.[81] They are constructed to describe the mean value, spread, and randomness of intensity by characterizing the histogram function. First-order radiomic features are calculated using a histogram constructed by dividing pixel intensities in an image into equally spaced bins and computing the proportion of pixels in each bin.[82] Note that binning methods affect the values of histogram=based radiomic features. Specifically, if different binning ranges are used on different images, then the resulting features would not be comparable for analyses.

In practice, first-order histogram features have been used in a wide range of studies. One such example was performed by Li et al., who used first-order histogram features, among others, to characterize a woman's risk of breast cancer.[66] Classifications were performed between women with a BRCA1/2 gene mutation and women with a low risk of breast cancer, and between unilateral cancer patients and women with a low risk of breast cancer.

Table 1.3. Summary of first-order histogram features explored in this dissertation, where $x_{i,j}$ is the value of the pixel in the i^{th} column and j^{th} row and n is the total number of pixels.

Feature Name	**Equation**	**Description**
Average	$Average = \frac{1}{n}\sum_{i}\sum_{j} x_{i,j}$	Mean value of all pixels contained in the region of interest
Maximum CDF	$0.95 = \int_{-\infty}^{MaxCDF} x(t)dt$	Gray value corresponding to the 95% region cutoff on the cumulative density function (CDF).
Minimum CDF	$0.05 = \int_{-\infty}^{MinCDF} x(t)dt$	Gray value corresponding to the 5% region cutoff on the CDF.
Balance	$Balance = \frac{MaxCDF - Average}{Average - MinCDF}$	Values less than one correspond to having an ROI that is skewed towards relatively denser values.
Seventy CDF	$0.70 = \int_{-\infty}^{SeventyCDF} x(t)dt$	Gray value corresponding to the 70% region cutoff on the CDF.
Thirty CDF	$0.30 = \int_{-\infty}^{ThirtyCDF} x(t)dt$	Gray value corresponding to the 30% region cutoff on the CDF.
Quasi Balance	$Quasi\ Balance = \frac{SeventyCDF - Average}{Average - ThirtyCDF}$	Values less than one correspond to having an ROI that is skewed towards relatively denser values.
Skewness	$Skewness = \frac{\frac{1}{n}\sum_{i=i}^{n}(x_i - \overline{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=i}^{n}(x_i - \overline{x})^2}}$	Symmetry of the histogram distribution

1.3.2. Fourier Transform Features

Features based on the Fourier transform of an image describe the spatial frequency distribution of an image.[83] The Fourier power spectrum is computed for discrete data by squaring the magnitude of the discrete Fourier transform of the image. The Fourier power spectrum describes the portion of a signal's power that occurs at various frequencies throughout the image. These features characterize the magnitude of the image pattern and its frequency content.[84]

The angular-dependent Fourier transform was computed and used in analysis to describe the directional dependence of parenchymal patterns. This was done by dividing the power spectrum into 16 sectors, and computing the root mean square of the power spectrum (FRMS) and the first moment of the power spectrum (FFMP) on each individual sector. While directionality in the breast parenchyma may not be particularly strong, directionality has shown utility in characterizing other diseases such as bone texture.[84]

A study by Huo et al. investigated the use of Fourier transform features, among other features, in the task of characterizing the risk of breast cancer.[85] In this study, features were used to characterize parenchymal texture to differentiate between age-matched BRCA1/2 and low-risk women.

Table 1.4. Fourier transform features explored in this dissertation, where F(u,v) is the 2D Fourier transform of the image.

FRMS	$\mathrm{FRMS} = \sqrt{\int_0^\infty \int_0^\infty \mathrm{\|F(u,v)\|^2 dudv}}$	Root mean square of the power spectrum
FFMP	$FFMP = \frac{\int_0^\infty \int_0^\infty \sqrt{u^2+v^2}\|F(u,v)^2\|dudv}{\int_0^\infty \int_0^\infty \|F(u,v)^2\|dudv}$	First moment of power spectrum

1.3.3. Fractal Dimension Features

Fractal dimension describes the self-similarity of an image.[86] Since its first suggestion by Mandelbrot, fractal dimension has seen use in various computer vision applications.[86] When used to describe the breast parenchyma, fractal dimension characterizes the complexity of the ductal tissue. Two methods of computing the fractal dimension include the box counting method and the Minkowski method.

Mandelbrot first described fractal dimension as a means of estimating the length of a coastline. He estimated the length of a coastline, L(ε), by constructing an area containing all points with distance ≤ ε from the coastline, and dividing this area by 2ε. As he decreased the distance ε, the estimated length of coastline, L(ε), increased. He found that Equation 1.1 held for many coastlines, where F and D are constants for a given coastline, and D is the box counting dimension.

$$L(\varepsilon) = F\varepsilon^{1-D} \tag{1.1}$$

Given a digital image of a surface, differential box counting estimates the fractal dimension of that surface. In this method first proposed by Chaudhuri and Sarkar, the image is partitioned into grids of varying scales and the number of boxes containing the image boundary line are counted. By comparing the number of boxes containing outline over the various grid scales, the fractal dimension can be estimated following Equation 1.2, where A(ε) is the surface area of the region of interest at scale size ε.[87]

$$D_{BC} = 2 - \lim_{\varepsilon \to 0} \frac{\log\left[A(\varepsilon)\right]}{\log\left[\varepsilon\right]} \tag{1.2}$$

The Minkowski dimension describes the same characteristics, however is computed in a different way. The Minkowski dimension is computed by applying morphological operators at various scales, including dilations and erosions, following Equations 1.3 and 1.4, where f is the

image, g is a structuring element, and r is the scale. Both the box counting and Minkowski methods for fractal dimension calculation were included in this dissertation, for completeness.

$$D_M(f) = \lim_{r \to 0} \frac{\log [V_g(r)/r^3]}{\log [1/r]} \tag{1.3}$$

where

$$V_g(r) = \sum_{i=1} \sum_{j=1} \{(f \oplus rg) - (f \otimes rg)\} \tag{1.4}$$

Fractal dimension features have been used in various clinical classification tasks, including evaluation of risk of breast cancer. A study by Li et al. investigated the use of fractal dimension features in the task of classifying high-risk women with BRCA1/2 gene mutations compared with low-risk women as judged by the Gail model.[65]

Table 1.5. Summary of fractal dimension features explored in this dissertation.

Feature Name	**Description**
Box counting Dimension	Calculated by fitting the slope using ε values of 0.1 mm, 0.2 mm, 0.4 mm, 0.8 mm, 1.6 mm, 3.2 mm, 6.4 mm, and 12.8 mm
Box counting Dimension 1	Calculated by fitting the slope using ε values of 0.1 mm, 0.2 mm, 0.4 mm and 0.8 mm.
Box counting Dimension 2	Calculated by fitting the slope using ε values of 1.6 mm, 3.2 mm, 6.4 mm, and 12.8 mm
Box counting Dimension 3	Calculated by fitting the slope using ε values of 0.1 mm, 0.2 mm, 0.4 mm, 0.8 mm, 1.6 mm, and 3.2 mm
Box counting Dimension 4	Calculated by fitting the slope using ε values of 0.1 mm, and 0.2 mm
Box counting Dimension 5	Calculated by fitting the slope using ε values of 3.2 mm, 6.4 mm, and 12.8 mm

1.3.4. Gray-level Co-occurrence Features (Second-Order Histogram)

While the previously discussed first-order histogram features describe the distribution of pixel intensities in an image, they do not characterize the spatial distribution of such pixels. Gray-level co-occurrence (GLCM) features address this shortcoming by characterizing both the spatial relationships and intensities between pixels in image regions. Haralick et al. suggests the construction of a matrix summarizing the spatial relationship between neighboring pixels.[70] Calculations can be performed on this matrix to derive various texture features which can then be used for image classification. First demonstrated on satellite images for classification of land-use type, these features have also seen utility in medical classifications.

The GLCM matrix is constructed by counting the number of nearest neighbor pixels consisting of various intensity combinations over an image. In constructing the GLCM matrix, neighborhood distance and connectivity, including the directionality of connectivity, can be specified. In this way, many different GLCM matrices can be constructed to describe a certain image. It is common to bin together ranges of pixel intensities in order to reduce the size of the GLCM matrix, thus reducing its sparsity. Figure 1.3 shows an example of an image and its corresponding GLCM. Typical texture features calculated using the GLCM are summarized in

Table 1.6, and their equations are given in terms of the GLCM matrix elements.

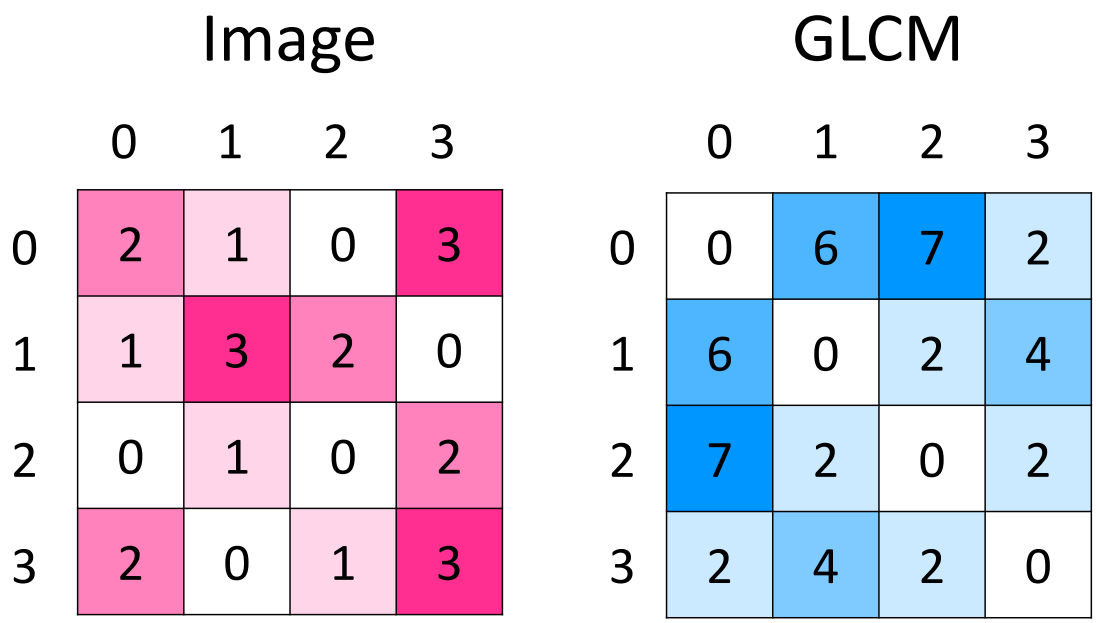


Figure 1.3. Example of an image and its corresponding GLCM matrix. In this example, 8-connectivity and nearest neighbor distance was used. No directionality restrictions were imposed, resulting in a symmetric GLCM.

Table 1.6. Summary of GLCM features explored in this dissertation, where $p(i,j)$ is the $(i,j)^{th}$ entry in a normalized gray-tone spatial-dependence matrix, $p_x(i)$ is the i^{th} entry in the marginal-probability matrix obtained by summing the rows of $p(i,j)$, N_g is the number of distinct gray-levels in the quantized image, and $p_y(j)$ is the j^{th} entry in the marginal-probability matrix obtained by summing the columns of $p(i,j)$.

Feature Name	Equation	Description
Contrast	$Contrast = \sum_{n=0}^{N_{g-1}} n^2 \left\{ \sum_{\substack{i=1 \\ \lvert i-j \rvert = n}}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\}$	Quantifies the combined effect of the gray-level differences in an ROI and the amount of local variation
Correlation	$Correlation = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$	Measure of gray-tone linear-dependencies in the image
Difference Entropy	$Diff\ Entropy = -\sum_{i=0}^{N_{g-1}} p_{x-y}(i) \log\{p_{x-y}(i)\}$ $p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g} \sum_{j=1}^{N_g} p(i,j)$	Measures the disorder related to the gray-level difference distribution of the image
Difference Variance	$Diff\ Variance = variance\ of\ p_{x-y}$ $p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g} \sum_{j=1}^{N_g} p(i,j)$	Measures the dispersion of the gray-level sum distribution of the image with regard to the mean
Energy	$Energy = \sum_i \sum_j (p(i,j))^2$	Measure of image homogeneity
Entropy	$Entropy = -\sum_i \sum_j p(i,j) \log(p(i,j))$	Measure of randomness of the image
Homogeneity	$Homogeneity = \sum_i \sum_j \frac{p(i,j)}{1 + (i-j)^2}$	Measures the smoothness of the gray-level distribution
IMC1	$IMC1 = \frac{HXY - HXY1}{\max\{HX, HY\}}$ $HXY = -\sum_i \sum_j p(i,j) \log(p(i,j))$	Characterizes the correlation between the probability distributions to quantify complexity of the image texture

	$HXY1 = -\sum_i \sum_j p(i,j) \log(p_x(i)p_y(j))$	
IMC2	$IMC2 = (1 - \exp[-2(HXY2 - HXY)])^{1/2}$ where $HXY2 = -\sum_i \sum_j p_x(i)p_y(j) \log(p_x(i)p_y(j))$	Quantifies the complexity of the image texture

Table 1.6, continued...

Maximum Correlation Coefficient	$Max\ CC = (2nd\ largest\ eigenvalue\ of\ Q)^{1/2}$ $Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$	Quantifies the complexity of the image texture
Sum Average	$Sum\ Average = \sum_{i=2}^{2N_g} ip_{x+y}(i)$	Measures the mean of the gray-level sum distribution of the image
Sum Entropy	$Sum\ Entropy = -\sum_{i=2}^{2N_g} \sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\}$	Measures the disorder related to the gray-level sum distribution of the image
Sum Variance	$Sum\ Variance = variance\ of\ p_{x-y}$	Measures the dispersion

A related set of features are the neighborhood gray tone difference (NGTDM) features. Like GLCM features, these features measure the variation between a central pixel and the surrounding pixels.[83]

1.3.5. Power-Law Beta

The power spectrum of an image is calculated from the Fourier transform of an image and describes the frequency content. In digital imaging, the discrete Fourier transform is used in the calculation of the power spectrum, since a digital image has a finite number of data samples. The power spectrum is defined by Equation 1.5, and the power law spectrum is defined by Equation 1.6. The exponent β is estimated by fitting the inverse exponential to the power spectrum, where F(u,v) is the Fourier transform of the image.[88,89]

$$P(f) = |F(u,v)|^2 \tag{1.5}$$

$$P(f) = \frac{B}{f^{\beta}} \tag{1.6}$$

Like the fractal dimension, power law features describe the complexity of an image. An image with coarser texture will have a larger β and smaller fractal dimension, and an image with finer texture will have a smaller β and larger fractal dimension.

Power law features have been shown to be descriptive of background parenchymal patterns, as well as predictive of a woman's genomics.[60,88,89] For example, Li et al. investigated the use of power spectral analysis in classifying BRCA1/2 gene mutation carriers from low-risk women.[60] Thus, this dissertation includes study of these feature types in examination of a woman's risk of cancer.

1.3.6. Edge Gradient Features

The gradient of pixel intensity over various distances between pixels is used to compute edge gradient features. Thus, edges are detected using small edge operator masks, and the distance-dependent texture description function is computed. The larger the average gradient, the finer the texture and the smaller the average gradient, the coarser the texture.[61,90] The maximum, minimum and standard deviation of the image also provide information about the scale of texture. Coarseness of the background parenchyma has been shown to be indicative of a woman's risk of breast cancer.[83] In general finer textures will have a high number of edges and high-contrast images will have large edge magnitudes.

Table 1.7. Summary of edge gradient features explored in this dissertation, where g(d) is the gradient between pixels at distance d, and n is the total number of pixels in the image.

Feature Name	Equation	Description
Mean of Gradient	$Mean\ Grad = \frac{1}{n}\sum g(d)$	Measure of the coarseness of the texture
Maximum of Gradient	$Max\ Grad = \max\left(g(d)\right)$	Greatest difference in adjacent pixel values
Minimum of Gradient	$Min\ Grad = \min\left(g(d)\right)$	Minimum difference in adjacent pixel values
Standard Deviation of Gradient	$Standard\ Dev\ Grad = \sqrt{\frac{\sum \lvert g(d) - \overline{g(d)}\rvert^2}{n}}$	Measure of the variation between pixel gradients

1.4. Technical Innovations of Presented Works

Radiomic texture analysis has been shown to perform well in classification tasks addressing clinical tasks such as malignant versus benign lesion classification and response to therapy.[91–95] More recently, they have also been shown to be useful in risk assessment.[63] However, radiomics approaches do not typically incorporate consideration of repeatability over imaging conditions in feature selection and evaluation of the resulting classifier. To address this, novel robustness metrics are proposed to characterize the equivalence, variability and correlation of radiomic features across varying imaging conditions. In mammography, imaging protocols vary across and within institutions. Parameters such as peak tube current, tube voltage, exposure time, half-value layer, quantization, and spatial resolution may vary across images within Mammography Quality Standards Act (MSQA) criteria.[96,97] These parameters may result in differences in texture values. In instances of feature comparison, it is critical to either correct for the impact of imaging parameters or to limit comparisons to only features with values that are affected less by the imaging parameters than the biological characteristic of interest.

Given the proposed robustness metrics, a method of feature selection is proposed that incorporates robustness assessment and classification evaluation in the construction of radiomics classifiers. This proposed method was designed to incorporate hierarchical clustering to identify redundant features. Then, robustness metrics are applied to identify a set of robust, non-redundant features. Once this set of features is identified, stepwise feature selection is applied to identify a subset of these features that are descriptive in the clinical task at hand. Thus, by proposing a feature selection method incorporating robustness, this work encourages a move towards consideration of imaging parameters in radiomics studies.

This work additionally explores implementation of emerging technologies in medical physics. One such exploration presented is the use of convolutional neural networks (CNNs) for transfer learning. As CNNs have gained popularity in the field of medical physics, here we present an examination of the use of transfer learning on DBT images as this modality is gaining ubiquity.

Furthermore, this work brings innovation to radiomic breast cancer risk assessment by incorporating temporal sequences of screening mammograms. While most studies in radiomics studies in medical imaging involve evaluation of a single image or a set of images acquired on a single date, we propose use of long short-term memory networks for use on temporal sequences of screening mammograms collected over a series of years for breast cancer screening. This method is illustrated in predicting risk of future malignant lesions.

1.5. Outline of Presented Works

Chapter 2 presents novel metrics for feature robustness characterization and describes their use in comparing radiomic features calculated on images from two different mammography vendors.

Chapter 3 presents a method for feature selection by applying the metrics proposed in Chapter 2 to select a set of features that are robust, non-redundant and relevant to the clinical task at hand. This method is illustrated in the clinical task of predicting the presence of high-risk factors. Chapter 4 implements transfer learning from convolutional neural networks to the task of classifying lesions detected in tomosynthesis images as either malignant or benign. Chapter 5 uses temporal sequences of mammograms acquired prior to detection of a mammographic abnormality to classify subsequently identified lesions as malignant or benign using long short-term memory networks.

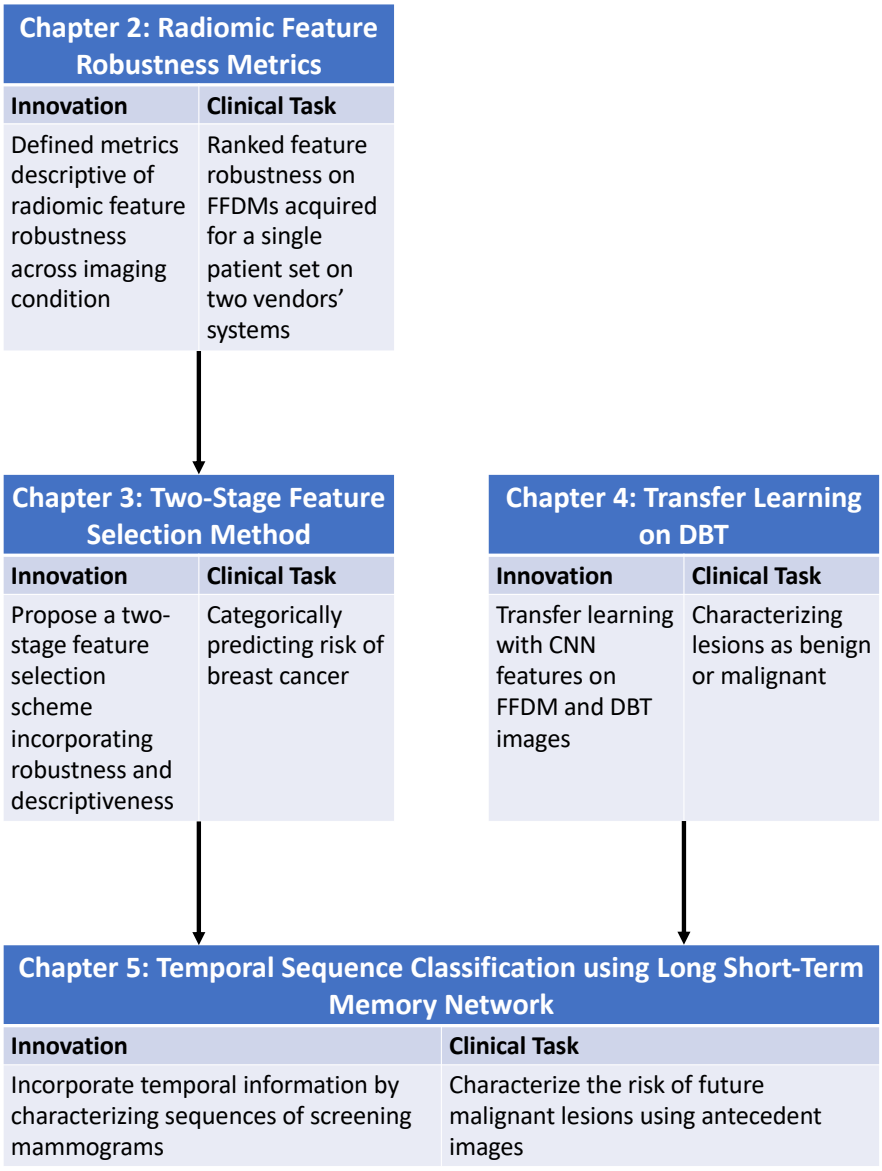


Figure 1.4. Summary of organizational structure of the presented works. Additionally, Chapter 1 is an introduction, and Chapter 6 is a conclusion of the works described here.

CHAPTER 2

Robustness of Radiomics Across Two Digital Mammography Manufacturers' Systems

2.1 Introduction

CAD has become a routine part of clinical workflow in mammography screening for early detection of breast cancer. Many clinical CAD systems employ radiomic feature extraction for detection and diagnosis of breast tumors.[98,99] However, the methods used in CAD have also been shown to aid in other clinical tasks, such as risk assessment. Radiomic features, such as breast density, have been shown to hold a role in predicting risk in asymptomatic women, prior to cancer induction.[100,101] Research studying risk predictors has also included radiomic texture analysis, which incorporates textural descriptors of the breast parenchyma.[102–104]

Within quantitative imaging, the goal of radiomics is to extract biologically meaningful information, as opposed to modifications caused by the image acquisition process. These types of modifications may be introduced by sources including image acquisition parameters, differing manufacturers or models, and differing hospital settings. Because these characteristics do not describe patient biology, efforts to standardize and harmonize imaging data can help remove these outside influences. Small datasets with heterogeneities in acquisition are often combined in the curation of large datasets, which are a key component of "big-data" methods such as machine learning and deep learning. Heterogeneities in data caused by non-biological factors are detrimental to utility of studies in this area. A first step to achieving data harmonization across imaging datasets is to develop an understanding of how some of the imaging variables affect feature calculations. Initial work to explore this was performed by Li et al. in comparing

classification performance on digitized screen-film mammograms to FFDM images in the task of risk assessment.

This study investigates the variation and robustness of radiomic features across two equipment manufacturers. Understanding feature variation and robustness may enable studies that lead to improved cancer risk assessment, detection, and diagnosis in the clinical arena. In this chapter, we propose novel metrics for robustness assessment and illustrate their use in characterizing feature robustness across two vendors' systems. Thus, this work contributes a method of evaluating feature robustness that may aid future radiomic studies by facilitating robustness considerations in feature selection.

2.2. Materials and Methods

2.2.1. Image Acquisition and Database Description

This retrospective study examined FFDM images acquired on a screening population of 111 women with low or average risk of breast cancer. Each case had screening mammography on a General Electric (GE) system and a Hologic system, separated in time by approximately one year (mean = 1.29 years, range: [0.86, 3.07] years). Of the 111 subjects, 104 were imaged first on a GE system, and 7 were imaged first on a Hologic system. No breast procedures were performed on subjects between the two studies. At the date of each subject's GE screening, subject ages ranged from 36 to 88 years (mean = 54.0 years, median = 52 years, standard deviation = 11 years). A description of the study population is included in Table 2.1. The FFDM images in this study were reviewed by an expert radiologist, and each case was included in this study only if no detectable abnormalities were observed in the images from both the GE and Hologic sets of mammograms. All images were acquired at the University of Chicago Medical Center.

Table 2.1. Demographics of the study population. The Hologic and GE imaging dates were separated by approximately one year and BI-RADS density is not always consistent between imaging exam dates, so the ages and breast density score are reported at the time of the GE exam. Data in parentheses are percentages.

Variable	Study Population Value
Mean age (y)	54.0
Age (y)	
<40	4 (2.70)
40-49	40 (36.04)
50-59	38 (34.23)
60-69	18 (16.22)
70-79	8 (7.21)
>80	3 (2.70)
Breast density score	
A	3 (2.70)
B	34 (30.63)
C	62 (55.86)
D	12 (10.81)

The FFDM images used in this study were collected retrospectively under an IRB-approved, HIPAA-compliant protocol. One set of images was acquired on a GE FFDM (Senographe 2000D) at 12-bit quantization with a pixel size of 100 × 100 μm. Another set of images was acquired on a Hologic FFDM (Lorad Selenia) at 12-bit quantization with a pixel size of 70 × 70 μm. Relevant system characteristics are summarized in Table 2.2.

Table 2.2. Summary of several key differences and similarities between the two systems examined in this paper.

Property	**GE Senographe**	**Hologic Selenia**
Pixel Size	100 μm × 100 μm	70 μm × 70 μm
Quantization	12-bit	12-bit
Anode Material	Rhodium	Tungsten
Detector Size	24 × 30.7	24 × 29
Detector Material	Amorphous Silicon	Amorphous Selenium
Conversion Method	Indirect	Direct

Regions of interest (ROIs) of size 256 × 256 pixels were selected manually from the central breast region directly posterior to the nipple in the craniocaudal projection of the mammographic images. Previous studies have shown that the radiomic features extracted from this ROI placement perform superiorly on risk assessment tasks.[61] ROI placement is illustrated in Figure 2.1, and the overall workflow is illustrated diagrammatically in Figure 2.2.

ROIs and radiomic texture features were extracted in a dedicated workstation. Subsequent analyses were performed in MATLAB (Version R2015b).

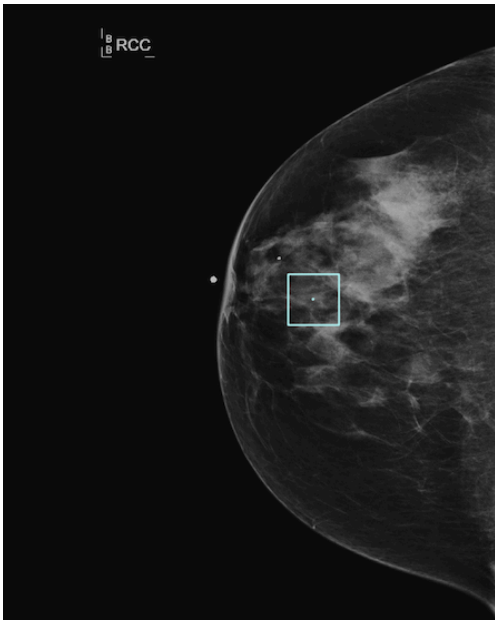


Figure 2.1. Region of interest placement, where the blue outline shows the region that is considered in feature extraction and analysis.

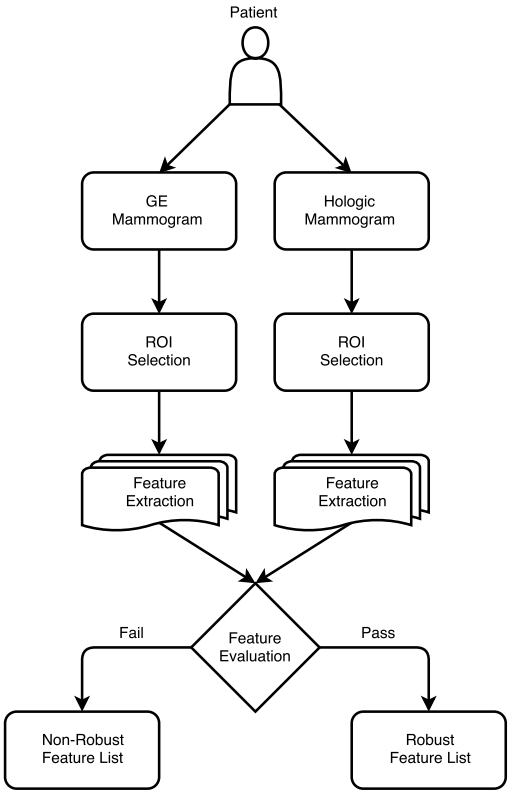


Figure 2.2. Flow chart illustrating the image analysis and feature extraction process employed in this study of feature robustness.

2.2.2. Radiomic Feature Extraction

Features describing mammographic parenchymal texture patterns were extracted automatically from each ROI. These computer-extracted features were based on algorithmic implementations of mathematical models for texture characteristics and have already been reported extensively in the literature. Texture characteristics were based on intensity, spatial pattern, and directionality within each ROI. Specifically, we computationally implemented (a) gray-level histogram analysis, (b) fractal dimensionality analysis, including box counting method and Minkowski method, (c) Fourier and power spectral analysis, (d) edge gradient analysis, and (e) GLCM features, as detailed elsewhere and in Chapter 1.4.[60,65,66,104] We categorized each feature computed

by these methods based on the texture quality that it described. These categories include intensity, spatial pattern, and directionality measures.

2.2.3. Statistical Evaluation

A goal of this quantitative imaging study was to develop metrics with which to evaluate the robustness of widely applied radiomic texture features as computed from mammograms across varying manufacturers. To that end, we selected four parameters for the assessment of the robustness of these features. These metrics were used to separate radiomic features into categories of robust and non-robust in terms of consistency across imaging manufacturer. The robustness metrics presented in this study included (a) the mean of feature ratios (MFR) to estimate equivalence, (b) the standard deviation of feature ratios (SDFR) to estimate variability, (c) the Spearman correlation coefficient (ρ_F) to describe correlation, and (d) the statistical significance of the Spearman correlation coefficient.

2.2.3.1. Equivalence: Mean of Feature Ratios (MFR)

The MFR was computed for each specific texture feature by comparing the ratio of the Hologic feature value to the GE feature value for each case, and then computing the mean value of these ratios across all pairs of images in the dataset. This calculation is given in Equation 2.1. Highly robust features are expected to produce similar values, regardless of imaging system employed, so a mean of ratios near unity indicates robustness in this regard. In this equation, $f_{i,v}$ indicates the value of a feature f for patient i on vendor system v. In this study, Hologic and GE were used for v1 and v2.

$$MFR = \frac{1}{N}\sum_{i=1}^{N}\frac{f_{i,v1}}{f_{i,v2}} \tag{2.1}$$

2.2.3.2 Variability: Standard Deviation of Feature Ratios (SDFR)

The variability between feature ratios was also examined to understand how consistent the feature ratio was across cases. To characterize this variation, the SDFR was computed as described in Equation 2.2. The ideal feature would have a constant ratio across manufacturers of one for every case, so a low variability near zero suggests robustness. In this equation, $f_{i,v}$ indicates the value of a feature f for patient i on vendor system v.

$$SDFR = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{f_{i,v1}}{f_{i,v2}} - MFR)^2} \tag{2.2}$$

2.2.3.3 Correlation: Feature Correlation (ρ_f)

The feature correlation across imaging systems was measured using the Spearman correlation coefficient (ρ_F). The statistical significance of each correlation was measured by the p-value of ρ_F.

Comparisons between these different robustness metrics were drawn, and ranges of acceptable criteria values as indicators of robustness were selected based on the distribution of features in this dataset. These proposed robustness metrics are presented in Table 2.3.

Table 2.3. Summary of the figures of merit used to characterize robustness, possible values each metric may hold, ideal values indicating "perfect" robustness, and cutoff ranges proposed in this study to indicate robustness.

Name	Description	Possible Values	Ideal Value
MFR	Mean of feature ratios	$MFR \in \mathbb{R}$	1
SDFR	Standard deviation of feature ratios	$SDFR \geq 0$	0
$\boldsymbol{\rho_F}$	Spearman correlation coefficient	$-1 \leq \rho_F \leq 1$	1
$\boldsymbol{p}$	P-value of spearman correlation coefficient	$p \geq 0$	0

2.2.4. Sensitivity to ROI Placement

This study recognizes the inherent limitation that placement variability may exist across images due to the deformable nature of breast tissue across repeat imaging. To understand the extent to which this limitation may have impacted our results, we investigated the sensitivity of radiomic features to ROI placement in terms of feature robustness. This was done by examining non-overlapping sub-samples of each ROI. We selected ROIs of size 128 × 128 pixels from non-overlapping positions within each original 256 × 256 pixel ROI. Radiomic features were computed in each of the sub-ROI positions and correlations between sub-ROIs from each individual case were examined, within each manufacturer image. A strong feature correlation indicated robustness to image region placement within a single image, while weak feature correlation indicated poor robustness to image registration and therefore high variability dependent on ROI placement. Features that remain relatively unchanged in response to shifts in region placement may be less impacted by imperfections in region alignment across the two images studied.

2.3. Results

The MFR, SDFR, ρ_F, and p-value of ρ_F of each feature were computed. The resulting values are summarized in Table 2.4. Through inspection, empirically based ranges in performance criteria for features were selected to differentiate robust from non-robust features. These criteria required (a) $0.8 < MFR < 1.2$, (b) $SDFR < 0.3$, (c) $\rho_F > 0.5$, and (d) $p < 0.05$ (statistically significant correlation). These cut-off values were data-driven and are highly specific to this feature set and data set. Subsequent studies may use different criteria to identify the most robust features.

Table 2.4. Statistics describing the robustness metrics used in this study describing all features calculated in this study. The large range and standard deviation of each robustness metric suggests a wide range in overall robustness of the features examined in this study. The large range and standard deviation of each robustness metric suggests a wide range in overall robustness of the features examined in this study.

Robustness Metric	Minimum	Maximum	Average	Standard Deviation	Cutoff Decision
MFR	-5.4	68.5	18.6	23.7	$0.8 < MFR < 1.2$
SDFR	0.0	81.1	16.9	22.3	$SDFR < 0.3$
$\boldsymbol{\rho_F}$	-0.1	0.7	0.3	0.2	$\rho_F > 0.5$

Features that met the criteria for robustness as proposed in this study included six box-counting fractal dimension features, 20 Minkowski fractal-dimension features, four power law features, and four GLCM features. The individual features that met the defined criteria for robustness tended to describe spatial patterns, as opposed to directionality or intensity in each image. A list of these features and corresponding values of the robustness criteria is included in Table 2.5. The same radiomic features were found to meet robustness criteria when a subset of

the study population of matched BI-RADS density was examined. These results are presented in Table 2.6.

Table 2.5. Summary of features found to meet the robustness criteria proposed in this chapter. The mean of each feature on the GE and Hologic ROIs are reported separately. The standard error (SE) reported on the MFR was calculated using replacement bootstrapping (100 samples).

Feature Name	μ_{GE}	$\mu_{Hologic}$	MFR ± SE	SDFR	ρ_F	*p*-value
Box counting Dimension (all 8 points)	2.8	2.9	1.040 ± 0.002	0.04	0.5	< 0.001
Box counting Dimension (first 4 points)	2.8	2.9	1.016 ± 0.002	0.03	0.7	< 0.001
Box counting Dimension (first 6 points)	2.7	2.8	1.025 ± 0.002	0.03	0.6	< 0.001
Box counting Dimension (first 3 points)	3.0	3.0	1.008 ± 0.002	0.03	0.7	< 0.001
Box counting Dimension (last 4 points)	2.6	2.7	1.037 ± 0.004	0.05	0.6	< 0.001
Box counting Dimension Regression y-intercept	-0.8	-0.9	1.157 ± 0.012	0.17	0.5	< 0.001
Global Minkowski Fractal Dimension	2.5	2.5	1.008 ± 0.001	0.02	0.7	< 0.001
Minkowski Minor Axis Diameter	9.5	11.2	1.179 ± 0.008	0.12	0.6	< 0.001
Powerlaw Beta (fine binning log of average)	2.0	1.9	0.972 ± 0.006	0.10	0.5	< 0.001
Powerlaw Beta (fine binning average of log)	1.9	1.9	0.973 ± 0.007	0.09	0.5	< 0.001
Powerlaw Beta (coarse binning log of average)	2.4	2.3	0.949 ± 0.005	0.08	0.6	< 0.001
Powerlaw Beta (coarse binning average of log)	2.4	2.3	0.952 ± 0.005	0.08	0.6	< 0.001
GLCM Correlation	1.0	0.9	0.979 ± 0.003	0.04	0.6	< 0.001
GLCM IMC 1	-0.4	-0.3	0.863 ± 0.014	0.21	0.6	< 0.001
GLCM IMC 2	1.0	0.9	0.980 ± 0.003	0.04	0.6	< 0.001
GLCM Maximum Correlation Coefficient	1.0	0.9	0.978 ± 0.002	0.04	0.6	< 0.001

Table 2.6. Summary of robustness metrics calculated on subsets of the full data population separated by radiologist-assigned BI-RADS density category. Due to limited quantities of data in these subsets, ρ_F and the corresponding p-value are not reported. The number of cases in each density category are reported, and both breasts for each patient were included in the calculation of robustness metrics.

	BI-RADS Density A n = 3		**BI-RADS Density B n = 34**		**BI-RADS Density C n = 62**		**BI-RADS Density D n = 12**	
Feature Name	MFR	SDFR	MFR	SDFR	MFR	SDFR	MFR	SDFR
Box counting Dimension (all 8)	1.01	0.05	1.03	0.04	1.04	0.05	1.03	0.03
Box counting Dimension (first 4)	1.01	0.02	1.02	0.03	1.01	0.04	1.03	0.04
Box counting Dimension (first 6)	1.01	0.02	1.03	0.03	1.02	0.04	1.03	0.03
Box counting Dimension (first 3)	1.00	0.02	1.01	0.03	1.00	0.04	1.03	0.05
Box counting Dimension (last 4)	0.99	0.01	1.04	0.05	1.04	0.05	1.02	0.04
Box counting Dimension Regression y-intercept	1.02	0.16	1.14	0.17	1.15	0.19	1.11	0.09
Global Minkowski Fractal Dimension*	1.00	0.01	1.01	0.01	1.01	0.02	1.01	0.02
Minkowski Minor Axis Diameter	1.17	0.16	*1.20*	*0.14*	1.16	0.13	1.18	0.09
Powerlaw Beta (fine binning log of average)	1.03	0.04	0.96	0.09	0.99	0.12	0.93	0.10
Powerlaw Beta (fine binning average of log)	1.04	0.04	0.97	0.10	0.99	0.11	0.93	0.10
Powerlaw Beta (coarse binning log of average)	1.00	0.04	0.93	0.08	0.97	0.10	0.93	0.08
Powerlaw Beta (coarse binning average of log)	1.00	0.03	0.94	0.08	0.96	0.09	0.93	0.08
GLCM Correlation	0.99	0.03	0.98	0.04	0.98	0.05	0.98	0.03
GLCM IMC 1	1.01	0.16	0.86	0.19	0.89	0.27	0.85	0.18
GLCM IMC 2	0.99	0.03	0.98	0.04	0.99	0.05	0.98	0.03
GLCM Maximum Correlation Coefficient	0.98	0.03	0.98	0.04	0.98	0.04	0.98	0.03

As illustrated in Figure 2.3, features included in this study displayed a wide range of ρ_F and MFR. By inspection of this graph, we determined that spatial pattern features produce values comparatively highly correlated, with similar values across manufactures compared with features computed to describe texture directionality or image intensity. Differences between the mean of ratios and ratio of means, as shown in Figure 2.4, demonstrate the existence of cases with ratios that far from the mean for several given features. Standard error of the MFR was calculated

using bootstrapping with replacement.[105] A mean of ratios near the ratio of means indicates robustness. Features with a ratio of means far from the mean of ratios may thus be considered non-robust.

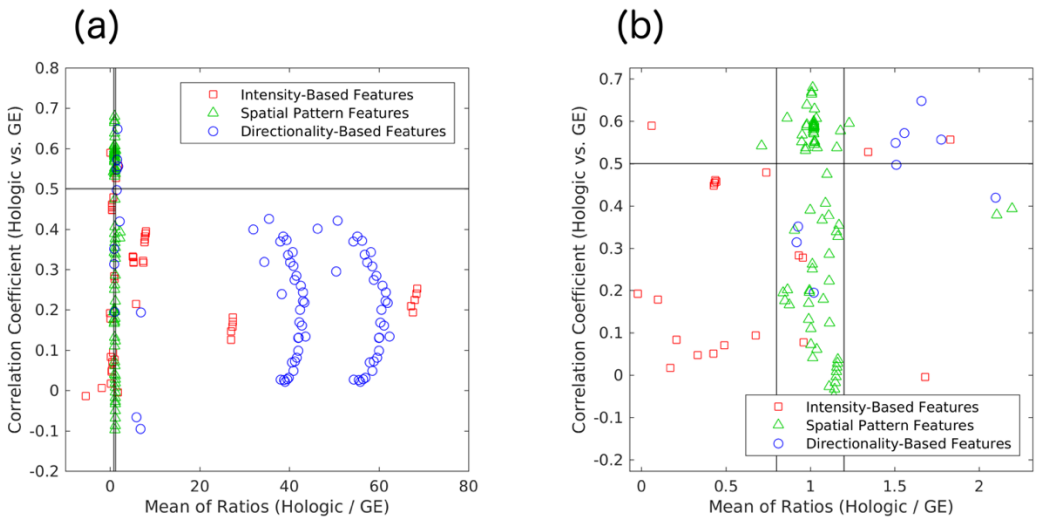


Figure 2.3. Scatter plot showing the correlation and mean of ratios of various features examined in this study. Graph (a) is scaled to show all features included in this study, and (b) is the same plot reproduced with a narrower scale to show in more detail the robustness of features included. The color of each point shows the categorical basis of that feature. Solid black lines show the ranges of the robustness selection criteria employed in this study. Most features with high correlation and mean of ratios near a value of one tend to be spatial pattern features, as opposed to intensity-based or directionality-based features.

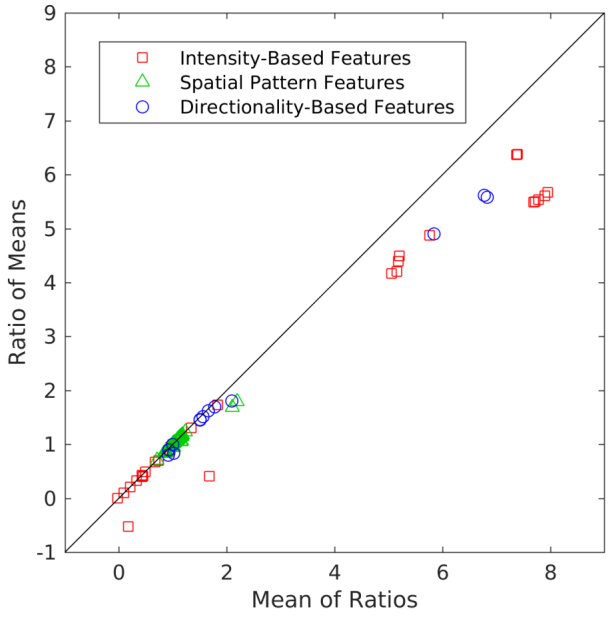


Figure 2.4. Scatter plot showing the mean of ratios and ratio of means of each feature investigated in this study. Features with a low number of outliers have similar values for the ratio of means and mean of ratios, while features with data deviating from a uniform pattern tended to have greater values for the mean of ratios compared to the ratio of means.

In an investigation of the impact of ROI placement on radiomic feature values, we found that correlation across non-overlapping ROIs in a single image were, on average, higher than the correlations across manufacturers. The first, second, and third quartiles of feature correlation computed on non-overlapping regions within a single image are shown in Figure 2.5.

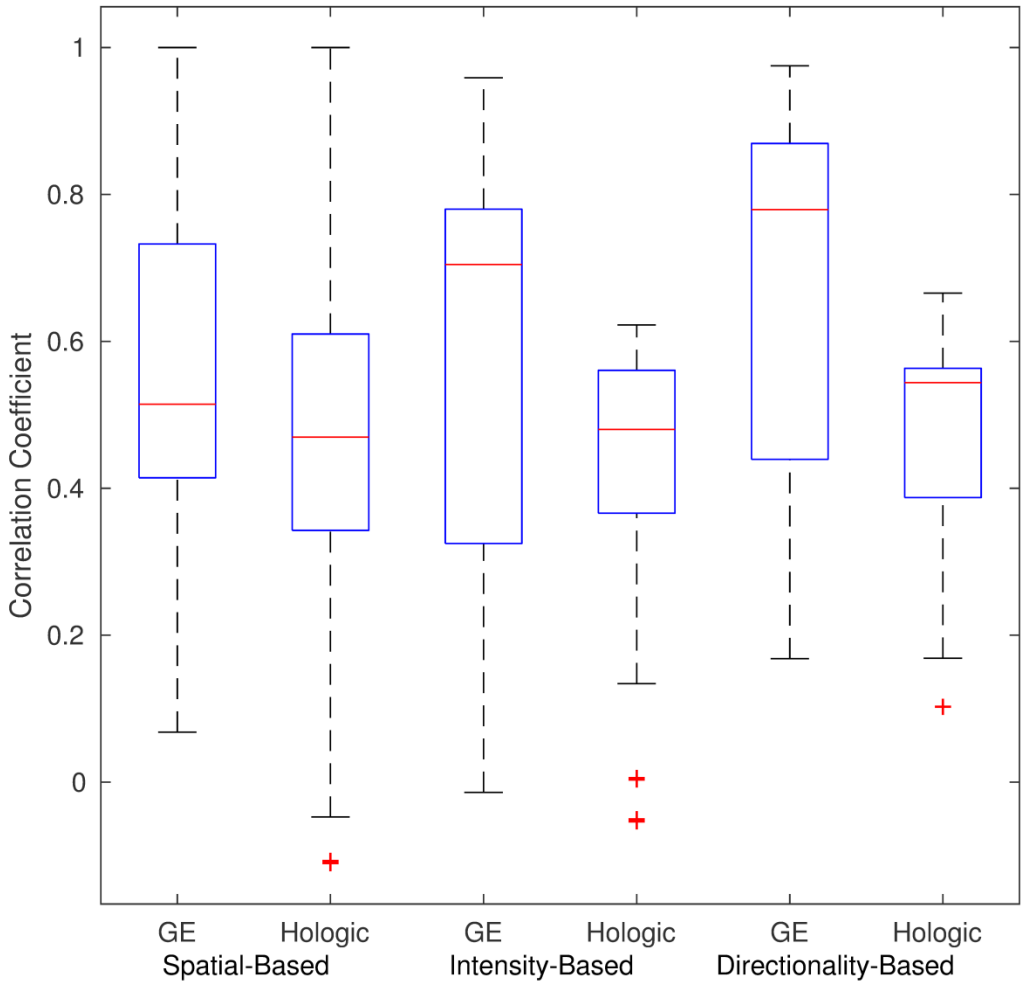


Figure 2.5. Boxplot showing the effect of region placement on radiomic features extracted from GE and Hologic FFDM images. In this figure, the blue boxes indicate the first and third quartiles of feature correlations when 64 × 64 pixel sub-ROIs of non-overlapping location in a single image are compared. The red horizontal lines in each box represent the median of feature correlations.

To explore whether the observed correlations were decremented by outliers or whether the observed correlations describe the overall relationship, we trimmed the data points included in correlation calculations for each feature by removing any point that lay in the first or fourth quartile in each the Hologic and GE system datasets in terms of the feature value quartiles.

Specifically, note that data points were removed if their value fell outside of the interquartile range compared to other patient feature values, but not based on the agreement across features calculated on the two vendors. Correlations were computed on the retained data points, and values of the ρ_F robustness metric were improved in 26 out of 217 features. Each of the 26 features that produced an increased correlation coefficient were defined as non-robust by our previously described metrics, and the correlation coefficient did not increase by enough to move any of these features into the robust category. An examination of a subset of robust and non-robust features with illustrative cases of good feature agreement and poor feature agreement are provided in Figure 2.6.

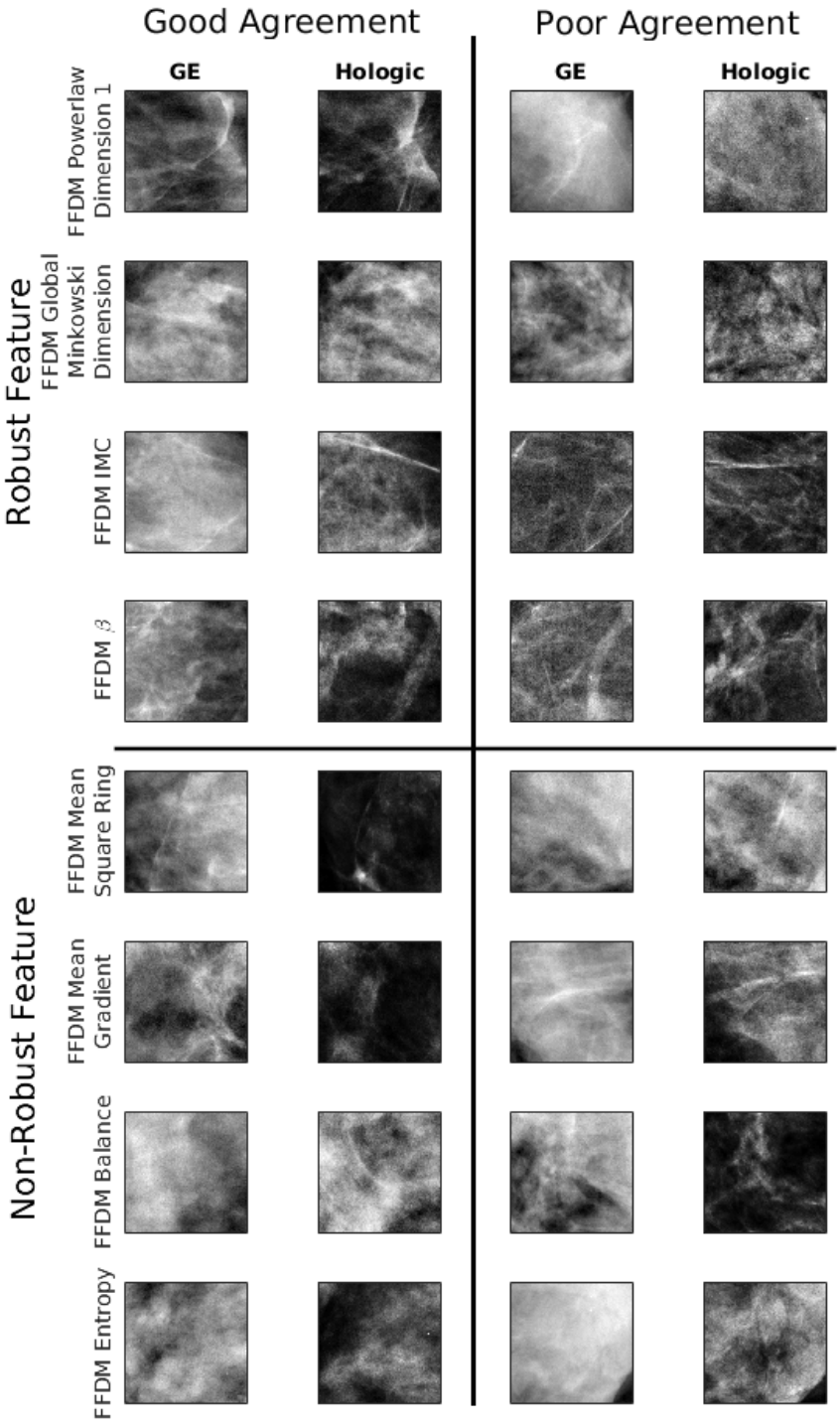


Figure 2.6. Illustration of example ROIs for the case of good and poor agreement across robust and non-robust features. Robust features are those that met the robustness criteria proposed in this study on a population level. Here, the examples shown are for the individual patient with the best and worst agreement of each feature within the dataset.

2.4. Discussion

In this study, we investigated feature robustness across different manufacturers of mammography equipment. We quantified heterogeneities in feature values caused by the variation of this parameter in the image acquisition process. This examination is intended to represent a first step in motivating a broader look at the quantitative effects that various steps of the image acquisition process have on extracted features and the role that this may play in CAD, evaluation of response to therapy, and other quantitative image analyses.

Our principal interest was in investigating measures of feature robustness across images acquired under different imaging parameters and with different imaging equipment. As such, we chose to use descriptors of equivalence, variability, and correlation in judging the agreement in feature values across acquisition on equipment from different manufacturers. We thus expect that robust features would demonstrate high equivalence across manufacturers. However, we accept that due to many factors, some of which are listed in Table 2.2, perfect equivalence is highly unlikely. Therefore, we expanded our criteria to assess the correlation between feature values as produced by the two different manufacturers.

Several factors affect the values of features examined in this study. As explored in this study, the placement of the ROI plays a role in the computation of feature values. This was illustrated by the range of correlation coefficients for features calculated in adjacent, but not overlapping, ROIs in a single image as illustrated in Figure 2.5. The effect of image acquisition parameters was removed by comparing correlation coefficients in a single image. Therefore, any deviation in feature value within a single image at different points demonstrates sensitivity of that feature to ROI placement. For features that are highly sensitive to ROI placement, any differences in feature placement across images from different manufacturers would produce changes leading to

induction of heterogeneity in the dataset. Li et al. has shown that changing the breast region in which an ROI is placed can significantly decrease the utility of texture features in the task of assessing risk.[61]

Certain radiomic features were shown to be sensitive to characteristics of the mammography vendor, such as equipment and post-processing algorithms. While pixel size is accounted for in the feature computation algorithm, it may still affect the values of different features. This is because larger pixels do not show finer detail, and some features that compute feature on a finer scale may lose this information.

A limitation in this study was the use of "for presentation" as opposed to "for processing" images. The images used in this study were processed by manufacturer-defined algorithms prior to feature extraction, and the algorithms used were not held constant over cases due to periodic updates of clinical systems throughout the period covered by our retrospective image collection. While the systems used in this study are no longer state-of-the-art, they are sufficient for us to present new metrics for use in assessing robustness.

It is interesting to note that many of the features found to be robust in this study are among the more frequently used features in literature. Specifically, the various calculations of box counting dimension, Minkowski dimension, power law beta, GLCM correlation, GLMC IMC1, GLCM IMC2, and GLCM maximum correlation coefficient are consistently considered in analyses by our group.[60,63,65,83,106]

It is important to also note that scans on GE and on Hologic machines were separated in time by about one year, as each was part of normal screening for the respective patient. It has been well documented that parenchymal patterns, including density, change over a woman's lifespan.[107] As we did not obtain scans on different mammography equipment on the same date,

age was not controlled for in our study. However, it would be interesting to explore the quantitative effect of age on the feature values explored in this study.

2.5. Conclusion

The field of radiomics depends on large datasets to draw robust conclusions about the detection, diagnosis, and therapy response of various diseases. One method by which large datasets might be produced is by combining smaller datasets acquired under different parameters, such as with different imaging equipment from different manufacturers. In doing so, there is the innate assumption that the differences between radiomic features are due to the patient biology and not differences in image acquisition. Therefore, this study aimed to draw conclusions as to which texture-based radiomic features are robust to the merging of datasets acquired on different mammography equipment. In this study, we proposed a set of robustness metrics including (a) the MFR to estimate equivalence, (b) the SDFR to estimate variability, (c) the Spearman correlation coefficient (ρ_F) to describe correlation, and (d) the statistical significance of the Spearman correlation coefficient to describe the robustness of radiomic features. These metrics were applied to FFDM texture features derived from cases imaged on two digital mammography systems from different manufacturers and used to differentiate robust from non-robust features in this regard.

By providing metrics to characterize the robustness of radiomic features across heterogeneous image sets, we hope to lay the groundwork for future efforts to develop methods of standardization and harmonization of data. This stands to produce more homogeneous datasets and further improve big data studies and implementation of machine learning in quantitative imaging.

CHAPTER 3

Robustness Assessment and Classification Evaluation (RACE) Demonstrated on Multi-Manufacturer FFDM

3.1. Introduction

Breast cancer is one of the most commonly screened for forms of cancer, with 65.3% of women age 40 and over reporting having had a mammogram in the past two years.[108] In addition to detecting cancer, mammograms provide imaging phenotypes that may inform lifetime risk. For example, it has been well documented that mammographic density can be useful in predicting breast cancer risk.[56,59,64] Typically, personalized risk models include characteristics such as age, family history and certain genetic mutations such as BRCA1/BRCA2. Developments in CAD suggest that parenchymal texture may also help inform risk along with these other demographic factors. Thus, stronger risk prediction models may be facilitated by considering imaging phenotypes and their relationships to cancer risk.

Quantitative measures of parenchymal texture have been successfully applied to evaluate the risk of cancer in asymptomatic females.[55,57,60,63,65–69] These studies used radiomic texture features including fractal dimension,[65] power-law spectral analysis,[60] absolute gray-level, gray-level histogram analysis, neighborhood gray tone difference matrix (NGTDM), and GLCM.[70]

Risk evaluation stands to be particularly impactful for patient care due to established high-risk screening recommendations that have been enacted by agencies such as the American Cancer Society. These recommendations help translate identification of high-risk individuals to actionable recommendations, which may lead to improved early detection of disease.[71] The availability of specialized screening modalities such as MRI and clinical impact of supplemental

screening on high-risk populations has elevated the demand for strong risk evaluation metrics. The actionable screening steps available to women at an elevated risk of breast cancer have motivated continued research in risk assessment in order to best utilize the available specialized screening modalities.

One challenge faced in developing widely generalizable imaging phenotypes is the sensitivity of individual texture features to imaging conditions. Imaging conditions such as manufacturer, kVp, and processing algorithms may each affect radiomic feature values.[109–113] Studies have been performed to evaluate repeatability (test-retest) and reproducibility of radiomic features in cancer imaging.

In a study by van Velden et al., the repeatability of radiomic features in non-small-cell lung cancer (NSCLC) using positron emission tomography (PET) computed tomography (CT) images was investigated.[114] The study reported high repeatability of radiomic features relative to standardized uptake value measures and found that features were more sensitive to region of interest placement than to changes in reconstruction.

Drukker et al. performed case-based analysis to evaluate repeatability on sonographic lesions in the task of classifying lesions as malignant or benign.[115] The study sought to investigate case-based classifier output over bootstrapped repetitions.

Hunter et al. studied reproducibility and redundancy of radiomic features of NSCLC patients and on a texture phantom from CT.[116] The study reported that feature redundancy and reproducibility was highly machine-sensitive.

Zhao et al. used same-day repeat CT scans of lung cancer patients to evaluate the impact of reconstruction settings, slice thickness, and reconstruction algorithms on feature

repeatability.[117] The study concluded that most texture features are repeatable, although they were significantly impacted by reconstruction parameters.

However, incorporation of repeatability and reproducibility into feature selection and classification construction procedures is relatively unexplored. Therefore, this study proposes methods by which to implement the findings of robustness studies to the improvement of CAD methodology. Here, we define repeatability as the variation between measurements using the same equipment, and reproducibility as the variation between measurements using different equipment.

Many feature signatures for risk are developed on homogeneous databases, and reproducibility over imaging conditions is not always evaluated in imaging phenotype studies. Efforts towards this end have been taken by the Quantitative Imaging Biomarkers Alliance (QIBA) by forming working groups to focus on metrology concepts, algorithm comparisons and technical performance of imaging biomarkers. These groups have published metrology papers describing best practices for studies involving analysis of imaging biomarkers.[118–120] The work presented here builds off of these works. To ensure generalizability of findings to heterogeneous imaging conditions, this study seeks to identify a parenchymal texture signature descriptive of risk of breast cancer by emphasizing both (a) robustness across imaging manufacturer and (b) classification accuracy in feature selection methodology. This is important because in clinical practice, images are acquired on a number of different models from many manufacturers, used with a range of settings. Our study seeks to present a method of identifying features that are repeatable over FFDM manufacturers and incorporate the subset of these that are descriptive and non-redundant into the construction of a classification model. We demonstrate this method using the clinical task of classifying collectively the presence of breast cancer risk factors. For brevity,

we refer to this two-stage method as RACE (robustness assessment, classification evaluation) as shown in Figure 3.1.

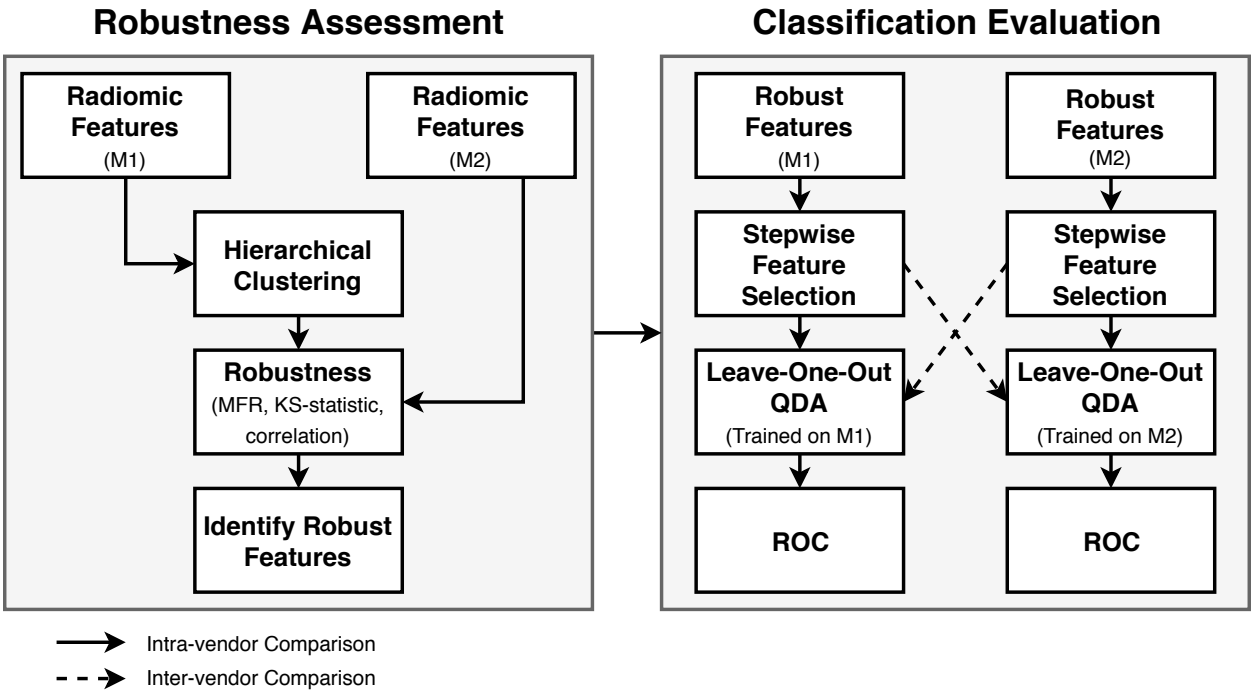


Figure 3.1. Diagrammatic illustration of steps involved in the RACE method. Texture features are first clustered and assessed in terms of robustness using only feature values and vendor information, remaining blinded to risk classification. The union of features identified by clustering features from M1 (machine one) and M2 (machine two) is the set considered to be robust and non-redundant. The most robust and non-redundant features are identified, and only these features are used as feature candidates in classification evaluation. Solid and dashed arrows show two different data pathways followed to evaluate the generalization of classification of the heterogeneous image datasets. The full analysis was repeated twice; once with the GE unit as M1 and the Hologic unit as M2, and then again with the GE unit as M2 and the Hologic unit as M1. The task in this study was to classify patients as having either a high or low risk of breast cancer.

3.2. Materials and Methods

3.2.1. Image Acquisition and Database Description

All images included in this study were retrospectively collected FFDM acquired under standard

clinical protocols. All images were acquired at the University of Chicago Medical Center. All images used in this study were collected under an institutional review board (IRB) approved, Health Insurance Portability and Accountability Act (HIPAA) compliant protocol. All subjects had no detectable mammographic abnormalities and were classified as either having or not having a risk factor of breast cancer. This classification was based on each subject's family history of breast cancer, family history of ovarian cancer, personal history of atypical ductal hyperplasia (ADH), and personal BRCA1/BRCA2 status. Additionally, patients with a personal history of breast cancer or any suspicious abnormality were excluded from this study. Thus, while risk varied between the two groups, no lesions were present in the images examined in this study as judged by the clinical radiologist report accompanying each image and verified by a breast image researcher. Each subject underwent screening mammography on a GE system and a Hologic system. The GE images were acquired on a GE Senographe 2000D at 12-bit quantization with a pixel size of 100 × 100 μm. The Hologic images were acquired on a Hologic Lorad Selenia at 12-bit quantization with a pixel size of 70× 70 μm. Sets of images were separated in time by about 1 year. The mean age of women without high risk factors present was 54.3 years (range = 39-86), and the mean age of women with high risk factors present was 49.7 years (range = 24-88). No breast procedures were performed on subjects between the two studies, and all images were assigned BIRADS 1 (negative) when reviewed by a clinical breast radiologist. Characteristics of the study population are summarized in Table 3.1.

Table 3.1. Demographics of the study population separated by risk of cancer. Data in parentheses are percentages. Radiologist-reported BIRADS density was not always consistent between the GE and Hologic imaging exam, so values in this table represent the density and age reported at the time of the GE exam. Also, summary of indication for high-risk designation is presented. Some subjects may be designated as high-risk for more than one factor. A breakdown of inclusion criteria is also shown. In this context, small breast is defined as breast area smaller than the size of a 512x512 pixel square as this limited our ability to compute features on images in this analysis.

Variable	**Number of Patients without Risk Factors Present**		**Number of Patients with Risk Factors Present**	
Mean age (SD)	54.3 (10.5)		49.7 (11.6)	
Age (year)				
<40	1		20	
40 to 49	31		29	
50 to 59	28		37	
60 to 69	15		10	
70 to 79	7		4	
≥80	1		2	
Breast density score				
A	2		4	
B	27		41	
C	44		54	
D	10		3	
Risk Factor				
Family history of breast cancer	--		107	
Family history of ovarian cancer	--		9	
BRCA1/BRCA2 mutations	--		3	
Personal history of ADH	--		1	
Breast Area Exclusion	**Patients**	**Images**	**Patients**	**Images**
Total in Database	86	172	112	224
# Small Breast	3	9	10	27
# Included in study	83	163	102	197

A small number of women were excluded from this study because the breast area in their images was smaller than that required for placement of the ROI. Small breast area could result from small breast volume, large pixel size, or the extent of breast compression during image acquisition. The quantity of such exclusions is reported in Table 3.1.

The distribution of time intervals between exams is shown in Figure 3.2. This histogram shows the interval of time between the GE and Hologic exam dates for each patient included in this study, separated by those with a high or low risk of breast cancer. The distribution of days between exams for the group with and without high risk factors were not shown to be significantly different by the two-sample t-test ($p = 0.29$).[121] Thus, this suggests that differences in time intervals between the two populations can be explained by random chance.

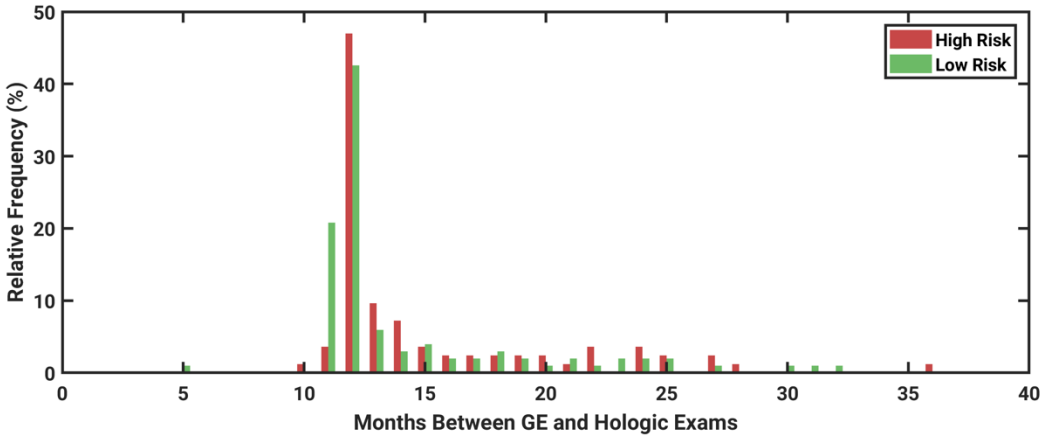


Figure 3.2. Histogram demonstrating the interval of time between the date of the GE exam and the Hologic exam, for each patient included in the study. The time between exams was not found to be significantly different between women with and without high risk factors present ($p = 0.29$). Note that the GE image was not always acquired before the Hologic image.

ROIs and radiomic texture features were extracted in a dedicated workstation. Subsequent analyses were performed in MATLAB (Version R2015b).

3.2.2. Radiomic Feature Extraction

Radiomic texture features were calculated on square ROIs of size 512 × 512 pixels which were manually placed in the central breast region posterior to the nipple. The CC images of the left and right breasts were used. Manual alignment over the two vendor images was done by visually examining both images simultaneously. Previous studies have shown that this ROI size and

placement scheme performed best across different locations in the breast.[61] Because all images had a negative or benign interpretation, ROIs were placed over normal background parenchymal tissue. Following manual ROI placement, features were automatically calculated on each ROI. The features were based on algorithmic implementations of mathematical texture descriptors that have been reported extensively in the literature.[60,65,70] Specifically, features were based on (a) gray-level histogram analysis, (b) fractal dimensionality analysis including the box-counting method and Minkowski method, (c) Fourier and power spectral analysis, (d) edge gradient analysis, and (e) GLCM. The quantity of features calculated from each group is summarized in Table 3.2. This set of quantitative features was evaluated because the constituent features have demonstrated utility in previous studies involving clinical classifications based on parenchyma regions in FFDM images.[60,61,65,66] These features are discussed in greater detail in Chapter 1.4.

Table 3.2. Quantity of feature types included in the feature set from which features were selected for classification analysis.

Feature Category	**Number of Features**
Fourier	148
Box counting fractal dimension	6
Edge-frequency	4
Histogram	38
Minkowski fractal dimension	32
Power-law beta	8
GLCM	14
First order	6
Total	256

3.2.3. Robustness Assessment

Hierarchical clustering was performed to identify groups of redundant features.[122] In hierarchical clustering, a hierarchy of clusters is built in an unsupervised manner such that objects in each cluster are similar with respect to a specified metric. In this study, the Pearson correlation

coefficient was used as the distance metric.[123] Clustering was performed in an agglomerative manner such that maximally similar clusters were iteratively merged in the construction of the feature hierarchy. Single linkage (nearest neighbor) was used to describe the distance between objects. The number of clusters used in grouping ranged from 18 to 256 in order to evaluate the impact of varying levels of strictness in robustness considerations, as the motivation of clustering is to select one robust feature from each cluster to consider in later stages of feature selection. The lower end of the range of the number of clusters was chosen based on the findings of Hua et al.[124] The upper end of the range of the number of clusters was chosen because a total of 256 radiomic features were considered in this study. Therefore, by sorting the features into 256 clusters, only a single feature persists in each cluster, which therefore would be analogous to omittance of robustness considerations as all features, regardless of their level of robustness, would persist.

A wide range of clusters was investigated in order to explore the trend in classification performance as restrictions on robustness varied, thereby permitting an evaluation of the relevancy of robustness considerations in classification performance. However, in practical application of the proposed method, it is expected that only one number of clusters need be considered. In general, the optimal number of clusters may depend on study-dependent factors such as the specific classification task, the redundancy between features in the full feature set explored, and the number of patients evaluated in the study.

Feature robustness across mammography vendors was evaluated using statistical metrics including (a) mean of feature ratio (MFR), (b) correlation coefficient, and (c) Kolmogorov-Smirnov test statistic as introduced in Chapter 2.[109,125] These robustness metrics were selected to describe equivalence, correlation and similarity of the sample distributions.

A composite indicator (CI) was developed to merge information from three robustness metrics investigated in Chapter 2 so as to include multiple aspects of robustness in evaluating features. The CI was calculated by taking the sum of metric values normalized by z-score. Metric z-scores were weighted by +1 when a high value indicates robustness, and by -1 when a low value indicates robustness. Therefore, the CI for feature f is defined by Equation 3.1, where $z_{corr,f}$ is the z-value of the correlation coefficient for feature f, $z_{MFR,f}$ is the z-value of the mean of feature ratios for feature f, and $z_{KS,f}$ is the z-value of the Kolmogorov-Smirnov test statistic for feature f. A higher CI value indicates relatively high robustness, and a low CI value indicates relatively low robustness compared with the other features considered.

$$CI_f = z_{corr,f} - z_{MFR,f} - z_{KS,f} \tag{3.1}$$

Note that high correlation values indicate high robustness, and low MFR and SDFR values indicate high robustness thus explaining the coefficient of ±1 in Equation 3.1. Relative robustness ranking of the investigated texture features was performed by ordering features based on their CI value in descending order, where more positive values of CI_f suggest strong robustness, and more negative values of CI_f suggest weak robustness. The feature with the highest CI_f in each cluster were identified and considered in the classification evaluation stage.

3.2.4. Feature Selection and Classification

The classification stage involved using robust, non-redundant features to predict a woman's risk of breast cancer based on the presence of risk factors. The workflow for the model is illustrated by Figure 3.1.

The robust, non-redundant features identified at stage one were fed into stepwise feature selection separately for each vendor. The stepwise feature selection method employed in this study applies a stepwise regression by iteratively adding and removing features from a multilinear model.[126] Features are added or removed from the model based on the statistical significance of the change in performance, with the p-value of the F-statistic used as the figure of merit. Feature selection was performed in a leave-one-out manner, and the top features were identified as those features selected the greatest number of times. The top 18 features were used in analysis, based on the findings of Hua et al.[124] Therefore, in each classification performed in this study, regardless of the number of clusters used in the robustness assessment step, exactly 18 features were ultimately selected. Note that as the number of clusters is altered, this will alter which 18 features are ultimately selected for use in the classifier, as the robustness constriction is tuned by the number of clusters.

It is standard in typical radiomics studies to perform feature selection, such as stepwise feature selection, on the full set of candidate features with no consideration for feature robustness. In our study, this approach is equivalent to having the number of clusters equal to the number of candidate features, thus causing all features to pass the first stage of robustness assessment. Specifically, for intra-vendor analyses in which the number of clusters is equal to the number of candidate features, no information from the second vendor system was used in feature selection. Therefore, in this study, intra-vendor classification with 256 clusters used shows the performance of an approach in which differences in FFDM systems is disregarded from analysis. Likewise, the inter-vendor classification with 256 clusters used shows the performance when no robustness criteria are used in limiting candidate features.

Following feature selection, selected features are used in leave-one-out quadratic discriminant analysis (QDA) to build a model for classification. QDA is similar to linear discriminant analysis but has a quadradic decision surface as opposed to a linear surface. Specifically, QDA does not hold the assumption that the co-variance matrix is common to all classes. Thus, it estimates co-variance matrices separately for each class. Models were built separately for GE and Hologic images. To evaluate the classification performance in the task of risk classification, the full classification evaluation analysis was performed in a leave-one-out manner (single fully nested loop). Receiver operating characteristic (ROC) analysis was used to calculate the area under the curve (AUC). The AUC was used as the figure of merit in this analysis in the task of classifying risk of breast cancer.

As illustrated by Figure 3.1, stepwise feature selection was performed on images from a single vendor. However, QDA was used to construct texture signatures merging the selected features on each of the two vendor image sets. Performance was evaluated in the task of classifying risk of breast cancer, and agreement in performance was used to characterize generalizability of the model across vendors. We will refer to the vendor from which the images were used to select features as machine one (M1) and the other vendor used to assess generalizability as machine two (M2). Robustness Assessment Classification Evaluation (RACE) was repeated with each GE and Hologic as M1. The entire feature selection process (clustering, robustness ranking, stepwise feature selection) was performed once based on clustering features from the GE unit and once using features from the Hologic unit. The full classification analysis (QDA, leave-one-out classification) was performed on the GE unit and the Hologic unit for each set of features identified using clustering from both the GE- and Hologic-identified features. When clusters were based on images from GE, then the GE unit was considered M1 in the

analysis scheme. Likewise, when clusters were based on images from Hologic, then the Hologic unit was considered M1 in the analysis scheme. Additionally, the classification evaluation stage was performed as a single loop for separated training and testing to avoid overestimation of the resulting classification performance.

3.2.5. Comparison to Harmonization Methods

While the approach to handling heterogeneous feature data in this chapter focuses on limiting feature selection to robust features through a two-stage analysis (RACE), other groups have approached the same issue by harmonizing (or standardizing) feature data across different imaging conditions. One such example is the ComBat harmonization method, originally developed to correct for the "batch effect" in the genomics field and later applied to PET radiomics studies.[127,128] In a study by Orlhac et al., the ComBat harmonization method was applied to standardize radiomic features extracted from PET images of breast cancer patients acquired at two different institutions in order to identify triple negative (TN) lesions.[127]

As first suggested by Johnson et al. and implemented for PET radiomics by Orlhac et al., the ComBat harmonization method functions by estimating the additive scanner effect, γ, and the multiplicative scanner effect, δ, using Empirical Bayes estimates.[127,128] Thus, the normalized value of features, y, are described by Equation 3.2, where y_{ij} is the standardized feature for ROI j and scanner i, α is the average value for feature y, γ is the additive effect of scanner i, δ is the multiplicative scanner effect, and ε is the error term.

$$y_{ij}^{ComBat} = \frac{y_{ij} - \hat{\alpha} - \gamma_i^*}{\delta_i^*} + \hat{\alpha} \tag{3.2}$$

In the evaluation of the ComBat method on our data, we normalized each of the 256 examined features according to Equation 3.2, and then performed stepwise feature selection and QDA for leave-one-out ROC analysis, mimicking the classification evaluation analysis of RACE (Figure 3.1 right). For evaluation of ComBat harmonization, the robustness assessment stage (Figure 3.1 left) was omitted, as feature harmonization is expected to improve feature robustness across imaging conditions. To match the analysis conditions from RACE, a total of 18 features were included in the final radiomic signature construction. Furthermore, to investigate the impact of ComBat harmonization on feature robustness, robustness metrics were computed after ComBat harmonization, and we compared robustness metric values before and after ComBat harmonization using a two-tailed t-test.

To explore the potential interplay between the ComBat and RACE methods, we also initially applied ComBat on features for harmonization, and then used these harmonized feature values in the RACE feature selection method. To match each of the individual methods, 46 clusters were used in the RACE method, and 18 features were ultimately selected for the final radiomic signature construction. These comparisons are summarized in Figure 3.3

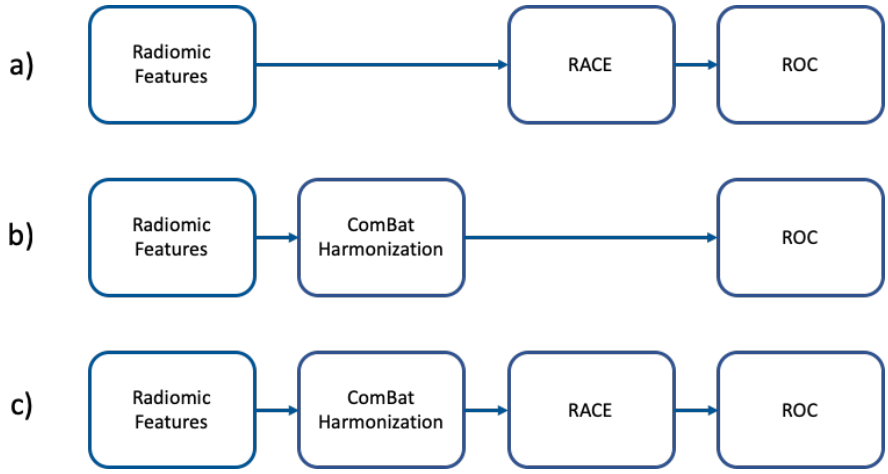


Figure 3.3. Diagram showing the different series of algorithms explored in comparing ComBat harmonization to the proposed RACE method. A) Radiomic features are extracted, and then the RACE algorithm is used. B) Radiomic features are extracted, and then only ComBat harmonization is used, along with feature selection. C) Radiomic features are first pre-processed using ComBat harmonization, and then the harmonized feature values are used in the RACE algorithm. The task for each workflow was to predict the risk of breast cancer among patients with no mammographically detectable lesion present at the time of screening.

3.2.6. Statistical Evaluation

In order to evaluate whether robustness considerations improved classification performance, data series over varying number of clusters were evaluated for the presence of a monotonic trend using the Mann-Kendall test for monotonic trend.[129] The Mann-Kendall test evaluates the presence of a statistically significant trend between the number of clusters used in analysis. The direction of the trend (increasing or decreasing) was then computed by the Thiel Sen estimator.[130] The sign of the Thiel-Sen estimator indicates whether the classification performance tended to increase or decrease as the number of clusters increases.

In comparing the RACE method to ComBat harmonization, the AUC values for both inter- and intra- vendor performance were calculated when features were selected on GE and on Hologic images in the task of predicting risk. The statistical significance of the difference between AUCs from ComBat harmonization and from the RACE was calculated using ROCKIT

software.[131] By applying the Holm-Bonferroni correction for multiple comparisons, $p \leq 0.017$ was required to demonstrate statistical significance.

3.3. Results

3.3.1. RACE Performance

Features found to be the most robust relative to the other features examined in this study are summarized in

Restricting candidate features to just the most robust features was also shown to have a significant impact on inter-vendor classification performance in the task of risk assessment. The classification performance was observed to monotonically decrease as the number of clusters increased (Mann-Kendall $p < 0.001$) as shown in Figure 3.4. An increase in the number of clusters corresponds to a reduction in the stringency of robustness criteria.

Table 3.3. Feature families that tended to have a large proportion of robust features included box counting fractal dimension, power-law beta, and GLCM features. Percentage density also was robust over vendors, relative to the other features examined here.

Restricting candidate features to just the most robust features was shown to have a significant impact on classification performance of the intra-vendor evaluation as demonstrated by a monotonic increase in AUC with increasing number of clusters in the task of classifying risk (Mann-Kendall $p = 0.0168$ and $p < 0.001$ for GE and Hologic, respectively). However, the Thiel-Sen estimator of the rate of increase was still very small (0.0000586 and 0.000120 for GE and Hologic, respectively).

Restricting candidate features to just the most robust features was also shown to have a significant impact on inter-vendor classification performance in the task of risk assessment. The classification performance was observed to monotonically decrease as the number of clusters

increased (Mann-Kendall $p < 0.001$) as shown in Figure 3.4. An increase in the number of clusters corresponds to a reduction in the stringency of robustness criteria.

Table 3.3. List of the most robust features over the two vendors examined in this study. The composite indicator is a measure of robustness, where larger values indicate a more robust feature relative to the others examined in this study. The composite indicator is computed according to Equation 3.1. Features that were observed to be robust in Chapter 2 are noted by *.

Feature Name	Feature Family	Composite Indicator (CI)
Sum Entropy	GLCM	5.81
Percentage Density	Density	5.52
Dim 5*	Box counting Fractal Dimension	5.33
Sum Variance	GLCM	5.26
Beta 3*	Power-law	5.23
Safmp	Fourier Features	5.18
Beta 1*	Power-law	5.18
Variance	GLCM	5.14
Dim 4*	Box counting Fractal Dimension	5.11
Beta 7*	Power-law	5.09
IMC 2*	GLCM	5.06
Maximum Correlation Coefficient*	GLCM	5.01
Dim 1*	Box counting Fractal Dimension	4.99
Correlation*	GLCM	4.96
Sarms	Fourier Features	4.91
Beta 5*	Power-law	4.81
Rrms	Fourier Features	4.80
Rfmp	Fourier Features	4.72
Global Minkowski Dimension*	Minkowski Fractal Dimension	4.71
Dim*	Box counting Fractal Dimension	4.66

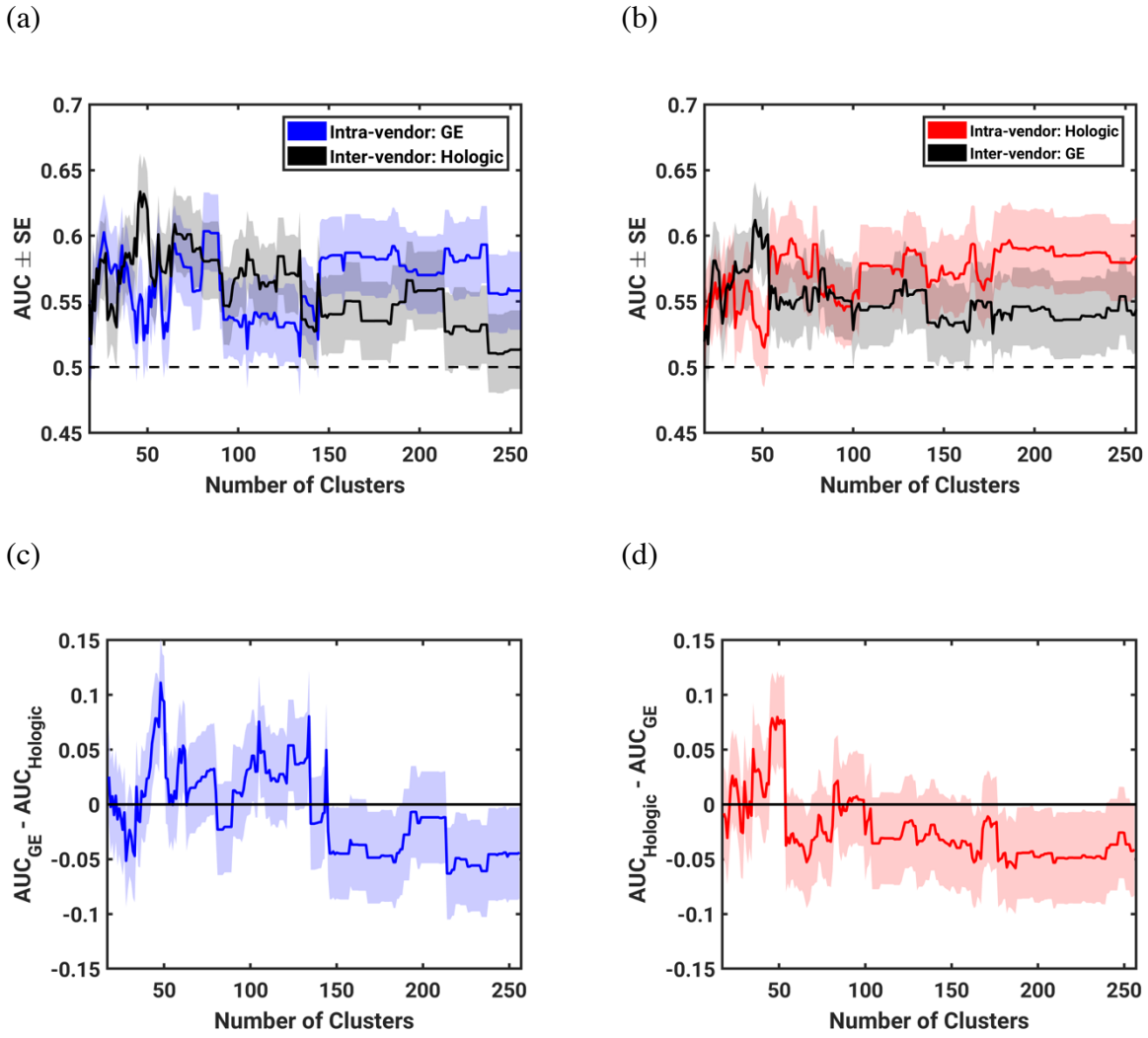


Figure 3.4. Resulting performance of classifiers trained on varying quantities of clusters and therefore varying degrees of stringency on the robustness of input features. Parts (a) and (b) show performance of intra- and inter-vendor feature selection and classifier construction as the number of clusters, and therefore stringency on robustness, is varied. Parts (c) and (d) show the difference between intra- and inter- vendor classifier performance to demonstrate generalizability. The task of each classification was predicting the risk of breast cancer. Parts (a) and (c) show results for when GE is designated M1 and Hologic is designated M2. Parts (b) and (d) show results for when Hologic is designated M1 and GE is designated M2.

As the number of clusters increases, more features were considered as candidate features. As more features were considered as candidate features, generalizability across vendors tended

to diminish. This is demonstrated by presence of a monotonic trend in the difference between intra- and inter- vendor classification performance as the number of clusters increased as shown in Figure 3.5 and (Mann-Kendall $p < 0.001$, Thiel Sen Estimator = -0.000321 and -0.000191 for GE and Hologic, respectively).

While these trends are interesting in a research setting, a fixed number of clusters would be more useful for practical application of RACE in subsequent radiomics studies. In this study, 46 clusters yielded peak inter-vendor classification performance when RACE is repeated with either GE or Hologic as M1. This is illustrated by a peak in the inter-vendor curves of Figure 3.4 parts (a) and (b). Therefore, while the full range of numbers of clusters can help describe trends in performance, this study will also include discussion of the particular application of RACE using 46 clusters.

		Training Data (M1)			
		GE		Hologic	
		Mann-Kendall P-Value	Thiel-Sen Estimator	Mann-Kendall P-Value	Thiel-Sen Estimator
Testing Data (M2)	GE	**p = 0.0168**	**-0.0000586**	**p < 0.001**	**-0.0000997**
	Hologic	**p < 0.001**	**-0.000296**	**p < 0.001**	**0.000120**
	Difference	**p < 0.001**	**-0.000321**	**p < 0.001**	**-0.000191**

Figure 3.5. Results of the Mann-Kendall test for the presence of monotonic trends and the Thiel-Sen Estimator of such trends for the performance as a function of the number of clusters. Statistically significant values are denoted by boldface font. Colored results (blue, red) correspond to intra-vendor comparisons using GE and Hologic images, respectively. Gray results correspond to inter-vendor comparisons.

3.3.2. Examination of Selected Features

Many of the same features were selected both when GE was designated as M1 and when Hologic was designated as M1. As illustrated by Figure 3.6, over half of the features selected on a given vendor's data were also selected when RACE was repeated on the other vendor. Considering that 256 total features were investigated, the commonalities across the two selected feature sets suggest that features selected are descriptive on images from both vendors. However, there are some instances in which features were not selected in both analyses. For example, the feature entropy was selected when using clusters from GE, but not Hologic. Energy was selected when using clusters from Hologic, but not GE. While these features are calculated by different formulas, each describes image homogeneity. Furthermore, these features are each highly correlated with one another. Therefore, while features selected may have varied, the features selected tended to come from the same feature families over the two vendors.

Not all features included in
Restricting candidate features to just the most robust features was also shown to have a significant impact on inter-vendor classification performance in the task of risk assessment. The classification performance was observed to monotonically decrease as the number of clusters increased (Mann-Kendall $p < 0.001$) as shown in Figure 3.4. An increase in the number of clusters corresponds to a reduction in the stringency of robustness criteria.

Table 3.3 were necessarily selected for inclusion in the classifier. Reasons for this may include redundancy or lack of discriminatory ability of the feature of interest. Namely, robustness is not a sufficient condition for inclusion in the final set of selected features. Features that are robust and not redundant with other feature candidates are passed from the robustness assessment to classification evaluation step of RACE, but in order to be included in the final set of features, the features must also

have discriminatory power in the clinical task in order to be selected by the stepwise feature selection method.

Figure 3.6. Summary of features selected for the classifier when RACE is performed either with GE designated as M1 or Hologic designated as M1. Colored boxes indicate that a certain feature was selected when data from the particular vendor images was used. The results presented in this figure are specifically from selection after grouping features into 46 clusters, as it provides the best inter-vendor performance for each manufacturer. Selected features were recorded from each leave-one-out iteration during stepwise feature selection, and the 18 features most frequently selected for each manufacturer is recorded here. Different colors are used to indicate different feature categories.

For instance, it can be observed from
Restricting candidate features to just the most robust features was also shown to have a significant impact on inter-vendor classification performance in the task of risk assessment. The classification performance was observed to monotonically decrease as the number of clusters increased (Mann-Kendall $p < 0.001$) as shown in Figure 3.4. An increase in the number of clusters corresponds to a reduction in the stringency of robustness criteria.

Table 3.3 and Figure 3.6 that while the 20 most robust features did not include any edge gradient, first-order or histogram features, some features from each of these categories were ultimately chosen for inclusion in the final classifier. This may happen because while the RACE method gives preference to the most robust redundant features, it does not remove features such as those with

moderate robustness from the set of candidates. If a feature with moderate robustness were clustered with features that had lower robustness, that moderate feature would be considered in stepwise feature selection, and thus may be ultimately included in the final model. This can be illustrated by the selection of the radiomic feature of minima, which is a histogram feature with a CI of -0.70, suggesting that it is below average in terms of its robustness. Minima was clustered with features including average, maximum CDF, minimum CDF, seventy percent CDF, and thirty percent CDF. Each of these features had a CI between -1.05 and -1.60, suggesting lower robustness than minima. Thus, minima would be the most robust feature of its cluster and would be considered in the next stage of classification evaluation.

Conversely, highly robust features are not guaranteed to be selected in stepwise feature selection for inclusion in a final feature set. For example, the box counting fractal dimension feature Dim1 was highly robust with a CI of 4.99. As the most robust feature in its cluster, it was considered in feature selection. However, during stepwise feature selection, Dim1 was not selected for inclusion in the final model, perhaps suggesting that it is not descriptive in the clinical task of risk assessment.

3.3.3. Comparison to Harmonization Methods

Classification performance of the RACE method was compared with the ComBat harmonization method used in previous studies.[127,128] The results, summarized in Figure 3.7, suggest that while the two methods perform similarly on intra-vendor comparisons, the performance of RACE was significantly different from ComBat harmonization method on inter-vendor comparisons when training on GE images and testing on Hologic images in the task of risk classification. However, given the overall low performance of each classification, these remarks are intended as preliminary observations as opposed to finite conclusions.

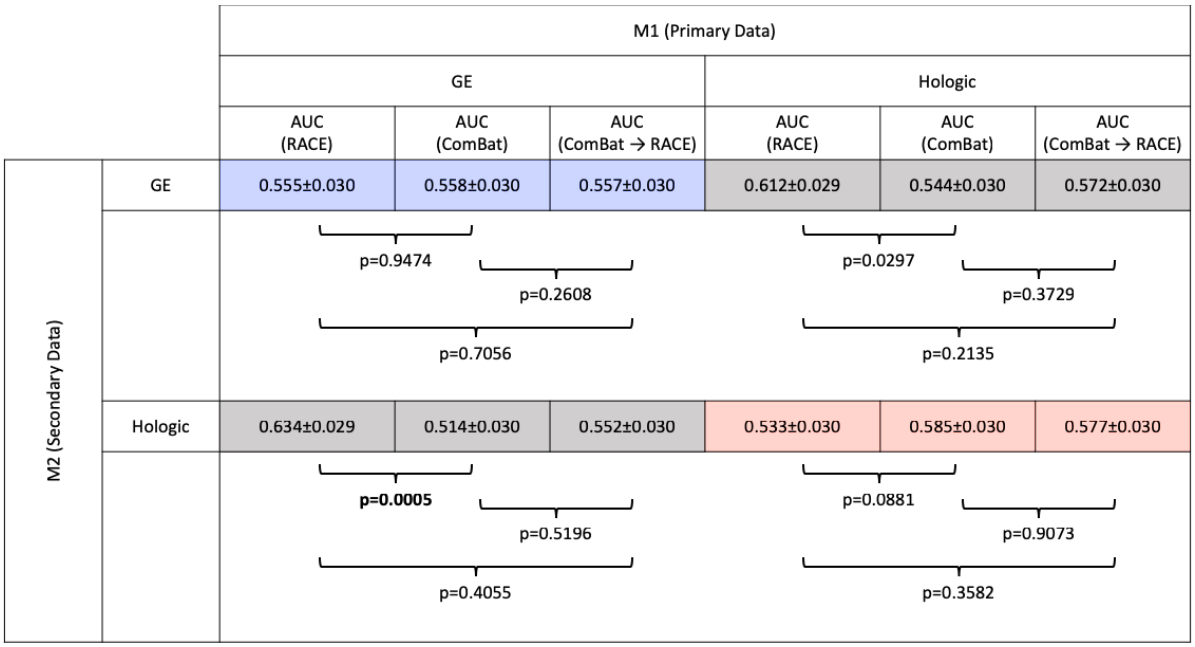


Figure 3.7. Performance in the task of classifying presence of risk factors of breast cancer of three analysis methods: (1) RACE, (2) ComBat, and (3) ComBat followed by RACE. In each method, 18 features were included in the ultimate radiomic signature construction, and leave-one-out cross-validation was performed. While intra-vendor comparisons failed to demonstrate significant differences between the three methods, inter-vendor comparisons did demonstrate significant differences, with the two-stage method performing better as judged by the AUC in the task of risk classification. M1 refers to the vendor on whose image features were selected and M2 refers to the vendor used to assess generalizability. By the Holm-Bonferroni correction for multiple comparisons, $p \leq 0.017$ is required to demonstrate statistical significance.

When ComBat harmonization is applied and followed by RACE, the results failed to demonstrate significant differences in performance from either ComBat alone or RACE alone in the task of risk classification. This trend held for each of the four combinations of training and testing data investigated in this study

In comparing the robustness metrics between raw feature values and harmonized feature values, it was observed that the MFR, which characterizes the agreement in feature magnitude, was significantly changed ($p < 0.001$), while the correlation coefficient of the feature values between vendors was not significantly different before and after harmonization ($p = 1.000$) as

shown in Figure 3.8. This result suggests that ComBat harmonization acts to improve agreement in feature magnitude across vendors but does not impact the correlation of features across the two systems.

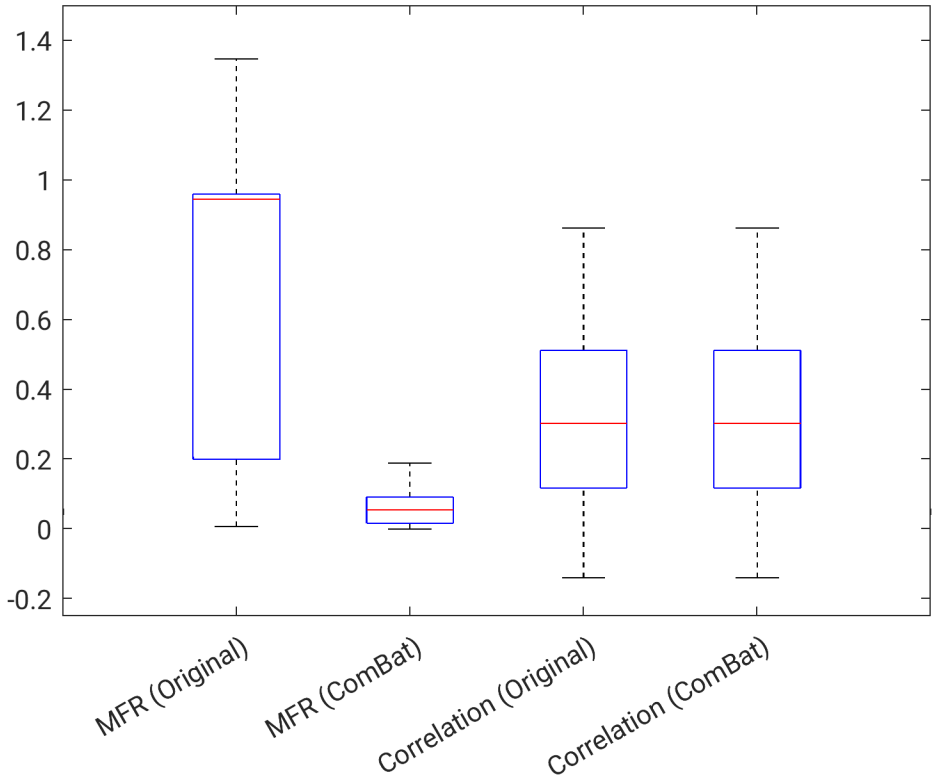


Figure 3.8. Summary of trends in robustness metrics computed on features before and after ComBat harmonization. MFR near zero indicates high robustness, and correlation near 1 indicates high robustness.

3.4. Discussion

This study found that many features describing spatial characteristics tended to be robust. For example, box counting fractal dimension features were observed to be highly robust over vendors. Fractal dimension characterizes the roughness and self-similarity of images.[87] In the case of breast parenchyma, this suggests that fractal dimension may describe the complexity of

dense tissue pattern as it appears in the mammogram. Power-law features were also observed to be robust over the two vendors in this study. Previous studies have suggested that power-law features are related to the background parenchymal pattern of breast structure.[60,132,133] The power-law exponent, β, indirectly characterizes the frequency content of the texture pattern and can be mathematically transformed to fractal dimension.[88,134] Therefore, it would be expected that power-law features would demonstrate similar trends to fractal dimension in terms of robustness. Derivative statistics from the GLCM matrix also demonstrated high robustness over mammographic units. GLCM features describe spatial relationships between pixels. By calculating how frequently pairs of pixels with specific values in specific spatial relationships occur throughout the region of interest, descriptors such as energy, entropy and correlation were computed.[70] Studies that have investigated the robustness of GLCM features over varying region segmentations and bin sizes for discretization also found that GLCM features demonstrate high robustness.[135]

It is likely that the technical characteristics of the Hologic and GE unit used to acquire images influence the feature values extracted and thus also influence the robustness of such features. A more focused examination of specific technical parameters that differ between the two units used in this study can be found in a study by Mendel et al.[109] Briefly, in Chapter 2, it was shown that the two units differ in pixel size, anode material, detector size, detector material, and conversion method as shown in Table 2.2.[109]

The monotonic trends observed in this study suggest that considerations of feature robustness in feature selection tends to improve generalizability of models across vendors. While a trend was observed, the nature of this trend was not explored beyond the monotonic direction (increasing or decreasing). In standard practice, robustness and repeatability of radiomic features

across imaging machines is typically not evaluated or included in the feature selection process. The results of this study suggest that conventional methodology may not be reproducible on data acquired on a different machine. This may also pose problems in studies that perform classification using data from multiple acquisition systems, if consideration is not given to feature robustness. However, the overall low performance of both methods render this as an initial observation in need of further verification on different tasks. The overall low performance may be attributed to the difficulty of this specific classification task. Use of mammograms of asymptomatic women to predict whether she has a high or low risk of breast cancer has not seen high performance in previous studies addressing this task.[60,65,101]

While this study observed significantly different inter-vendor results between RACE and ComBat harmonization, this could be due in part to the intended use of the two algorithms. While this study developed the two-stage method with the goal of developing a texture signature that is robust in inter-vendor comparisons, the study by Orlhac et al. only performed classifications using single features, as opposed to feature signatures.[127] Specifically, the calculation of a radiomic signature may accommodate for normalization differences over the two vendors, thus reducing the utility of a harmonization step prior to signature calculation. Therefore, this study suggests that RACE is useful in producing inter-vendor radiomic signatures, while ComBat may be useful when performing classifications with a single feature.

When applying ComBat harmonization followed by RACE, we failed to show a significant difference between the performance compared to using RACE alone in the task of risk classification. The failure to observe a significant difference suggests that after performing feature harmonization, further application of robustness assessment does not significantly impact the classification performance in the task of classifying risk. Likewise, this suggests that adding

the standardization step of feature harmonization does not significantly change the performance of robustness assessment for classification. A possible explanation for this could be that by following ComBat harmonization with RACE, feature reduction is being performed on the candidate feature set prior to stepwise feature selection. Thus, it is possible that the stepwise feature selection algorithm employed in this study is better suited to applications on already reduced feature sets. Another possible explanation could be the limited size of our dataset. Future studies could investigate benefit of stepwise feature selection when applied on feature sets containing various numbers of features.

Importantly, the pixel size of images collected on the two systems was not consistent. As demonstrated separately by Mendel et al. and Mackin et al., varying the pixel size impacts the radiomic features calculated from images.[109,111] In harmonizing pixel size for feature calculations, Mendel et al. observed reduced feature robustness following pixel interpolation on FFDM to produce consistent pixel size, compared to no pre-processing.[109] Likewise, Mackin et al. observed increased feature variability following voxel resampling on computed tomography images to produce consistent pixel size, compared to no pre-processing.[111] Previous studies have used methods such as image resizing followed by Butterworth filtering, low-pass filters, band-pass filters, or Gaussian filters in order to harmonize images prior to feature calculation.[111,136–139] In this study, we chose not to apply these steps to force consistent pixel size as the presented method seeks to address image heterogeneities through feature selection as opposed to image processing. However, we acknowledge that some post-acquisition harmonization steps may improve feature agreement over imaging conditions.

The two-stage feature selection process (RACE) is not without limitations. First, reproducibility datasets consisting of patients imaged on separate machines are relatively

uncommon in medical imaging. Therefore, this initial investigation applied the methods to only one task on only one dataset. In general, the methods could be applied to a number of applications if repeatability data is available.

Additionally, this study applied the methods to a challenging task, in which high performance is not necessarily expected. Several studies have used radiomic texture analysis to address risk of breast cancer based on screening images, however even studies with well separated patient populations (i.e. unilateral cancer versus low-risk) had only moderate performance[66]. Previous studies have also measured reductions in classification performance when women with different types of risk factors are analyzed together (e.g. classifying BRCA2 versus low-risk controls compared with classifying BRCA1/2 versus low-risk controls). In this study, the group of patients with high risk factors present had a range of factors including BRCA1/2 gene mutations, family history of breast or ovarian cancer, and personal history of ADH, meaning that this group was likely heterogeneous in its true overall lifetime risk of breast cancer. Therefore, there existed greater variability within groups in this study that may have contributed to low performance values. This study did not test the classification of specific risk factors from low-risk controls because of the limited database size. Furthermore, differences other than presence of high-risk factors existed between the two groups. One such measured difference was the difference in mean age (54.3 and 49.7 years for risk factors absent and present, respectively). Parenchymal texture has been observed to change over a woman's lifetime, and therefore this confounding factor could impact the results of this study.

Another consideration for further optimization of the methods in this study includes closer examination of the optimization of ROI size and location. In this study, regions with size 512x512 pixels were placed in the central region directly behind the nipple as this location was

shown to perform well in previous risk assessment studies.[61] However, because of differences in database and analysis, these parameters are not necessarily optimal in the present study. Thus, while it has not been proven that ROI size and location used in this study are optimal, their utility in previous studies makes them logical choices. This study used radiomic features of the breast parenchyma to predict risk. Dense component, if present in the breast, is typically located in the central region immediately behind the nipple. Thus, the location used is a practical choice as it was where the tissue of interest is typically located. Furthermore, the study by Li et al. found that classification performance did not significantly decline as the ROI size changed. Instead, the study reported that there was no statistically significant difference observed as the size of the ROI decreased.[61] Therefore, substantial differences in the outcome of this study would not be expected if a different ROI size were used.

As this was a retrospective study, the imaging units on which images were collected are no longer considered state-of-the-art. The GE Senographe 2000D unit was first released in 2000. Compared with newer GE units, the GE Senographe 2000D has a smaller field of view and lower detective quantum efficiency (DQE) and normalized noise power spectrum (NNPS) due to improvements in electronic noise in latter models.[140] The Hologic Selenia is also different from later models, as the unit used in this study had a molybdenum-molybdenum (Mo-Mo) target-filter. This target-filter material has been shown to result in higher average glandular dose compared with molybdenum-rhodium (Mo-Rh) or rhodium-rhodium (Rh-Rh) target-filter combinations.[141] This is because Mo-Mo target-filter combinations result in a softer x-ray beam. Mo-Mo has also been shown to result in lower contrast in dense breasts compared with the other target-filter material combinations, making it less optimal.[141] Newer Hologic systems use a tungsten-silver (W-Ag) target-filter combination, which results in a harder x-ray beam.[142] These

physical differences in image acquisition between the models used in this study and the models used clinically today may cause differences in image feature values and appearance, yet the methods would likely remain relevant for varying image parameters or system vendor.

Additionally, the average mean glandular dose (MGD) of the two vendors' units is different. As reported by Hendrick et al., the GE system had a MGD of 1.69 mGy per view and 4.02 mGy per exam, and the Hologic system had a MGD of 2.50 mGy per view and 5.03 mGy per exam.

3.5. Conclusion

This study proposed a two-stage method (RACE) for robust radiomic signature construction. RACE was demonstrated in the task of breast cancer risk assessment. The results suggest that feature generalizability monotonically decreases as feature reproducibility over vendors decreases. This trend shows that considerations of feature robustness could improve classifier generalizability in multifarious datasets collected on multiple mammography units. Furthermore, the same trend was observed when either vendor was used for feature clustering, thus supporting that this finding can be generalized. An investigated harmonization method (ComBat) was not shown to have strong classification performance when used on its own, but when ComBat harmonization was followed by RACE, classification results appeared similar to RACE alone. Thus, harmonization steps in conjunction with robustness assessment warrant future investigation in feature selection and classifier construction methods. In conclusion, implementation of the RACE method for robust classification was shown to lead to improved classification performance over harmonization steps alone.

CHAPTER 4

Transfer Learning from Convolutional Neural Networks for Computer-Aided Diagnosis on DBT and FFDM Breast Images

4.1. Introduction

As discussed in Chapter 1, DBT has emerged as a promising modality to improve screening sensitivity and accuracy. DBT produces pseudo-3D images by rotating an x-ray source in a partial arc around the breast while acquiring projection images. A growing number of studies have shown that tomosynthesis significantly reduces screening recall rates and increases cancer detection rates.[29][30][31][32] By providing volume data as opposed to single projection images, DBT gives a clearer visualization of regions of interest by minimizing overlaying tissue compared with 2D FFDM. Therefore, DBT is expected to be particularly useful for women with dense breasts for whom overlaying parenchymal tissue may obscure breast lesions.[33] However, human observer studies inherently involve qualitative and subjective interpretations. The objectivity of computer vision methods may therefore provide supportive evidence concerning the use of DBT in breast cancer screening.

The growing adoption of DBT in screening protocols makes the prospect of CADx on DBT images clinically impactful. Therefore, it is informative to compare performance on DBT to that on FFDM. Several groups have studied CADe of lesions using DBT images with conventional radiomic methods, yielding promising results.[143][144][145] These conventional methods are being superseded in some applications by emerging artificial intelligence approaches such as deep learning.

Deep learning is a machine learning method that is rapidly growing in usage in the image processing field. Deep convolutional neural networks (CNNs) have seen the most widespread use in object detection and image classification tasks.[146] These methods involve computing high dimensional, unintuitive features from large databases. This contrasts with previous CADx and CADe research that compute relatively small numbers of handcrafted intuitive features as CNNs can extract features through convolutional, pooling, and connected layers.[63][147]

Deep learning is now being used in medical imaging classification tasks.[148–151] Compared with natural object sets such as ImageNet[152], annotated medical datasets are limited in size. To handle small databases, approaches for medical classification tasks typically involve transfer learning through the application of a pre-trained CNN. The pre-trained CNN is typically intended for multiclass object classification on a database such as ImageNet, as illustrated in Figure 4.1.[153] Essentially, pre-trained neural networks act as feature extractors for image sets in different domains. Different domains typically have different population characteristics and different classification categories, thus necessitating a classifier such as a support vector machine (SVM).[150][149] Transfer learning has been applied in DBT lesion detection tasks, with applications for detecting both masses and calcifications.[154,155] Transfer learning has also been applied to lesion characterization with DBT, however comparisons with FFDM, synthesized 2D images, and key slice images were not performed.[156]

In order to compare the efficacy of transfer learning-based CADx on DBT and FFDM, transfer parameters were used to build classification models for each image type. Evaluation of the performance of deep learning features on FFDM and DBT images may provide further support in the utilization of extending deep learning-based CADx to DBT applications. This may improve the precision and accuracy of characterizing breast lesions. The aim of this study was to

provide an objective comparison among the diagnostic performance of FFDM, the synthesized 2D image, and the DBT key slice in the tomosynthesis cine loop through CADx in differentiating malignant from benign breast lesions. This type of comparison is innovative as while it is common to compare performance over different algorithms, comparison of performance across different image types is relatively unexplored.

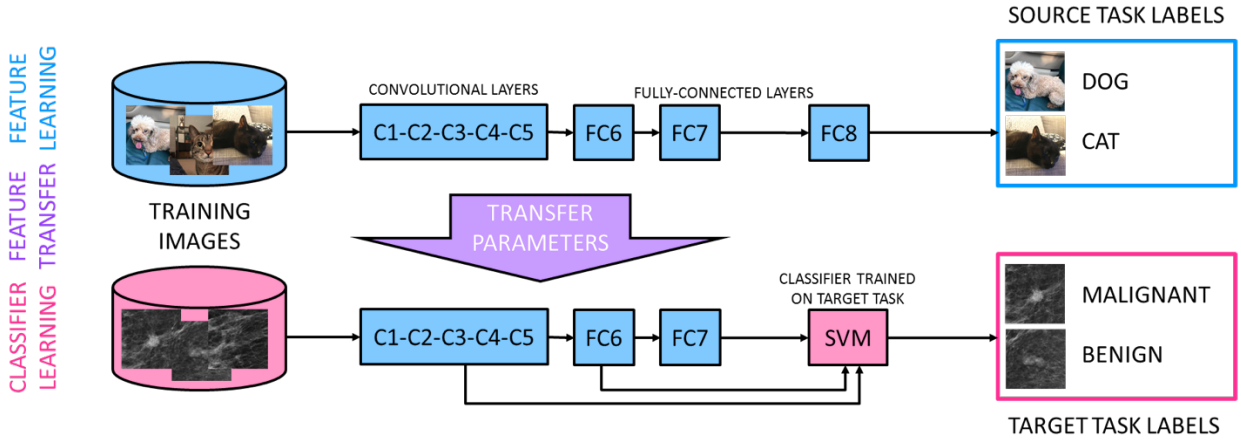


Figure 4.1. Illustration of the general deep learning approach of transfer learning through feature extraction. Parameters are transferred from a pre-trained neural network. Features are then extracted from the various network layers on input images from a separate domain, such as medical imaging.

4.2. Materials and Methods

4.2.1. Image Acquisition and Database Description

A retrospective review was performed on all patients who had undergone both FFDM and DBT in a single visit at the University of Chicago, and the subset of exams that resulted in a mammographically detected lesion that was ultimately biopsied for final surgical pathology was identified. FFDM and DBT imaging were performed on a Hologic Selenia Dimensions unit (Marlborough, Massachusetts, United States). All aspects of the diagnostic workup were performed at the University of Chicago Medicine, and images were retrospectively collected

under HIPAA-approved and IRB-approved protocols. A total of 76 patients with 78 lesions were included in this study, with exams ranging in date from August 2015 to June 2017. The average age of included patients was 54.7 years (standard deviation = 10.1 years). Of the 78 lesions, 30 lesions were biopsy proven to be malignant and 48 lesions were biopsy proven to be either high risk or benign. A summary of patient and lesion characteristics is included in Table 4.1. Each lesion was identified in the CC and MLO view on the 1) FFDM image, 2) synthesized 2D image, and 3) DBT key slice in the tomosynthesis cine loop. A fellowship-trained breast imager manually identified the key slice of each lesion from the tomosynthesis cine loop. For mass lesions, the key slice was defined as the slice nearest to the center of the lesion with the largest lesion diameter and/or when the lesion was best in focus. For architectural distortion lesions, the key slice was defined as that in which the largest number of spiculations were seen and/or in which the lesion was best in focus. For calcifications, the key slice was defined as that in which the greatest number of calcifications were in focus. We acknowledge that manual selection of a key slice may introduce bias in analysis, particularly if the mass lesion is not circumscribed or the calcifications are not along the plane of image acquisition. In future work, methods of evaluating the full lesion volume could reduce this source of bias and should be explored. Such methods may have the potential to further improve classification performance beyond that observed in this study.

In terms of lesion categorization, the high-risk lesions and the benign lesions were grouped into the “benign” category. All high-risk lesion patients either ultimately underwent surgical excision or had at least two years of imaging follow-up. No patients were upgraded to malignancy in this high-risk lesion category. Thus, the task of interest in this study was to classify lesions as malignant or benign.

Table 4.1. Summary of patient ages, lesion types, and lesion molecular subtypes.

	Frequency (%)	
	Malignant	**Benign**
Age		
≤39	--	1 (2.1)
40-49	6 (20.0)	20 (41.7)
50-59	7 (23.3)	18 (37.5)
60-69	12 (40.0)	9 (18.8)
≥70	5 (16.7)	--
Average Age (SD)	59.6 (10.3)	51.5 (8.6)
Lesion Type		
Mass	10 (33.3)	23 (47.9)
Architectural Distortion (ARD)	9 (30.0)	6 (12.5)
Calcifications	11 (36.7)	19 (39.6)
Molecular Subtype		
DCIS	14	
IDC	12	
ILC	3	
Invasive Mammary	1	
Papillary Carcinoma	1	
Atypical Ductal Hyperplasia (ADH)		7
Complex Sclerosing Lesion		3
Fibroadenoma (FA)		9
Fibrocystic Change		7
Normal Breast Parenchyma		5
Cyst		1
Apocrine Metaplasia		4
Stromal Fibrosis		3
Intraductal Papilloma		5
Sclerosing Adenosis		3
Usual Ductal Hyperplasia (UDH)		1
Total	**30**	**48**

ROIs were extracted in a dedicated workstation. CNN-based features were extracted using Keras (Version 2.1.2) with a TensorFlow (Version 1.10.0) backend framework. Model

training and subsequent analyses were performed in MATLAB (Version R2015b).

4.2.2. Deep Feature Extraction

A fellowship-trained breast radiologist identified each lesion on all three image types: FFDM, DBT synthesized image, and DBT key slice. A square region of interest (ROI) measuring 512 x 512 pixels was manually placed to fully cover the lesion on both the CC and MLO views for each image type. ROIs were then bicubically interpolated to a size of 224 x 224 pixels to conform to the size of training images used in the initial training of VGG19. Examples of malignant and benign ROIs from each image type are shown in Figure 4.2.

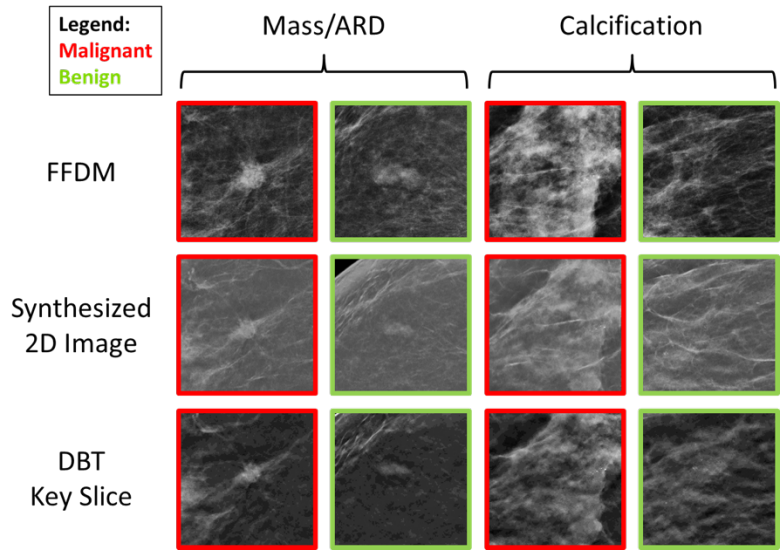


Figure 4.2. Examples of malignant and benign ROIs selected to use for classification of two masses and two calcifications. ROIs for the four example lesions are shown in each of the image types explored. Malignant lesions are outlined in red, and benign lesions are outlined in green.

The VGG19 deep CNN, which consists of 19 weight layers, was used to extract features in this study.[157] VGG19 was pre-trained on over one million images from the ImageNet dataset that consists of natural objects used for multiclass object classification.[152][147] Learned weights

obtained during pre-training were applied to the network in this study, and features were extracted from various layers of the network. While the natural images on which VGG-19 was trained involved RGB channels, the medical images used in this study were gray scale. Thus, the gray scale images were replicated in each of the RGB channels. Extracted features were used in the research task of classifying breast lesions as malignant or benign. Note that due to the small database size, the network was not trained or fine-tuned in order to avoid overfitting. Instead, images were fed through the existing pre-trained architecture, and quantitative features were extracted from various layers in the network.[158]

Features were extracted from each max pooling layer of the VGG19 convolutional network for each of these modalities, and features from each maxpool layer were fed through a meanpool layer to reduce the number of features. Following initial feature extraction, feature dimension reduction was further conducted by eliminating features with zero variance over all lesions considered in this study.

4.2.3. Feature Selection and Classification

Following feature extraction and reduction, leave-one-out stepwise feature selection was performed to identify a non-redundant set of informative features.[126] To identify such a feature set, stepwise feature selection was performed in a leave-one-out manner over the training data, with one training case left out of each round. At each round, stepwise feature selection was performed by iteratively adding and removing features from the classification feature set, using the p-value of the F-statistic as a metric to measure significance in improvement of the model. The null hypothesis was that a candidate feature would have a zero coefficient in the multilinear model, and if there existed sufficient evidence to reject the null hypothesis, then the candidate

feature was added to the model. Conversely, if there existed insufficient evidence to reject the null hypothesis, then the candidate feature was removed from the model. This iterative algorithm continued until no single step improved the model. Stepwise feature selection is described in greater detail elsewhere.[126]

After repeating the stepwise feature selection algorithm for each left out training case, the cumulative frequency of the selection of individual features was considered. The most frequently selected features over the leave-one-out iterations were selected for use in the classification model. The motivation behind this iterative method of feature selection was to keep the number of features included in models constant when comparing across image types. Stepwise feature selection on its own produces variable quantities of selected features, which might introduce bias into the evaluation of the performance of classifiers, as it has been shown that classification performance varies with the number of included features.[159] By using the frequency of selection to identify a fixed number of features, this potential source of bias was removed.

For combined analysis of masses and calcifications, the four most frequently selected features were used in the final classifier. For analysis of either mass/architectural distortions or calcifications, the two most frequently-selected features were maintained. The numbers of features used in this study were selected to be near the optimal number of features for classification with SVM for the dataset size based on recommendations by Hua et al.[124] The reduced feature set was used to train an SVM classifier with a linear kernel[160]. SVM was selected for use in this study due to its ability to handle sparse data, which is characteristic of CNN data. While SVM also has the ability to handle high-dimensional data, the use of feature selection prior to training the SVM rendered this typically advantageous characteristic unnecessary in this study. Outputs from the SVM were used to perform ROC analysis and to

determine the AUC in the task of classifying lesions as malignant or benign, which was used as a figure of merit in this study[161]. The standard error of the AUC was calculated to estimate the range of values for the population. Note that analysis was performed in a leave-one-out manner as opposed to independent training and testing sets due to the small size of available data. The resulting classification performances reported in this study are therefore viewed as an overestimation of performance, as separate training and testing may yield lower performance. However, since the aim of this study was to compare performance, this likely has a minimal impact on the study's conclusions.

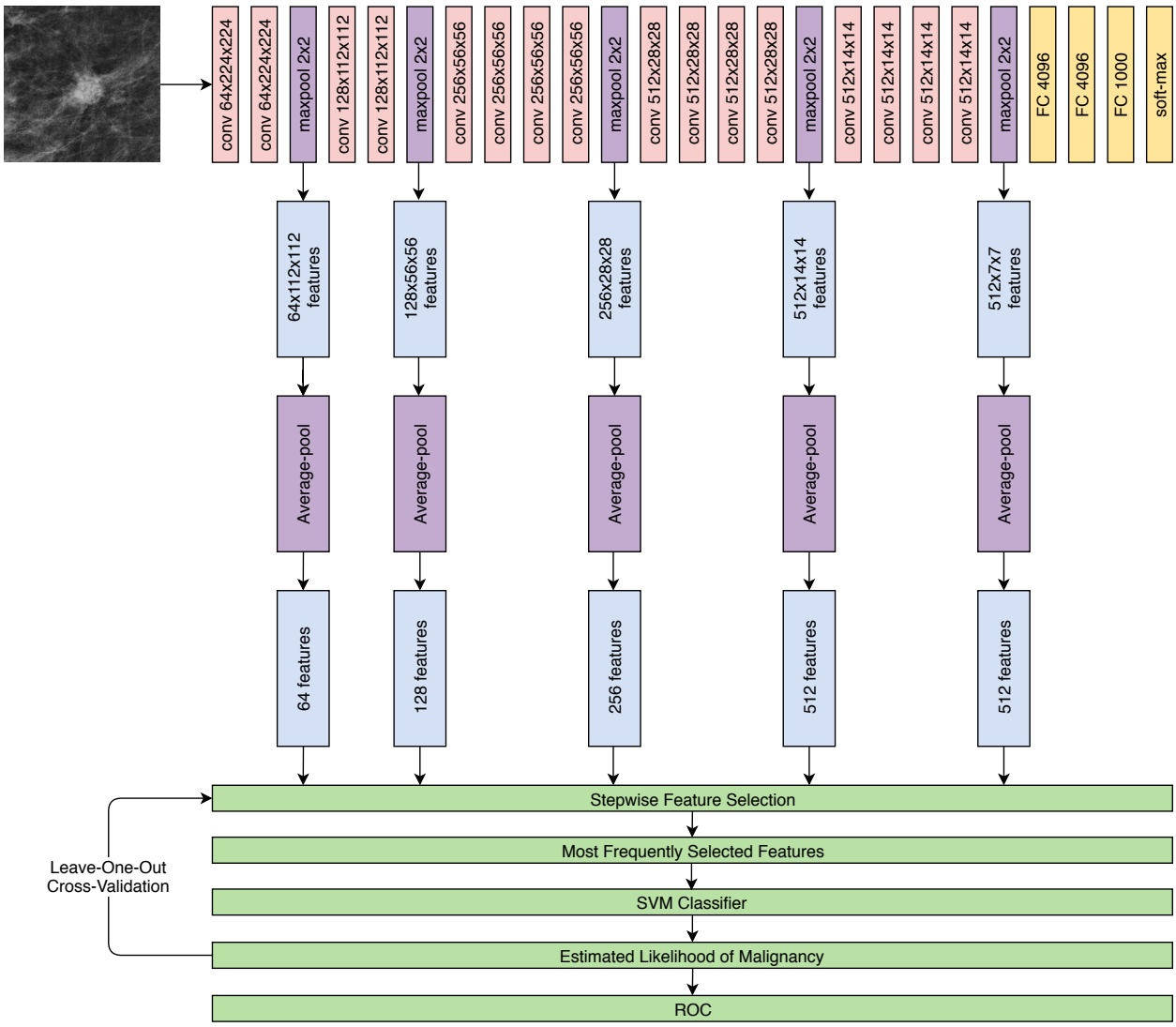


Figure 4.3. Structure of the VGG19 convolutional neural network and illustration of the layers from which features were extracted and input to the SVM classifier to yield an output classification decision in this study. Features were extracted from each maxpool layer, and an average-pool layer was applied to reduce feature dimensionality. Feature reduction was performed, and remaining features were input to an SVM classifier in a leave-one-out manner.

An extension of this analysis was needed to further understand the value of complementary information provided by the two standard screening views of each breast to the task of classifying lesions as malignant or benign. To this end, we investigated the performance of a merged-view classifier by combining signatures from the CC and MLO views of each lesion.

The merged-view classifier was constructed through soft voting of the SVM output of classifiers trained separately on the CC and MLO view images.[162] This analysis was repeated separately for each of the three imaging modalities.

The visual characteristics of masses and calcifications vary, and therefore this analysis sought to explore whether corresponding characteristics of malignancy vary as well.[163] Thus, the imaging data were additionally examined in subsets based on lesion type (mass/architectural distortion or calcifications). Training and classification were repeated following the same methodology as when performed on the full dataset in the task of classifying lesions as malignant or benign.

4.2.4. Statistical Evaluation

Statistical significance of the difference of each task's AUC from random guessing was calculated for each classifier using a statistical z-test.[164][165] The statistical significance of the difference between classifiers was evaluated using the p-value of a univariate z-score statistical test calculated using ROCKIT software.[161] Corrections for multiple comparisons were performed following the Holm-Bonferroni method.[166]

4.3. Results

Table 4.2. Summary of AUC values observed for classifying lesions as malignant or benign.

Images analyzed		All (n=78)	Masses/ARD (n=48)	Calcifications (n=30)
FFDM	**CC and MLO**	0.81±0.05	0.88±0.05	0.88±0.06
	CC View	0.76±0.05	0.90±0.07	0.83±0.08
	MLO View	0.76±0.06	0.82±0.06	0.82±0.08
Synthesized 2D Image	**CC and MLO**	0.86±0.04	0.91±0.04	0.94±0.04
	CC View	0.81±0.05	0.75±0.08	0.88±0.10
	MLO View	0.88±0.04	0.87±0.06	0.90±0.06
DBT	**CC and MLO**	0.89±0.04	0.98±0.01	0.97±0.03
	CC View	0.74±0.05	0.79±0.08	0.82±0.08
	MLO View	0.83±0.05	0.80±0.07	0.84±0.07

4.3.1. Single-View Lesion Characterization

The AUC was determined for the classification of malignant and benign lesions for each breast imaging modality (FFDM and DBT) and for each view (CC and MLO) in the task of classifying lesions as malignant or benign. The resulting AUC and standard error values are presented in Table 4.2. For the MLO view, the performance of synthesized 2D images was higher than the performance of FFDM or DBT key slice for both calcifications and mass/ARD lesions in the task of classifying lesions as malignant or benign. For the CC view, the performance of synthesized 2D images was highest for calcification lesions, and performance of FFDM was highest for mass/ARD lesions.

4.3.2. Merged-View Lesion Characterization

Lesions may be best characterized in one of the two standard views used in screening mammography (CC and MLO). Therefore, incorporation of information from both views may provide complementary information motivating this study's use of a merged classifier.

The merged classifier for DBT key slice images consistently outperformed DBT key-slice single-view classifiers in each lesion subset in the task of classifying lesions as malignant or benign, suggesting that the two views of DBT images provide complementary information. For FFDM and synthesized 2D images, the merged classifier did not consistently perform better than single-view classifiers. Thus, the merged classifier was not decidedly preferred on this dataset for these image types in classifying lesions as malignant or benign. Examples of lesions that were correctly and incorrectly classified by the various classifiers are shown in Figure 4.4.

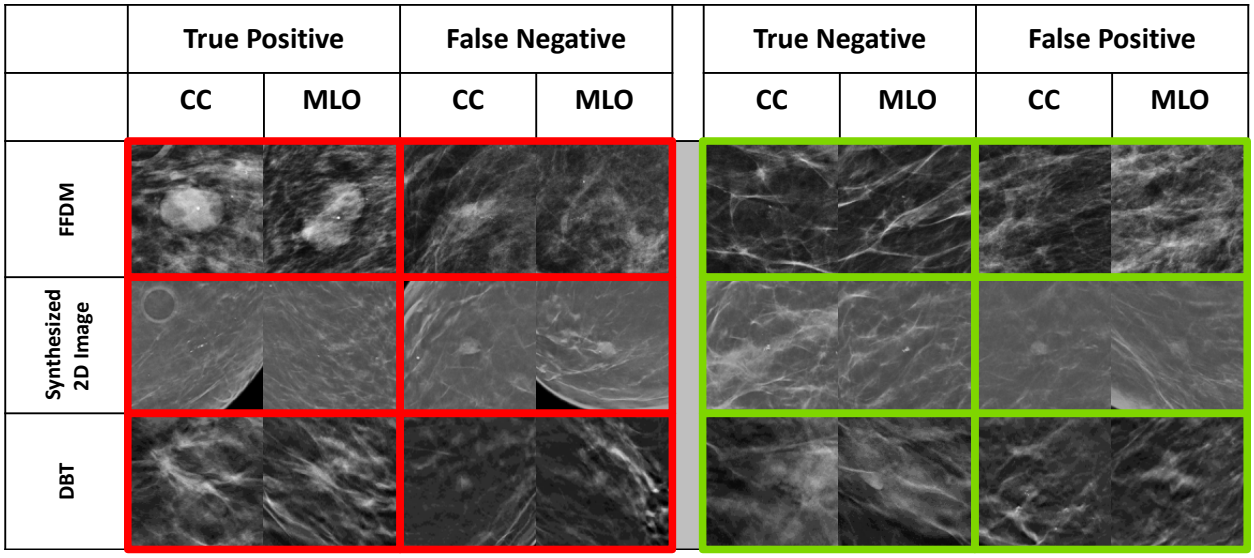


Figure 4.4. ROIs of lesions that were correctly or incorrectly classified by classifiers trained for each image type in the task of classifying lesions as malignant or benign. The most extreme lesion (i.e., highest or lowest probability of malignancy) was used to select the representative lesion shown here for illustrative purposes. Malignant lesions are outlined in red, and benign lesions are outlined in green.

Performance of the merged classifier in the task of classifying lesions as malignant or benign is reported in Table 4.2 and illustrated in Figure 4.5 and Figure 4.6. Performance of each the synthesized 2D images and DBT key slices was compared with that of FFDM in the task of lesion characterization, using the merged-view classifiers. After correcting for multiple

comparisons through the Holm-Bonferroni method, the performance of DBT key slice was significantly superior to the performance of FFDM in the task of lesion characterization.

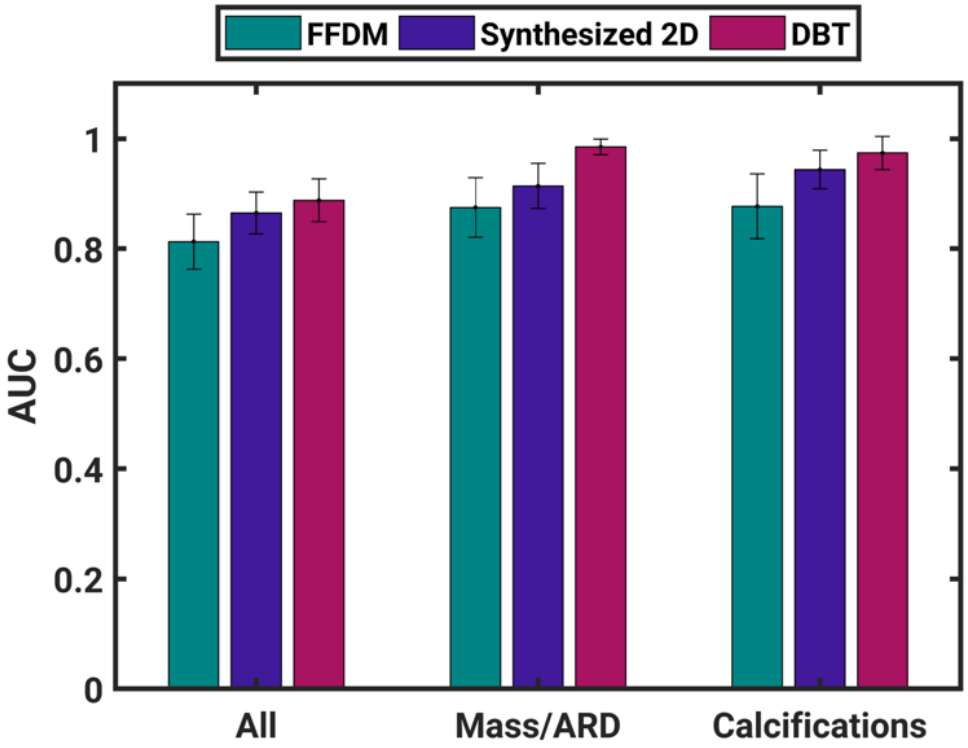


Figure 4.5. Classification performance of the merged-view classifier on each subset of lesions considered in this study in the task of classifying lesions as benign or malignant. AUC is plotted with error bars showing one standard error. This figure summarizes the performance of the merged-view classifier as reported in Table 4.2.

	FFDM	Synthesized 2D Image	DBT
All (n=78)	0.81±0.05	0.87±0.04	0.89±0.04
	p=0.77		
	p=0.04		
Mass/ARD (n=48)	0.88±0.05	0.91±0.041	0.95±0.01
	p=0.26		
	p=0.023		
Calcifications (n=30)	0.88±0.06	0.94±0.04	0.97±0.03
	p=0.13		
	p=0.07		

Figure 4.6. Significance of difference between AUC values using merged CC and MLO data for classification in the task of predicting malignancy. After corrections for multiple comparisons, a p-value of 0.025 is significant at the α=0.05 significance level.[166]

4.4. Discussion

This study involved an exploration of the potential of using pre-trained CNNs via feature extraction in the task of classifying malignant from benign breast lesions on (a) FFDM, (b) synthesized 2D images, and (c) DBT key slice images. To our knowledge, comparisons of CNN transfer learning performance across mammographic imaging modalities for lesion diagnosis has not yet been conducted. With the growing presence of DBT in breast cancer screening, it is increasingly important to understand how best to tailor CAD methodologies to best accommodate this new modality.

In this study, the application of CNNs to classifying lesions on FFDM and DBT images as either malignant or benign was explored. Transfer learning methodology was applied, using a pre-trained CNN to extract features from FFDM and DBT ROIs. The extracted features were input to a support vector machine classifier, and the diagnostic performance of resulting class probabilities was determined in terms of AUC in the task of lesion characterization.

When mass and ARD lesions are considered, DBT performed significantly better than FFDM in the task of classifying lesions as malignant or benign. This is in agreement with observer studies and conventional radiomics studies that also found that DBT had similar or better performance than FFDM in the task of lesion characterization.[31][167][168][30][169] Yet, it would be informative for a human expert to perform classifications on the data used in this study to perform comparisons with the reported results.

The increased lesion conspicuity in DBT is particularly beneficial when imaging dense breast tissue because it can be difficult to perceive suspicious lesions in extremely dense breast tissue. The border of masses, number of masses, and associated findings such as dilated ducts or vessels around a mass have been reported to be better depicted on DBT than FFDM images, especially in dense breasts.[170] Because of the ability of DBT to reduce tissue superimposition, one benefit of DBT is a reduction in the recall rate in women with dense breasts. Haas and colleagues[32] reported that the addition of DBT reduced recall rates for all breast density groups and age groups, with significant differences in recall rates for scattered heterogeneously dense and extremely dense breasts. Their study findings reiterate the conclusion that DBT is beneficial for patients with dense breast tissue and for those with nondense breast tissue. In our dataset, the DBT key slice yielded the highest AUC when individually classifying masses/ARD and calcifications, suggesting that tomosynthesis may indeed be helpful in reducing overlapping parenchymal tissue

in the analysis and classification of lesions. We acknowledge that the feature dimensionality was large compared with the number of lesions included in this study. Therefore, the results reported here are treated as initial findings and warrant further confirmation on a larger dataset.

Most findings on DBT are apparent on both the CC and MLO projections, but one-view-only findings can occur on DBT, and breast cancer still occasionally may be visible on only one projection. Previous studies involving DBT have estimated that 5–9% of breast malignancies are seen only on the CC projection, whereas 1–2% of breast malignancies are apparent only on the MLO projection.[171] Moreover, 12–15% of findings noted on both projections are more readily apparent on one view compared with the other.[172] In our study, performance tended to be higher when the CC and MLO views were merged, as shown in Table 4.2, suggesting that each view provides unique and synergistic information to aid in the classification of a lesion. When individually assessed, DBT synthesized 2D images performed better than all other imaging methods in both the CC and MLO view.

Resulting classification performances observed in this study were comparable to those reported by other DBT-based deep learning CADx studies. For example, while an independent data set was used, Samala et al. observed an AUC of 0.90 in the task of classifying mass lesions through an evolutionary pruning approach[156]. However, this study did not compare classification performance of DBT to that of FFDM or 2D synthesized images. Therefore, while the study by Samala et al. provides support for the feasibility of transfer learned deep CNNs for CADx on DBT images, the results observed in our study complement this finding by comparing performance over different image types. Similarly, Kim et al. implemented latent feature representation of breast lesions on DBT on a dataset independent from the one used in this study, and observed an AUC of 0.919 in characterizing breast masses, which agrees with the results of

this study.[173] Additionally, a study by Antropova et al.[149] demonstrated that multi-time-point inclusion from dynamic contrast-enhanced MRI into the three color channels of a pre-trained CNN. This approach could be applied to incorporate three key slices of the DBT image set.

4.5. Conclusion

While this study focused on investigating deep learning methods, the approach taken in this study could be extended to incorporate radiomics features. Repeating lesion characterization on DBT images using a standard radiomics approach may explicate whether radiomics and deep learning extract redundant or complementary information. Characterizing the relationship between these algorithms may be constructive in developing CADx systems for clinical use in breast imaging. While such an investigation would clearly be of value, this study omitted such a comparison as it focused instead on comparing value of breast image types, as opposed to comparing computer vision algorithm methodologies. As more images are collected at our institution, we plan to incorporate more sophisticated deep learning methods such as fine tuning and training from scratch. By continually improving CADx of breast lesions, we hope to improve diagnostic accuracy and patient management for breast cancer patients.

CHAPTER 5

Long Short-Term Memory Recurrent Neural Networks for Risk Prediction on Time Series on FFDM

5.1. Introduction

While agencies such as the American College of Radiology, American College of Physicians and American Cancer Society have different recommendations for breast screening frequency guidelines, they all suggest mammographic screening with some frequency over some portion of a woman's lifetime.[174–176] Women who follow these guidelines thereby produce temporal sequences of mammographic images. When interpreting screening exams, radiologists may compare current mammograms with prior mammograms to qualitatively assess interval change of breast tissue. This is done because interval change may be related to the development of breast cancer.[177]

It has been demonstrated that comparing current and prior mammograms improves the performance of screening. A study that compared performance on over one million images found that the use of comparison mammograms at screening resulted in lower recall rates (6.9% with comparison mammograms vs. 14.9% without comparison mammograms) and higher specificity (93.5% with comparison mammograms vs. 85.7% without comparison mammograms).[177] This suggests that in ambiguous cases where it is not obvious whether an abnormality poses a threat, the changes in mammograms over time provide the radiologist with discriminatory information that helps inform the decision of whether or not to send a patient for follow-up. For example, if a suspicious region is judged to be visible and unchanged in prior mammograms, then this may lower the risk of malignancy as judged by the radiologist. The utility of prior images in

radiologist review suggests that prior images contain useful information, and thus it is possible that they may also be informative in cancer prediction system whose goal is to assist the radiologist in detecting and diagnosing cancer based on imaging.

Previous studies have also explored the incorporation of prior imaging exams in clinical classification tasks. A study by Santeramo et al.[178] implemented a time-modulated LSTM network to detect abnormalities in a database of 745,480 chest x-rays. The study compared the performance of a CNN (Inception v3) trained on single images as a baseline to an LSTM network using the single images plus prior longitudinal observations. The clinical task in this study was to classify chest x-ray abnormalities as either cardiomegaly, consolidation, pleural effusion, or hiatus hernia. Using the F-measure as a figure of merit, the study observed, on average over the four abnormality types, that the LSTM resulted in a 7% increase in F-measure over the baseline CNN, and a 9% increase in PPV over the baseline CNN. Thus, while this study investigated temporal analysis for a different disease type and site, it provides evidence that temporal imaging information may inform detection of disease.

Another study by Shao et al.[179] investigate the use of temporal radiomics to predict the development of white matter hypersensitizes among elderly patients with normal-appearing white matter. This study constructed radiomic signatures on regions of interest covering various tissue types among a patient cohort each of which had undergone two or more MRI exams on the same scanner with a time period of at least one year between scans. The study reported an AUC of 0.954 (95% confidence interval: 0.876-0.989) when using ROIs placed to cover normal white matter and developing normal appearing white matter. This study provides evidence that computer analysis of temporal images may assist in predicting future disease.

In response to the importance of interval change on detection and diagnosis of breast cancer, this study applies radiomics and deep learning methods to the task of classifying future lesions as malignant or benign. Interval change may be related to future cancer incidence and may be an indicator of early carcinogenesis.

In order to incorporate information collected over a time series of FFDMs, we chose to use a long short-term memory (LSTM) network in this study, as it is capable of learning long-term dependencies for data organized as a series. As a recurrent neural network (RNN), LSTM networks are able to retain information about previous time points in a series and use this information to inform decisions on present time points of that same series. LSTM networks can take in feature vectors from various sources, and so this study explored the performance of an LSTM trained on features extracted from a CNN and the performance of an LSTM trained on conventional handcrafted features extracted from the same images. Utility of LSTM networks for use in dynamic contrast-enhanced MRI has been previously demonstrated.[180] Specifically, Antropova et al. demonstrated higher classification performance on lesion characterization with MRI using LSTM than using a fine-tuned feed-forward network with a single time point.[180] Additionally, the performance of LSTM is compared with the performance obtained by extracting features from a single time-point and merging features using SVM. In this way, a comparison is performed between deep features and conventional features as well as between time series data and single-time-point data for classification. This study compared LSTM network performance with a single time point with features merged using SVM as this method has shown strong performance on lesion characterization tasks as demonstrated in Chapter 4.

5.2. Materials and Methods

5.2.1. Image Acquisition and Database Description

Mammograms were retrospectively collected from women who underwent screening exams for two or more years prior to detection of a mammographic abnormality. Images were obtained at MD Anderson Cancer Center under a prospective study and at University of Chicago Cancer Center under a retrospective review. All images collected at each institution were collected in compliance with the Health Insurance Portability and Accountability Act and under institutional review board-approved protocols.

For each patient exam, the CC images of the left and right breast were retrospectively collected and used in analysis. Each patient included in this study ultimately underwent core biopsy with histologically confirmed findings of a malignant or benign lesion. However, all images used in this study were acquired prior to the detection of a mammographic abnormality. The laterality of the mammographic abnormality was noted, and the affected and contralateral breasts were treated separately in the analyses.

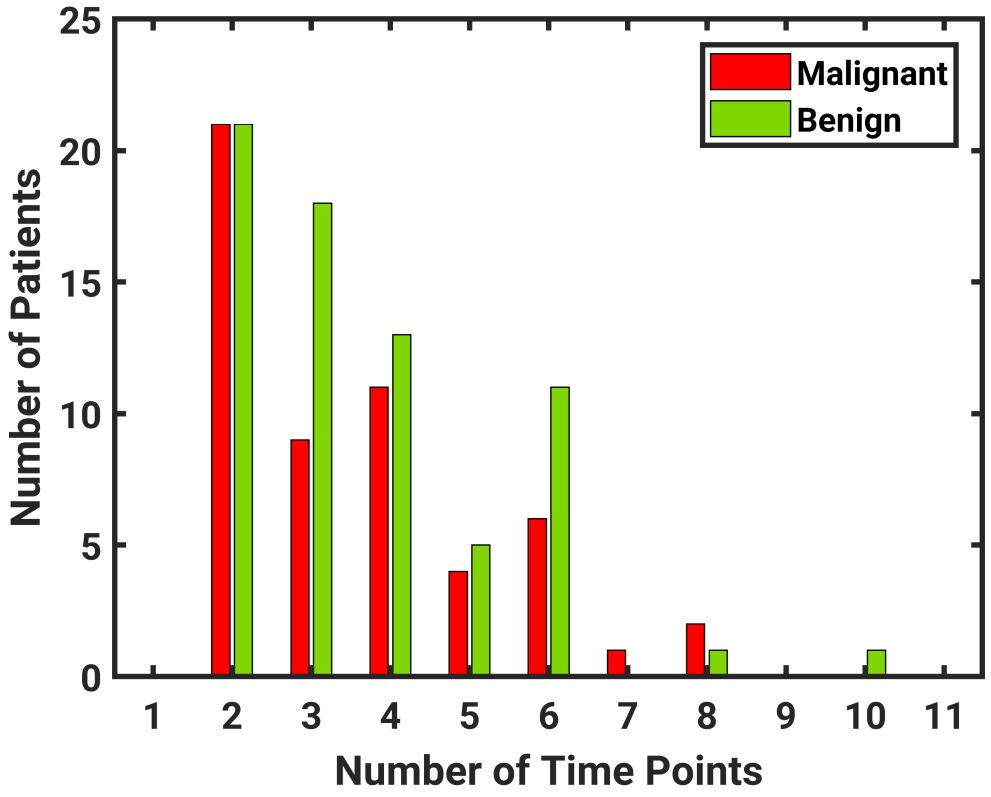


Figure 5.1. Histogram of the number of time points included in the study for patients with either malignant or benign lesions. All images included were acquired prior to the screening exam which led to diagnosis.

The number of mammographic exams per participant ranged from 2 to 10, and the distribution of the number of exams is shown in Figure 5.1. An example of a temporal series of images collected for a single patient is shown in Figure 5.2. Note that the period of time between subsequent screening exams was not always constant for each patient. The average time between exams was 1.19 years.

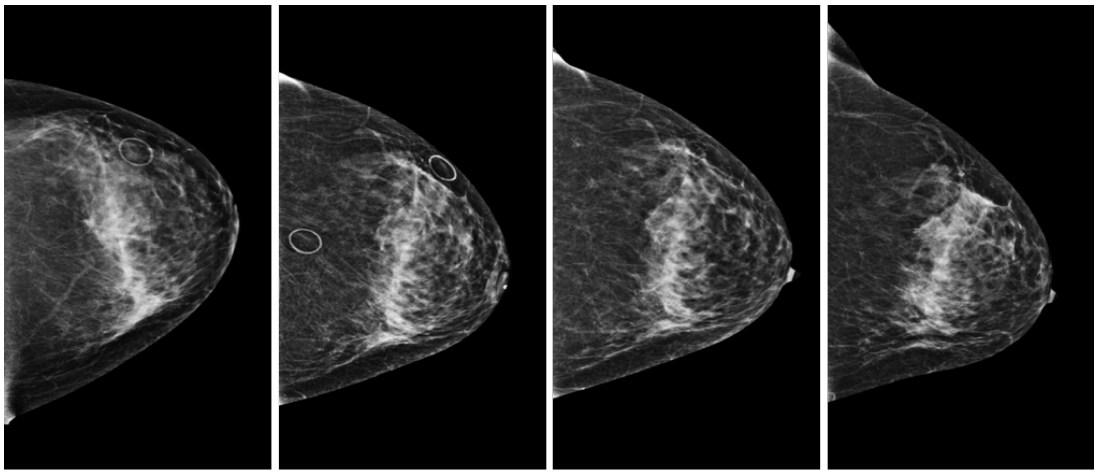


Figure 5.2. Temporal mammograms for one patient, collected annually over a span of four years. Note that the orientation of the breast changes in each image as does the presence of markers.

A total of 453 images from 125 patients were collected. Of these, 55 patients were diagnosed as malignant and 70 were diagnosed as benign. Of these, 87 patients were screened at MD Anderson Cancer Center and 38 patients were screened at the University of Chicago Medical Center. Of those screened at MD Anderson Cancer Center, 25 were malignant. Of those screened at the University of Chicago Medical Center, 30 were malignant. All images were collected on Hologic systems with pixel size of 70 μm x 70 μm and were processed according to the clinical standard at the patient's screening institution.

Table 5.1. Ages of patients included in this study, separated by malignant or benign lesion findings.

	Malignant	**Benign**
Mean age (SD)	57.2 (9.9)	54.5 (9.4)
Age (year)		
<40	2	3
40 to 49	13	23
50 to 59	18	26
60 to 69	14	13
70 to 79	7	4
≥80	1	1
Total	55	70

5.2.2. Region of Interest Placement

Due to the highly deformable nature of breast tissue, the positioning of breast tissue may vary widely between exams, especially in the 2D projection view produced by FFDM. Therefore, registration of physical regions in the breast over time points is an important yet challenging first step in computerized analysis of changes in breast parenchyma over time in both radiomic texture analysis and deep learning analysis.

Previous approaches taken to register time series of mammograms range from local to global. Some basic intensity-based registration methods include rigid and affine registrations, which involve translations and stretching of images in order to improve alignment.[181] These geometric transformations are applied to the full image in a global manner. Local registration methods include b-splines, polyrigid and Demons registration.[181] These methods act locally on images, modifying different regions of the image in a different manner to achieve improved registration. While local methods may be appropriate to handle the registration of highly deformable breast tissue, they can be highly time intensive. Therefore, it has been suggested that a multi-resolution approach, in which resolution of the images is iteratively improve until the

original image resolution is achieved, may converge to optimal registration performance in less time. In our comparison on b-splines and multi-resolution registration on a subset of this database, we found that multi-resolution registration resulted in a lower mean-squared error and shorter registration time.[182]

However, in this study, it is necessary only to align a small ROI across breast images. Further, it is desirable to avoid image manipulation, including deformations as part of the image registration process so as to not alter image texture. Thus, while automatic registration may be desirable in a clinical workflow, this preliminary study used manual registration of ROI centers across mammogram images acquired at different time points. Region placement was performed in a dedicated workstation that allowed sequential images to be displayed simultaneously to improve manual region placement. Placement was performed in a dedicated temporal radiomics workstation illustrated in Figure 5.3, by a research assistant with four years of breast imaging research experience.

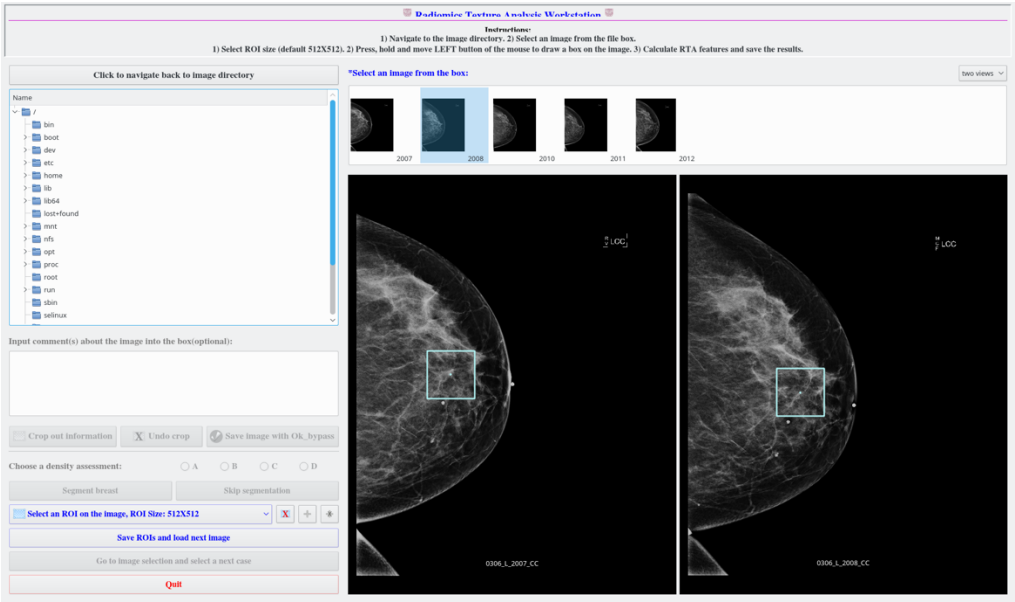


Figure 5.3. Dedicated temporal workstation illustrating ROI placement on images acquired of a single patient at two different time points.

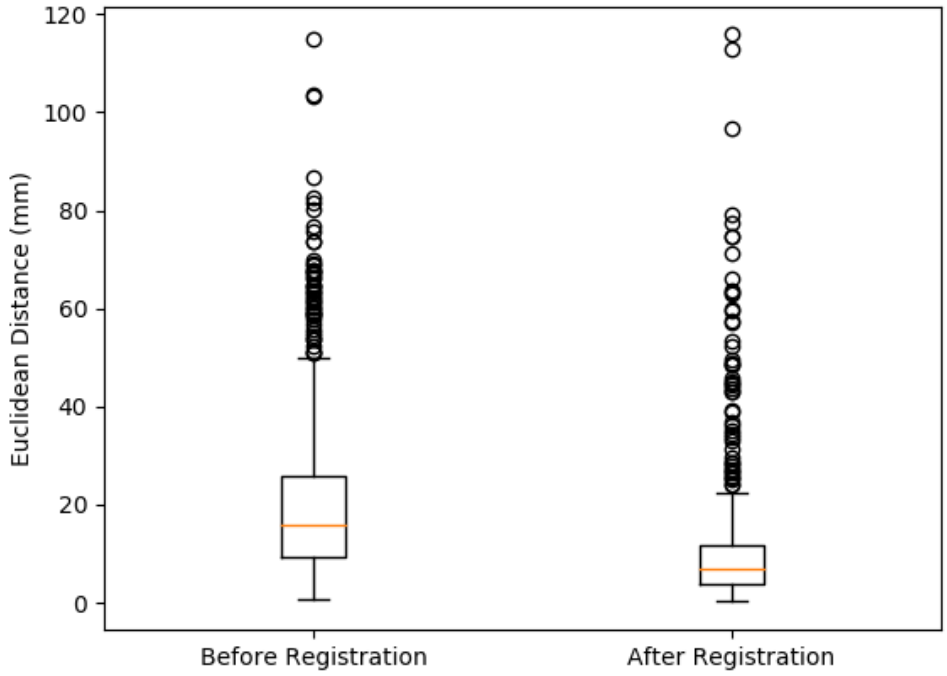


Figure 5.4. Boxplot showing Euclidean distance between registered ROI center and human identified ROI center.

An initial investigation into automatic mammogram registration over temporal sequences was performed. In this study, the gold standard was human alignment.[182,183] Application of automatic mammogram registration resulted in substantial differences between the manual ROI placement and registered ROI placement, as illustrated in Figure 5.2. Thus, qualitative human alignment based on anatomical landmarks was used for regions in this study. However, future efforts to explore additional mammogram registration algorithms to work towards improved automatic ROI placement is encouraged

5.2.3. Radiomic Feature Extraction

Following the methods described in Chapter 3, radiomic texture features were calculated on square ROIs of size 512 × 512 pixels that were manually placed in the central breast region posterior to the nipple. From each region, 45 features were extracted (Table 5.2). These features are described in more detail in Chapter 1.4.

Table 5.2. Summary of features included for analysis in the radiomics feature set.

Feature Category	**Number of Features**
Box counting fractal dimension	6
Edge gradient	4
Histogram	12
Minkowski fractal dimension	1
Powerlaw beta	8
GLCM	14
Total	**45**

5.2.4. Deep Feature Extraction

Deep learning-based feature extraction was performed on the same ROIs used for radiomic feature calculation. Similar to the work described in Chapter 4, features were extracted using the

pre-trained VGG-19 neural network. Features were extracted in the same way as the study described in Chapter 4.2.2, resulting in a total of 1,472 features for each image.

5.2.5. Long Short-Term Memory Network

5.2.5.1. Recurrent neural networks

Recurrent neural networks (RNNs) are designed for making classifications and predictions based on a time series of data. RNNs are composed of a series of identical feed-forward neural networks. In this series of networks, each individual network is used to analyze a single time point and is known as an RNN cell. Each RNN cell produces a recurrent output that is passed on to the next time step. Likewise, each RNN cell accepts a prior state as input. In this way, information from prior time-points informs the output of future time-points.

Mathematically, an RNN cell can be represented by Equation 5.1, where s_t is the current state, s_{t-1} is the prior state, x_t is the current input, and f is the recurrent function. Thus, a basic single layer RNN can be written as in Equation 5.2, where ϕ is the activation function, and W, U and b are the weights and biases of the network.

$$\begin{pmatrix} s_t \\ o_t \end{pmatrix} = f \begin{pmatrix} s_{t-1} \\ x_t \end{pmatrix} \tag{5.1}$$

$$s_t = \phi(W s_{t-1} + U x_t + b) \tag{5.2}$$

The general recurrent structure of an RNN is illustrated in Figure 5.5, where it is shown that information from the RNN cell for one time point in the series is passed along to the cell for the next input from the series.

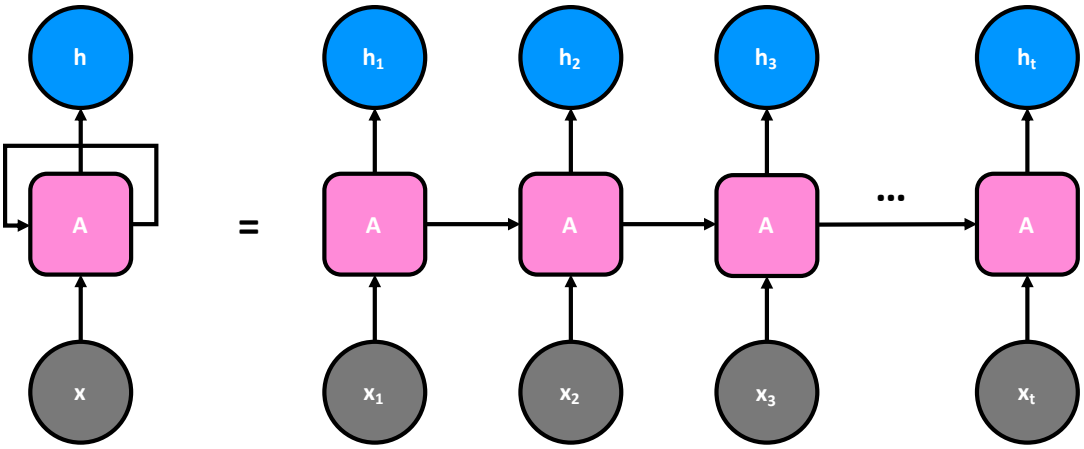


Figure 5.5. General architecture of an RNN cell component, where A represents the neural network, x_t represents some input, and h_t represents the output value

5.2.5.2. Information Morphing and the Vanishing and Exploding Gradient Problem

While useful in many applications, the simplest RNNs have weaknesses in practical use. One such problem is that of information morphing. With each step of the RNN, the network state inevitably evolves. While this evolution is typically desirable, the most useful information state may have occurred in a previous iteration and thus would be unavailable to the current state. This is known as the degradation problem.[184]

Another problem with simple RNNs arises from their use of backpropagation. Backpropagation is gradient based, and thus if the gradient of a network weight becomes very small, then the learning rate of the network will diminish. Sufficiency of basic RNNs to experience a so-called vanishing gradient can be proven by applying the Mean Value Theorem. We therefore can establish the presence, under certain conditions, of the vanishing gradient problem.[185,186] In this demonstration, s_t is the current state, s_{t-1} is the prior state, x_t is the current input, ϕ is the activation function, and W, U and b are the weights and biases of the network.

We first define s_t and z_t following the general RNN cell structure such that the activation function takes the weighted current state as input, as well as the network biases, as input to compute the updated state vector.

$s_{t+1} = \phi(z_t)$

$z_t = Ws_t + Ux_{t+1} + b$

By application of the Mean Value Theorem over several variables, there exists $c \in [z_t, z_t + \Delta z_t]$ s.t.

$\Delta s_{t+1} = [\phi'(c)]\Delta z_t$

Thus, it follows that,

$\Delta s_{t+1} = [\phi'(c)]\Delta(Ws_t)$

$\Delta s_{t+1} = [\phi'(c)]\mathrm{W}\Delta(s_t)$

Define, using the matrix 2-norm,

$$\gamma = \sup_{c \in [z_t, z_t + \Delta z_t]} \|\phi'(c)\|$$

where γ is the largest value of $[\phi'(c)]$ on the interval $[z_t, z_t+\Delta z_t]$. If the activation function is bound for all real input, then the singular values of $[\phi'(c)]$ will therefore also be bounded.

Next, taking the vector norm of each side,

$|\Delta s_{t+1}| = |[\phi'(c)]\mathrm{W}\Delta(s_t)|$

$|\Delta s_{t+1}| \leq \|\phi'(c)\| \, \|\mathrm{W}\| \, |\Delta(s_t)|$

$|\Delta s_{t+1}| \leq \gamma \|\mathrm{W}\| \, |\Delta(s_t)|$

$|\Delta s_{t+1}| \leq \|\gamma\mathrm{W}\| \, |\Delta(s_t)|$

Expanding over k time steps,

$|\Delta s_{t+k}| \leq \|\gamma\mathrm{W}\|^k \, |\Delta(s_t)|$

Thus,

$$\frac{|\Delta s_{t+k}|}{|\Delta s_t|} \leq \|\gamma \mathrm{W}\|^k$$

Notice that the logistic sigmoid function reaches a maximum derivative of ¼ at zero, and the tanh function reaches a maximum derivative of 1 at zero, thus,

$$\gamma = \begin{cases} 1/4 & for\ logistic\ sigmoid \\ 1 & for\ tanh \end{cases}$$

It then follows that,

given $\|\mathrm{W}\| \leq \begin{cases} 4 & for\ logistic\ sigmoid \\ 1 & for\ tanh \end{cases}$

we can conclude that over many time steps, the gradient approaches zero:

$$\therefore \lim_{k \to \infty} \frac{\Delta s_{t+k}}{\Delta s_t} = 0$$

Thus, we have concluded that sufficient conditions exist such that the gradient of the state vector may diminish over sequential steps of the network. Therefore, if the weight initializations for W are too small, the RNN may be unable to learn due to the vanishing gradient.

5.2.5.3. LSTM Gates

In order to avoid the potential pitfalls of information morphing and of the vanishing gradient problem, LSTM cells are designed to contain three gates that are not typically present in conventional RNNs: the input gate, output gate and forget gate. These three gates monitor the extent to which information is read in from an adjacent time point, how much of this information to write out, and to what extent the information is remembered and passed on to the next time point. The input gate (i_t), output gate (o_t) and forget gate (f_t) are defined as:

$$i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i) \tag{5.3}$$

$$o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o) \tag{5.4}$$

$$f_t = \sigma(W_f s_{t-1} + U_f x_t + b_f) \tag{5.5}$$

where s_{t-1} is the prior state, x_t is the current input, σ is the sigmoid function and W, U and b are the weights and biases of the network.

The fundamental principle of the LSTM network is that the output is incrementally changed from the prior state. Therefore, rather than computing the time-evolved state, the network computes changes in that state. The candidate time evolution of the state will be defined as $\widetilde{s_t}$. This state evolution is computed for LSTM in a way similar to the standard RNN, however we will first perform element-wise multiplication on the prior state by the output gate, notated by $\odot$. Before updating the prior state, selective forgetting is achieved by element-wise multiplication of the candidate time evolution with the forget gate. Here, s_t is the current state, s_{t-1} is the prior state, x_t is the current input, ϕ is the activation function, and W, U and b are the weights and biases of the network.

$$\widetilde{s_t} = \phi(W(o_t \odot s_{t-1}) + U x_t + b) \tag{5.6}$$

$$s_t = f_t \odot s_{t-1} + i_t \odot \widetilde{s_t} \tag{5.7}$$

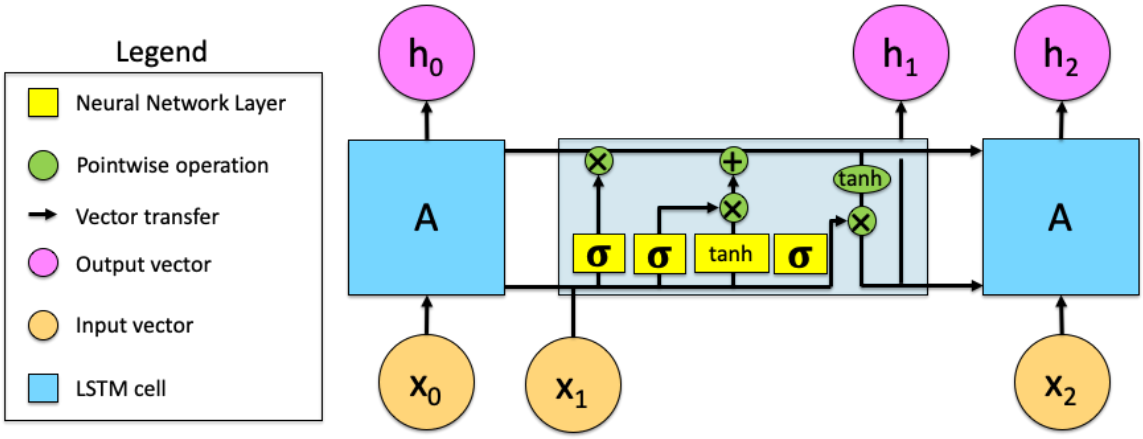


Figure 5.6. Diagram illustrating the components of an LSTM cell. Specifically, input is passed to the cell, which then is passed through sigmoid and hyperbolic tangent neural network layers. Pointwise operations are then performed to merge input with the current state vector to update the state vector and produce an output. In this diagram, x_t is an input vector at time point t, and h_t is an output vector at time point t.

5.2.6. Classification and Evaluation

In order to evaluate the value of temporal information relative to single time-point analysis, classifications were performed using both SVM (single time-point) and LSTM (multiple time-points) in the task of predicting the malignancy of future lesions. The same feature set used for training the LSTM and SVM networks. To characterize repeatability, 5-fold cross validation will be used for each classifier, with folds kept consistent over each classifier and consistent proportions of malignant and benign cases in each fold. This way, we ensure that training and testing splits were kept consistent for pairwise comparisons between classifiers. Each classifier will be trained separately on images of the affected and contralateral breasts in the task of classifying the malignancy of a future lesion using antecedent images.

ROI placement and radiomic feature extraction was performed on a dedicated workstation. CNN feature extraction and network training were performed in Keras (Version 2.1.2) using a TensorFlow (Version 1.10.0) backend framework.

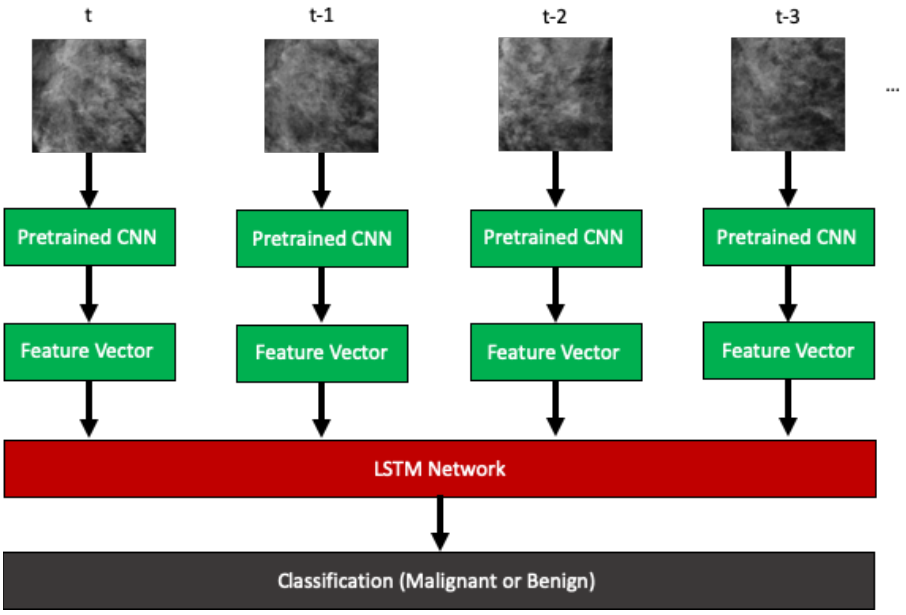
5.2.7. Temporal Sequence Classification with LSTM Network

In order to evaluate classification performance with the inclusion of multiple mammographic time-points, features extracted from each image were used as inputs to the LSTM network. To consider the value of handcrafted radiomic features compared with CNN features, separate networks were trained using each of these two as input features, as illustrated in Figure 5.7. Each classifier described was trained in the task of classifying future malignant lesions using antecedent images.

The LSTM network in this study was trained using a stochastic gradient descent (SGD) optimizer. In SDG, optimal weights are determined by choosing a random sample of training vectors and using these to compute an estimate of the gradient at each step of the training procedure. Given a random batch of training objects, the update by SGD is given by equation 5.8, where θ is the parameter to update, α is the learning rate, J is the objective function, and $(x^{(i)},y^{(i)})$ are the training feature vectors.

$$\theta = \theta - \alpha \nabla_{\theta} J(\theta; x^{(i)}, y^{(i)}) \tag{5.8}$$

a)



b)

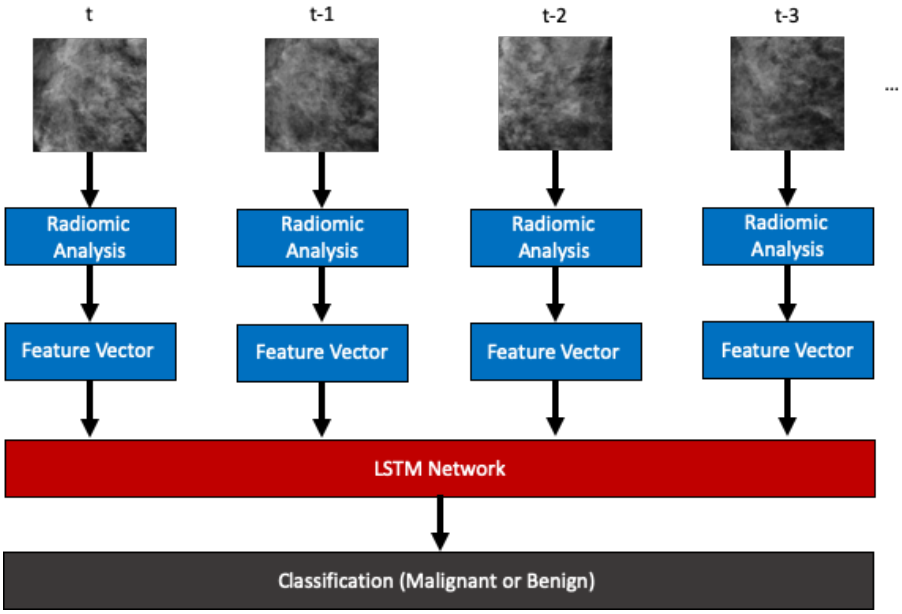


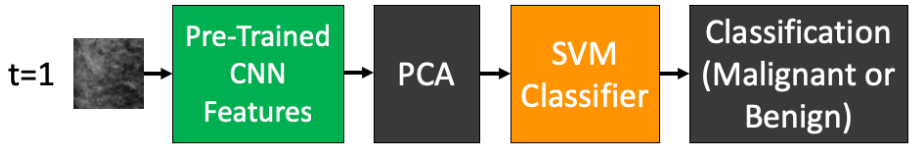
Figure 5.7. Summary of the workflow involved in using LSTM networks to classify temporal sequences of mammograms in this study. (a) Workflow for CNN-extracted features, and (b) workflow for radiomic features. Classifications were performed to predict the probabilities of future malignant lesions based on antecedent images.

Hyperparameters were selected by performing a limited sweep of learning rate and hidden dimension parameters. After sweeping over hidden dimensions of 512, 1024 and 2048, and sweeping over learning rates of 10^{-3}, 10^{-4} and 10^{-5}, the combination of parameters yielding the highest classification performance as judged by the AUC was selected for the task of classifying future malignant lesions using antecedent images. The selected hyperparameters for each set of input features is described in Table 5.3. For each LSTM network, 50 epochs were used in training.

5.2.8. Single Time-Point Classification with Support Vector Machine

In order to provide a reference for which to compare the LSTM model constructed using a time series of input images, classification was also performed using the image collected one year prior to diagnosis in the task of classifying the likelihood of malignancy of the future lesion. As only one single time point is used for classification, a support vector machine (SVM) was trained in a 5-fold out manner to construct predictions. To reduce dimensionality, principal component analysis (PCA) was performed to reduce feature space to 25 principal components prior to training the SVM. Training and classification were performed once using handcrafted radiomic features as input, and once using features extracted by a pre-trained CNN as input. This process is illustrated in Figure 5.8.

a)



b)

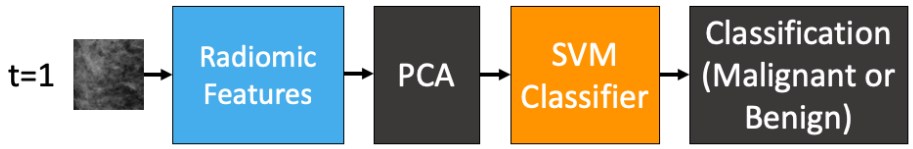


Figure 5.8. Summary of the workflow used to classify single time points of mammograms in this study. (A) Workflow for CNN-extracted features, and (B) workflow for radiomic features. Classifications were performed to classify future lesions as malignant or benign.

5.2.9. Statistical Evaluation

From ROC analysis, the AUC was used as a figure of merit in the task of predicting malignant lesions using antecedent images, and the statistical difference between the AUC values for different models was computed using ROCKIT software. A two-tailed t-test was used as there was no a priori assumption for the superior classifier.

To further investigate the role of laterality in classifying future lesions as malignant or benign using imaging phenotypes, a non-inferiority test was performed to compare the difference in AUC between the contralateral and affected breast for each of the classifications considered in this analysis. In this non-inferiority test, a maximum allowable difference to suggest non-inferiority (δ) in AUC of 0.1 was selected. Ideally, meta-analysis of literature concerning the clinical task of classifying future lesions as malignant or benign based on parenchymal analysis of the affected compared to contralateral breast would be performed to determine δ, however there exists limited works on this highly specific task, thus precluding such a meta-analysis.

Therefore, we chose to follow the suggestion of a difference in AUC of 0.1 in the demonstration of moderate differences as suggested elsewhere.[187] In evaluating the results of the non-inferiority test, the 95% confidence interval for the difference in AUC was compared with the non-inferiority boundary to determine whether noninferiority was demonstrated. If the confidence intervals included the non-inferiority boundary, then the results were interpreted as inconclusive as suggested by previous works.[188] This is illustrated in Figure 5.9.

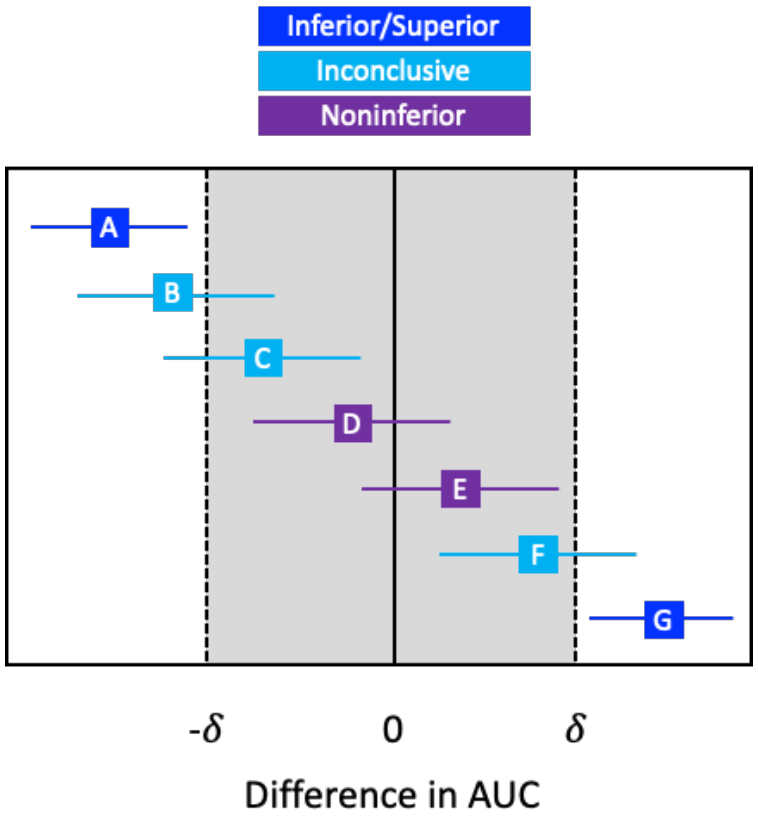


Figure 5.9. Illustration of the interpretation of non-inferiority results based on the range of the 95% confidence intervals of observed differences between AUC. In this demonstration, A would be inferior to the reference, G would be superior to the reference, D and E would be non-inferior to the reference, and B, C and F would be judged to be inconclusive.

Additionally, the odds ratio is investigated as a metric for the performance of the classifiers evaluated in this study. We chose to study the odds ratio as this is a commonly

accepted metric for clinical studies and is appropriate for retrospective studies as it is a symmetric measure of association that does not depend on sampling scheme. In other words, it does not require that the proportion of disease states represent the proportion of disease states of the greater population. Typically, the odds ratio is used to compare binary outcomes among binary or categorical exposure groups. It is computed by taking the ratio of the odds of disease state for two different exposure groups, and thus describes how many times higher the odds of disease are for one group compared to the other.

In this study, however, predictors were not binary, but rather were continuous variables. Thus, a regression was used to estimate the change in logistic odds ratio as the classifier output varies. A positive change in logistic odds ratio with increasing classifier value suggests a positive association between classifier output and malignancy. Thus, a positive value would suggest that a classifier appropriately predicts disease state. A negative change in logistic odds ratio suggests a negative association, and therefore suggests poor performance of a classifier.

5.3. Results

The hyperparameters that were ultimately selected for use in training the LSTM networks are summarized in Table 5.3. These hyperparameters were each selected as they led to the highest performance in the task of predicting future malignancy compared to the other hyperparameters considered in the grid search, with the AUC as a figure of merit in the task of predicting malignancy of future lesions based on antecedent images. Additionally, the performance of the classifier over hyperparameter space is summarized in Figure 5.10. When networks were trained with various sets of input features (radiomics, VGG-19 extracted features), different hidden dimensions and learning rates yielded higher performance. Interestingly, a larger hidden

dimension of 2048 was optimal for networks trained on radiomic features, which consisted of only 45 features per time point. However, a smaller hidden dimension of 512 was optimal for the network trained on VGG-19 extracted features when all time points were used. This suggests that, for this data, higher dimensional feature vectors performed best with a lower hidden dimensionality based on grid search results as shown in Table 5.3 and Figure 5.10.

Table 5.3. Hyperparameters selected for training of each LSTM network in this study.

Feature Set	Hidden Dimension	Learning Rate
Radiomics (All time points)	2048	10^{-5}
Radiomics (Two time points)	2048	10^{-4}
Pre-trained CNN (All time points)	512	10^{-4}
Pre-trained CNN (Two time points)	2048	10^{-3}

The performance of each classification included in this study is summarized in Figure 5.11. In general, classifiers using multiple time points with LSTM outperformed classifiers using a single time point through SVM. When features extracted from a pre-trained CNN were used, LSTM classifiers using either two or all available time points were observed to have statistically significantly different performance compared with a single time point on SVM, for each of the affected and the contralateral breast. When radiomic features were used, LSTM classifiers using either two or all available time points performed statistically significantly differently from a single time point on SVM for the contralateral breast but not the affected breast after correcting for multiple comparisons.

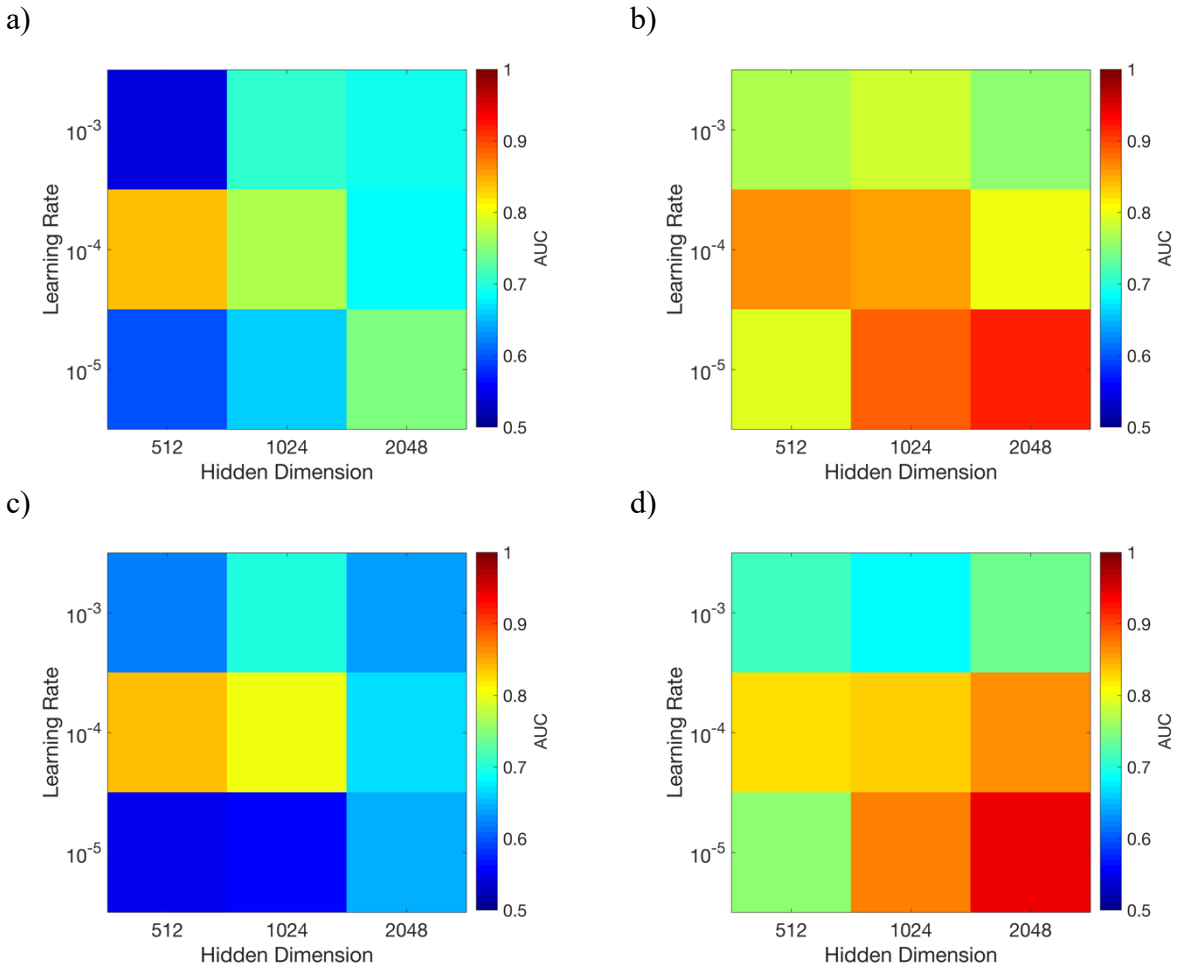


Figure 5.10. Intermediate performance output for LSTM network training on various combinations of hyperparameters. Each plot describes the (A) affected breast with CNN features, (B) affected breast with radiomic features, (C) contralateral breast with CNN features, and (D) contralateral breast with radiomic features. Color indicates the performance as judged by the AUC in the task of characterizing future lesions. Ultimately, only one hyperparameter set was selected based on the set that yielded the highest AUC. This figure shows the results for training the LSTM using all available time points, and this optimization was repeated again using only the two most recent time points.

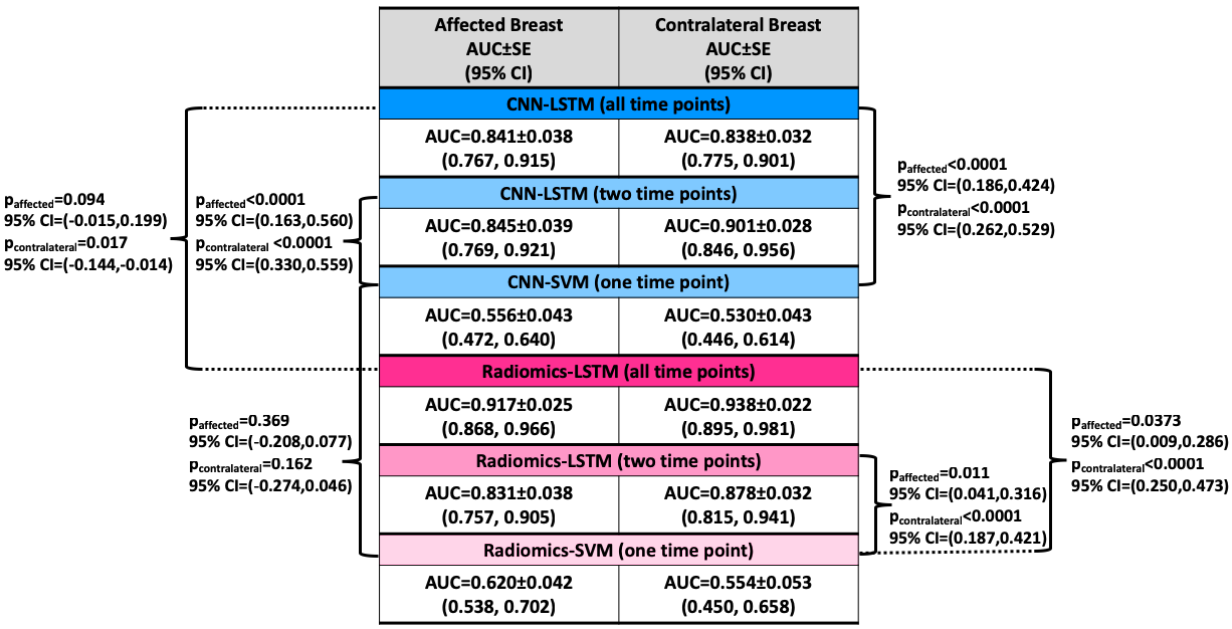


Figure 5.11. Performance of each classification performed in this work, and the p-value of the two-tailed t-test for comparison of the difference in AUC between classifications. After the Holm-Bonferroni correction for multiple comparisons, p-values of less than 0.008 suggest statistically significant differences. Additionally, 95% confidence intervals are given for the difference in AUC value for each comparison in the task of characterizing lesions as malignant or benign based on antecedent images. In each classification 125 cases were used and 5-fold cross-validation was performed.

Furthermore, we failed to show a significant difference in AUC between the LSTM network trained using CNN-extracted features and using radiomics features after correcting for multiple comparisons. This trend held for both classifications using the affected breast and using the contralateral breast.

The performance of each classifier is summarized by their respective ROC curves in Figure 5.12 in the task of characterizing the malignancy of future lesions using antecedent images. These curves demonstrate that the performance of LSTM classifiers is greater than the performance of SVM classifiers at each operating point of the curves. Notably, the LSTM curves, one based on CNN-extracted features and one based on radiomic features, cross for data extracted from images of the affected breast. This suggests ambiguity in their relative

performances, namely that CNN-extracted features perform better at low sensitivity and radiomic features perform better at high sensitivity.

a) b)

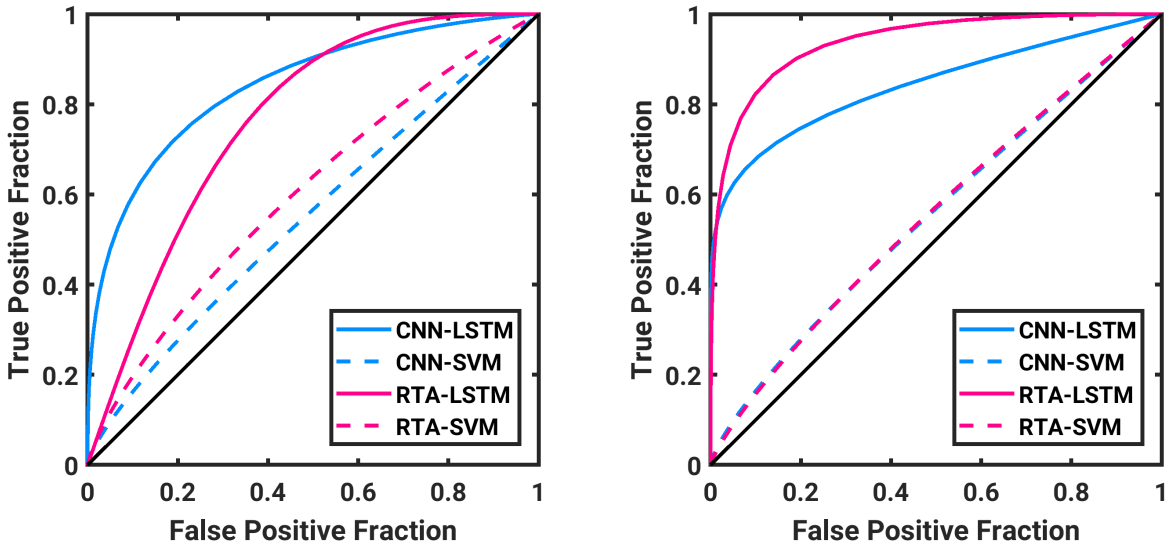


Figure 5.12. ROC curves for each comparison performed in this study. (A) shows classification performance on images of the affected breast, and (B) shows classification performance on images of the contralateral breast in the task of characterizing future lesions as malignant or benign. Performance is shown for an LSTM trained using all available time points, and an SVM trained using the single most recent time point.

In clinical practice, it is unknown whether the future lesion will develop in the right or left breast. Therefore, it is more clinically relevant to examine a merged classifier, which takes into account the classifier output on each the left and right breast in the task of predicting whether the future lesion will be malignant or benign. Furthermore, given that we failed to demonstrate significant difference between classifiers trained using pre-trained CNN-extracted features and radiomic features, it is also of interest to explore the classification performance in the task of characterizing future lesion malignancy when the output from each breast and each feature type is averaged. These results are presented in Figure 5.13. Statistical comparisons were

not performed on the merged classifier output in order to maintain statistical power by limiting the quantity of pairwise comparisons performed. However, it appears that merging information from the left and right breast tended to improve performance of the multi-time-point LSTM networks and merging information from the radiomic and pre-trained CNN classifiers also tended to improve performance in the task of characterizing lesions using antecedent images.

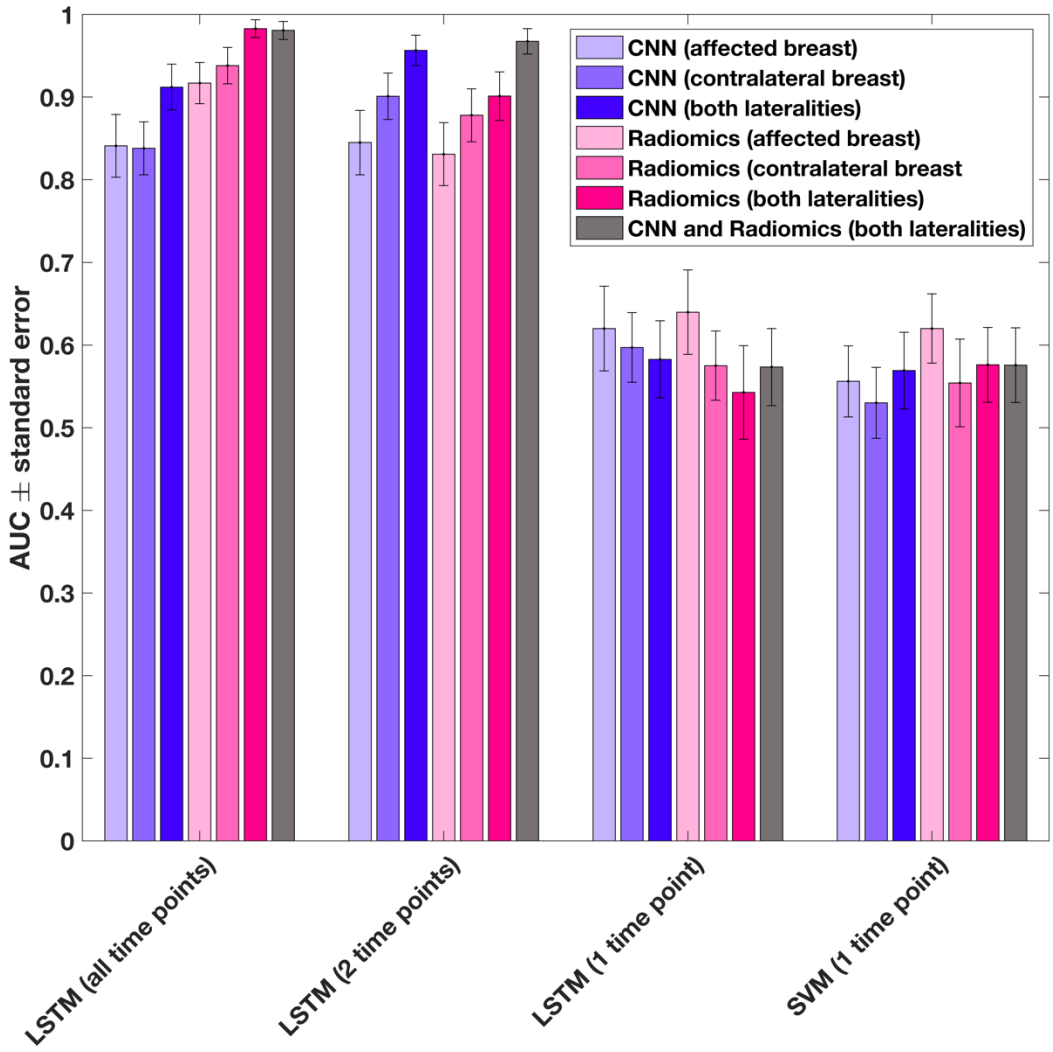
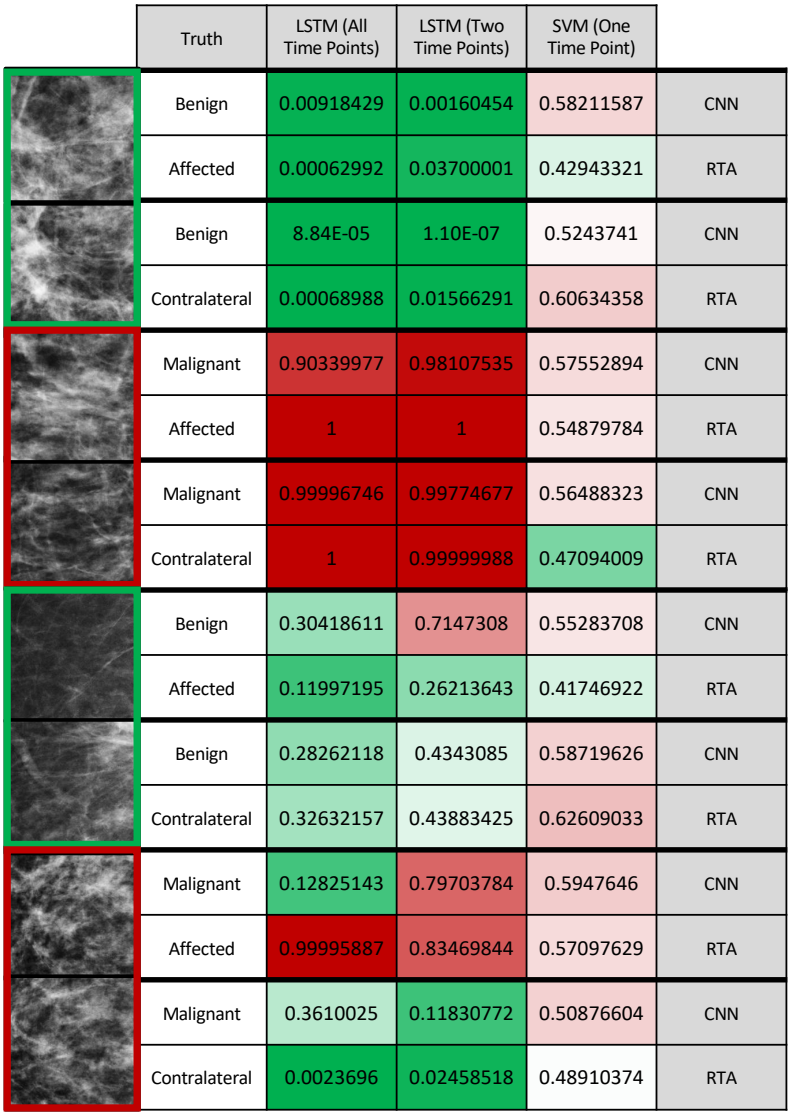


Figure 5.13. AUC values for each classifier compared, including merged classifiers. Each merged classifier was constructed by taking the average classifier output from two different classifiers for each individual case, and then performing ROC analysis on the averaged output

values in the task of characterizing future lesions as malignant or benign. Error bars show one standard error.

Example images for patients whose ROIs were either successfully or unsuccessfully classified in the task of classifying malignancy of lesions using antecedent images are shown in Figure 5.14.



Truth	LSTM (All Time Points)	LSTM (Two Time Points)	SVM (One Time Point)	
Benign	0.00918429	0.00160454	0.58211587	CNN
Affected	0.00062992	0.03700001	0.42943321	RTA
Benign	8.84E-05	1.10E-07	0.5243741	CNN
Contralateral	0.00068988	0.01566291	0.60634358	RTA
Malignant	0.90339977	0.98107535	0.57552894	CNN
Affected	1	1	0.54879784	RTA
Malignant	0.99996746	0.99774677	0.56488323	CNN
Contralateral	1	0.99999988	0.47094009	RTA
Benign	0.30418611	0.7147308	0.55283708	CNN
Affected	0.11997195	0.26213643	0.41746922	RTA
Benign	0.28262118	0.4343085	0.58719626	CNN
Contralateral	0.32632157	0.43883425	0.62609033	RTA
Malignant	0.12825143	0.79703784	0.5947646	CNN
Affected	0.99995887	0.83469844	0.57097629	RTA
Malignant	0.3610025	0.11830772	0.50876604	CNN
Contralateral	0.0023696	0.02458518	0.48910374	RTA

Figure 5.14. Illustration of several images that were either successfully or unsuccessfully classified by the various classifiers explored in the task of predicting future malignancy. For each of four patients, the ROI of the affected and contralateral breast are shown, and the probability of malignancy as output by each classifier is reported. Malignant lesions are outlined in red, and benign lesions are outlined in green.

To further investigate the differences observed between classification performance on the affected and the contralateral breast, a non-inferiority test was performed in the task of characterizing future lesions as malignant or benign. The results of this test are summarized in Figure 5.15. The 95% confidence interval for each comparison between affected and contralateral breast crossed into the non-inferiority margin, thus suggesting that the evidence is inconclusive due to low statistical power. Typically, non-inferiority margins are based on existing literature covering the task of interest, and the margin is estimated for clinical impact. However, given the limited quantity of literature investigating the task of predicting future malignancy, a margin of 0.1 was used in this study.

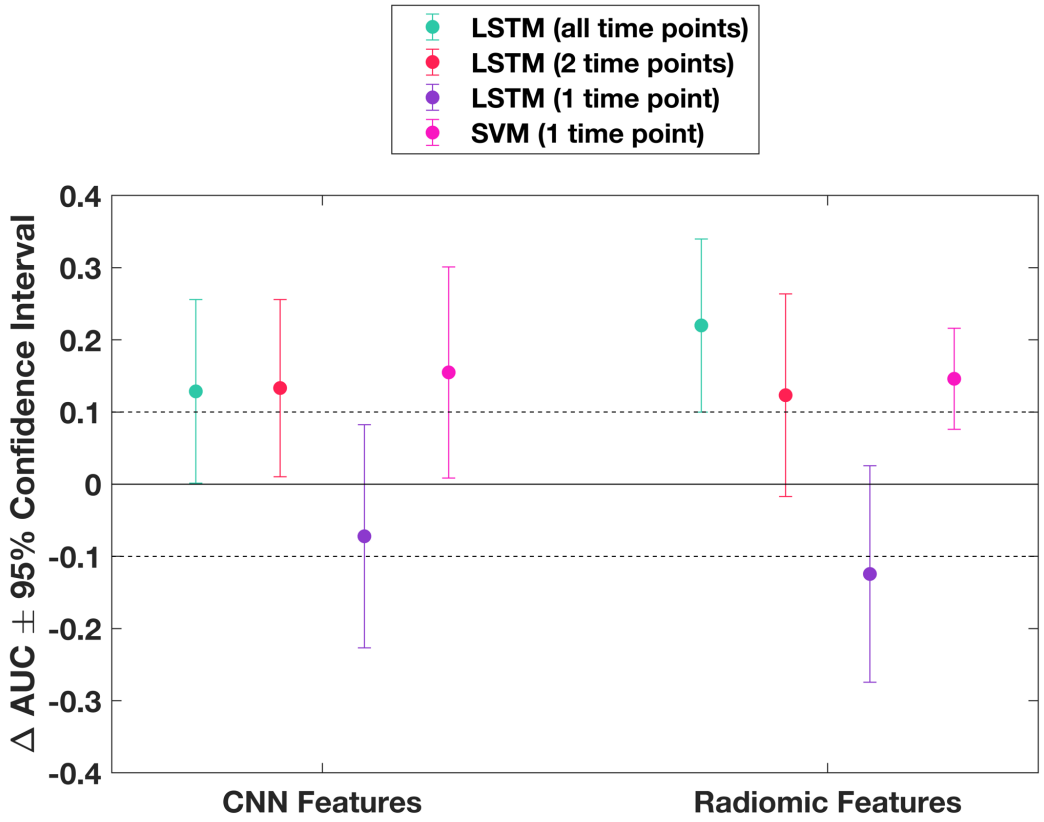


Figure 5.15. Difference between AUC calculated on the affected breast and the contralateral breast in the task of characterizing future lesions as malignant or benign based on antecedent images. Error bars show the 95% confidence interval of the difference. The dotted lines show the non-inferiority margin. Each classification comparison crosses into the non-inferiority margin, suggesting that the evidence of non-inferiority is inconclusive due to low statistical power.

To examine more closely the marginal descriptive impact of the trained classifiers in predicting malignant versus benign lesions in each of the classifiers examined in this study, the unit increase in logistic odds ratio per increase in standard error in probability score was calculated. The change in logistic odds ratio was calculated using a logistic regression and weighting according to the standard error of the classifier output.[189,190] The results are presented in Table 5.4.

Table 5.4. Unit increase in logistic odds ratio per standard error in probability score from trained classifier in the task of characterizing future lesions as malignant or benign based on antecedent images. The 95% confidence interval is shown for each value.

OR (95% CI)	**LSTM (all time points)**	**LSTM (2 time points)**	**LSTM (1 time point)**	**SVM (1 time point)**
CNN (affected breast)	0.122 (0.074, 0.172)	0.126 (0.078, 0.174)	0.042 (0.001, 0.083)	0.015 (-0.026, 0.056)
CNN (contralateral breast)	0.134 (0.083, 0.186)	0.186 (0.217, 0.250)	-0.008 (-0.047, 0.031)	0.019 (-0.026, 0.053)
CNN (both lateralities)	0.232 (0.145, 0.319)	0.397 (0.217, 0.576)	0.025 (-0.015, 0.065)	0.019 (-0.020, 0.059)
Radiomics (affected breast)	0.182 (0.123, 0.240)	0.104 (0.059, 0.150)	0.024 (-0.071, 0.022)	0.047 (0.004, 0.090)
Radiomics (contralateral breast)	0.214 (0.144, 0.283)	0.171 (0.113, 0.230)	0.052 (-0.001, 0.104)	-0.020 (-0.059, 0.020)
Radiomics (both lateralities)	0.352 (0.186, 0.519)	0.203 (0.130, 0.277)	-0.007 (-0.048, 0.033)	0.014 (-0.025, 0.053)
CNN and Radiomics (both lateralities)	0.436 (0.251, 0.621)	0.385 (0.231, 0.540)	0.002 (-0.036, 0.040)	0.020 (-0.019, 0.059)

In Table 5.4, classifiers with positive unit increase in logistic odds ratio per standard error in classifier output demonstrate positive association between classifier and disease status, thus suggesting utility of the classifier. In contrast, classifiers whose odds ratio confidence intervals include zero do not demonstrate association between the classifier output and disease status, suggesting poor performance of the classifier. Therefore, the odds ratio suggest similar conclusions to the AUC values, insofar as LSTM classifiers with two or all time points considered performed well, whereas LSTM classifiers with one time point and SVM classifiers did not perform well.

5.4. Discussion

While LSTM tended to outperform SVM classifiers in the task of characterizing future lesions as malignant or benign, exceptions existed for radiomics features of the affected breast. In this case, whether the LSTM was trained with two time points or all available time points, a statistically significant difference was not observed between the LSTM and SVM classifiers. This suggests that the marginal benefit of a classifier incorporating temporal information is greatest in cases when features are extracted from the contralateral breast and are extracted using a CNN as opposed to conventional radiomic features.

Classification performance in the task of predicting future lesion malignancy was not observed to be statistically significantly different when an LSTM network was trained using either VGG-19 features or using radiomic features. This suggests that, while these feature sets are different in their origin and how they are extracted, they achieve similar results. Thus, either feature set may be appropriate for classifications with temporal LSTM networks.

While statistical comparisons were not performed between performance on the affected and contralateral breast in predicting malignancy of future lesions in order to maintain statistical power, some general observations can be made. Generally, classification performance on features extracted from images of the affected breast was not always higher than the same classifications on the contralateral breast. Because only antecedent images were used in this analysis, no mammographic abnormalities were present. Thus, while it is possible that the affected breast had a precancerous texture change leading up to lesion detection, these results suggest that a change also occurred in the contralateral breast that may indicate future malignancy. Thus, this observation suggests that a field effect, changes in the breast extending

beyond the lesion itself, is present in both the affected and contralateral breasts in antecedent imaging leading up to detection of a lesion.

Non-inferiority tests gave inconclusive results, thus limiting our ability to draw conclusions about the relative superiority or inferiority of classifiers. With a larger dataset, we expect that confidence intervals may narrow, thus allowing conclusions to be drawn about superiority and inferiority of the classifiers investigated in this study.

This investigation into the use of temporal sequences of data for malignancy prediction involves several limitations. First, this study used a dataset of limited size compared with other implementations of LSTM networks. Curation of large data sets is more challenging and expensive in the medical domain compared with natural images, thus resulting in our small number of cases included. Additionally, the data used in this study was collected at two separate institutions. While all images were acquired on Hologic units, differences in image acquisition procedures may have varied between the two medical centers, resulting in some differences in image characteristics. Incidence rates were not matched at the two institutions, with the University of Chicago Medical Center data consisting of majority malignant lesions, and MD Anderson Cancer Center data consisting of majority benign lesions. Thus, the results observed may be an overestimate in the performance of the classifiers explored.

Additionally, the intervals at which women underwent screening were not consistent. While national agencies suggest screening at regular intervals of time, patient compliance was not consistent in the data. Furthermore, women may have undergone screening at an institution outside of the two involved in this study, and therefore this additional image was omitted from this investigation. Collecting images from consistent time intervals may affect, and potentially improve, the performance observed in this study.

The nature of screening exams involves repeat imaging on separate exam dates, thus inherently involving repositioning of the patient in the imager. As a result of this, images are not spatially registered to one another. While this may be solved through deformable registration methods, it is likely that such image processing would alter the radiomic features extracted, potentially reducing the efficacy of such features. The approach taken in this study was to manually align ROIs on undeformed images, however this method only results in approximate spatial registration across exam dates. While previous studies have shown that radiomic features tend to be only minimally impacted by small changes in spatial placement of an ROI, there may still be some effect present.[191]

Finally, note that this study compared a new method, using LSTM networks to incorporate temporal information, with a conventional supervised learning approach (SVM) that does not involve deep learning. The transfer learning approach of using SVM to merge CNN-extracted image features has also shown promise in other FFDM studies.[150,192]

5.5. Conclusion

This chapter presents an imaging-based breast cancer prediction method that captures temporal information about parenchymal texture on FFDM. These temporal sequences are used to classify future lesions as malignant or benign.

Compared with the previous methods presented in chapters 3 and 4, this work allowed for the incorporation of imaging information from multiple antecedent images, as opposed to just a single image. Thus, this method evaluated not only the appearance of the parenchyma, but also changes in the parenchyma over time. This work explored temporal network performance when

using features extracted either by conventional radiomics methods and from the VGG-19 pre-trained network.

Based on the analyses performed in this study, LSTM networks were observed to significantly outperform SVM classifiers when using features extracted using VGG-19 and to outperform SVM classifiers when using radiomic features of the contralateral breast. However, no statistically significant difference was observed between performance of LSTM and SVM classifiers when radiomic features of the affected breast were used.

The main motivation for selection of LSTM networks for use in characterizing temporal image sequences is their ability to prevent vanishing or exploding gradients during error backpropagation. Additionally, LSTM networks are well suited to handle sequences of varying length, as women have varying numbers of screening mammograms throughout their lifetimes.

The method used in this study was motivated by the fact that human experts compare current screening mammograms with previous screening mammograms to assist in the detection of abnormality. This suggests that prior images may provide additional information to the current image.[177] Thus, changes in texture over time may be indicative of an elevated probability of developing a malignant breast lesion.

The deep learning methods employed here captured temporal data patterns that are not typically examined in conventional radiomics approaches. This work has shown that the temporal data patterns capture clinically useful information in evaluating the classification of future lesions based on screening mammography.

CHAPTER 6

Summary and Future Directions

In this chapter, the main contributions of this dissertation are summarized. Limitations of the presented works are identified, and future work is suggested to address these limitations.

Chapter 2 proposed novel metrics for use in characterizing the robustness of radiomic features in pairs of images acquired under two different imaging conditions. These metrics were then utilized to identify a set of features that are robust over mammograms acquired with units from two different vendors: GE and Hologic. In general, features descriptive of spatial patterns tended to be more robust than those descriptive of image intensity or directionality. In particular, as evaluated using the proposed robustness metrics, the most robust features included box counting fractal dimension, Minkowski fractal dimension, power law beta, and the GLCM features of correlation and information. This chapter emphasized methods for evaluating robustness of radiomic features, as different features may be more or less robust on different modalities, image sites, or imaging parameters.

Chapter 3 built off of the work of Chapter 2 by incorporating measures of robustness into feature selection methods. Specifically, a two-stage method for radiomic analysis was implemented in the task of predicting risk of breast cancer. The first stage of this method used hierarchical clustering and the robustness metrics of Chapter 2 to identify a subset of features that were robust and non-redundant. This first stage functioned in an unsupervised manner. In the second stage, the utility of candidate features was evaluated in a supervised manner using stepwise feature selection. Thus, in this manner, a set of descriptive features was identified from the subset of already robust and non-redundant features. In order to demonstrate the efficacy of

the proposed method, this two-stage method is applied in the task of classifying risk of breast cancer. This study concluded that as restrictions on feature robustness were more stringent, classification performance increased in a monotonic manner. Additionally, we failed to show significant differences between the proposed two-stage method when used on raw feature values and the proposed two-stage method when used on pre-harmonized features through application of ComBat harmonization methods. While this chapter demonstrated the proposed two-stage method on paired data, this method could be extended for use with unpaired data by performing robustness assessment with metrics that do not require paired data. Therefore, application of this method suggests that radiomics-based classifications may be improved by incorporating robustness considerations.

While Chapters 2 and 3 focused on intuitive radiomic features, Chapter 4 extended beyond this to investigate the use of deep learning for image analysis. This chapter implemented the concept of transfer learning to medical image analysis in the task of characterizing lesions as malignant or benign. To handle the limited size of data sets available in the medical imaging domain, we use a pre-trained convolutional neural network to extract features, and then train a conventional classifier to use these features for image classification. This image analysis method was performed on FFDM images, DBT key slice images, and synthesized 2D images derived from DBT projection data. Classification performance was compared among the three different image types. The results of this study found that for mass and architectural distortion lesions DBT key slices performed better than FFDM images in the task of characterizing lesions as malignant or benign.

Having established the value of deep learning methods in medical image analysis, Chapter 5 applies deep learning methods to evaluation of temporal sequences of FFDMs

acquired over the course of years as part of consistent breast cancer screening. Through the use of LSTM networks, patterns in image characteristics over time were investigated for use in the task of classifying future lesions as malignant or benign using antecedent images. Furthermore, this study compared the use of radiomic features and features extracted from convolutional neural networks in LSTM networks. This study concluded that evaluation of temporal sequences resulted in significantly better performance than use of a single time point, however the study failed to demonstrate statistical significance between LSTM networks using radiomic or deep learning features. Therefore, these results suggest that incorporation of temporal information through a deep learning network significantly improves upon conventional approaches of single time point analysis through a simple classifier, such as SVM, in the task of classifying future malignant lesions. However, this study concluded that radiomic features and deep learning-extracted features are both appropriate for use in training an LSTM network.

Having completed these studies, several limitations were present that may be addressed in future studies. These limitations and suggested future work are summarized here.
In evaluating the robustness of radiomic features, images collected on two different vendors (GE and Hologic) were compared. This leads to further investigations of the specific factors impacting robustness. Specifically, robustness can be evaluated on imaging modalities other than mammography. These may include MRI, CT or others. Additionally, further investigation into potential harmonization techniques designed to handle the identified differences in features across imaging conditions could be developed.

The method of incorporating robustness metrics into classification of images was designed such that it required images of each patient under each of the two imaging conditions, which in this study were mammography unit vendor. Extension of this method to allow for more

flexible inputs could be investigated. Specifically, the feature selection method could be extended to incorporate image data under more than two imaging conditions. Additionally, the feature selection method could be extended to incorporate image data in cases where each patient is not necessarily imaged under the two (or more) imaging conditions under investigation.

Investigations into the application of deep learning feature extraction on DBT images yielded promising results, but these observations should be further investigated to expand upon our understanding, as the study involved a small dataset, incorporated only a single slice of the DBT, did not incorporate radiomic features, and used a simple SVM for classification. Specifically, as the work presented in this dissertation used only a single key slice of the DBT, future studies could incorporate volumetric information of the lesion by merging data from multiple slices of the DBT image set. Additionally, the methods in this study performed feature extraction followed by classification using SVM. Future studies could compare this approach to transfer learning through fine-tuning or by training a network from scratch.

In the investigation of temporal analysis of sequential mammograms, a small multi-institutional dataset with varying intra-exam time intervals was used, leaving room for future investigations. Specifically, exams in this study were not acquired at consistently spaced intervals of time. Further investigation into methods for handling variability in the time between exams, such as time-modulated LSTM, would be of interest. Additionally, this study involved predicting likelihood of malignancy in a cohort of patients who all ultimately underwent biopsy. Future studies could extend this investigation by comparing sequences of mammograms for women who ultimately developed a malignant lesion to those who did not develop any mammographic abnormality.

Computerized analysis of medical images has the potential to aid in breast cancer risk assessment. Its utility has been demonstrated on various modalities, including FFDM and DBT. By quantitatively evaluating parenchymal patterns and abnormality characteristics, radiomics and deep learning methods have the potential to provide information to clinicians that may assist in patient management decisions. The work presented in this dissertation provides evidence of the potential of radiomics and deep learning to improve breast cancer risk assessment.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68(1):7–30.

2. Tabár L, Gad A, Holmberg LH, Ljungquist U, Fagerberg CJG, Baldetorp L, et al. REDUCTION IN MORTALITY FROM BREAST CANCER AFTER MASS SCREENING WITH MAMMOGRAPHY: Randomised Trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. The Lancet 1985;325(8433):829–32.

3. Institute of Medicine (US) and National Research Council (US) Committee on the Early Detection of Breast Cancer. Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer: A Non-Technical Summary [Internet]. Washington (DC): National Academies Press (US); 2001 [cited 2018 Sep 10]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK223391/

4. Hakama M, Pukkala E, Heikkilä M, Kallio M. Effectiveness of the public health policy for breast cancer screening in Finland: population based cohort study. BMJ 1997;314(7084):864–7.

5. Mettlin CJ, Heath CW, Chu KC, Feig SA, Henderson CI, Hoover RN, et al. Biologic variations, incidence by age, and risk assessment of breast cancer screening outcomes. Cancer 1992;69(7 Suppl):1999–2000.

6. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to Fourteen-Year Effect of Screening on Breast Cancer Mortality. JNCI J Natl Cancer Inst 1982;69(2):349–55.

7. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. - PubMed - NCBI [Internet]. [cited 2019 Jan 31];Available from: https://www.ncbi.nlm.nih.gov/pubmed/3142562

8. Tabar L, Fagerberg G, Duffy SW, Day NE. The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. J Epidemiol Community Health 1989;43(2):107–14.

9. The Stockholm breast cancer screening trial--5-year results and stage at discovery. - PubMed - NCBI [Internet]. [cited 2019 Jan 31];Available from: https://www.ncbi.nlm.nih.gov/pubmed/2706329

10. Bjurstam N, Björneld L, Duffy SW, Smith TC, Cahlin E, Erikson O, et al. The Gothenburg Breast Cancer Screening Trial: preliminary results on breast cancer mortality for women aged 39-49. J Natl Cancer Inst Monogr 1997;(22):53–5.

11. Roberts MM, Alexander FE, Anderson TJ, Forrest AP, Hepburn W, Huggins A, et al. The Edinburgh randomised trial of screening for breast cancer: description of method. Br J Cancer 1984;50(1):1–6.

12. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. CMAJ Can Med Assoc J 1992;147(10):1459–76.

13. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. CMAJ Can Med Assoc J 1992;147(10):1477–88.

14. Kerlikowske K. Efficacy of screening mammography among women aged 40 to 49 years and 50 to 69 years: comparison of relative and absolute benefit. J Natl Cancer Inst Monogr 1997;(22):79–86.

15. Kerlikowske K, Grady D, Ernster V. Benefit of mammography screening in women ages 40-49 years: current evidence from randomized controlled trials. Cancer 1995;76(9):1679–81.

16. Feig S. Radiation risk from mammography: is it clinically significant? Am J Roentgenol 1984;143(3):469–75.

17. Niklason LT, Christian BT, Niklason LE, Kopans DB, Castleberry DE, Opsahl-Ong BH, et al. Digital tomosynthesis in breast imaging. Radiology 1997;205(2):399–406.

18. Friedewald SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA 2014;311(24):2499–507.

19. Rose SL, Tidwell AL, Ice MF, Nordmann AS, Sexton R, Song R. A reader study comparing prospective tomosynthesis interpretations with retrospective readings of the corresponding FFDM examinations. Acad Radiol 2014;21(9):1204–10.

20. Kopans DB. Digital Breast Tomosynthesis From Concept to Clinical Care. Am J Roentgenol 2014;202(2):299–308.

21. Reading Behavior for Screening Digital Breast Tomosynthesis (DBT) Compared to Conventional 2D Mammography (CM) [Internet]. [cited 2018 Sep 10];Available from: http://archive.rsna.org/2006/4434057.html

22. Wellings E, Vassiliades L, Abdalla R. Breast Cancer Screening for High-Risk Patients of Different Ages and Risk - Which Modality Is Most Effective? Cureus [Internet] [cited 2018 Sep 10];8(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5268380/

23. Lin D, Moy L, Axelrod D, Smith J. Utilization of magnetic resonance imaging in breast cancer screening. Curr Oncol 2015;22(5):e332–5.

24. Kuhl CK, Bieling H, Strobel K, Leutner C, Schild HH, Schrading S. Breast MRI screening of women at average risk of breast cancer: An observational cohort study. J Clin Oncol 2015;33(28_suppl):1–1.

25. Maher K (Kieran). Basic Physics of Digital Radiography. http://en.wikibooks.org/wiki/Basic_Physics_of_Digital_Radiography [Internet] 2014 [cited 2018 Sep 10];Available from: http://doer.col.org/handle/123456789/4196

26. Pisano ED, Cole EB, Hemminger BM, Yaffe MJ, Aylward SR, Maidment ADA, et al. Image Processing Algorithms for Digital Mammography: A Pictorial Essay. RadioGraphics 2000;20(5):1479–91.

27. Warren LM, Given-Wilson RM, Wallis MG, Cooke J, Halling-Brown MD, Mackenzie A, et al. The Effect of Image Processing on the Detection of Cancers in Digital Mammography. Am J Roentgenol 2014;203(2):387–93.

28. Hooley RJ, Durand MA, Philpotts LE. Advances in Digital Breast Tomosynthesis. Am J Roentgenol 2016;208(2):256–66.

29. Gur D, Abrams GS, Chough DM, Ganott MA, Hakim CM, Perrin RL, et al. Digital Breast Tomosynthesis: Observer Performance Study. Am J Roentgenol 2009;193(2):586–91.

30. Gennaro G, Toledano A, Maggio C di, Baldan E, Bezzon E, Grassa ML, et al. Digital breast tomosynthesis versus digital mammography: a clinical performance study. Eur Radiol 2010;20(7):1545–53.

31. Wallis MG, Moa E, Zanca F, Leifland K, Danielsson M. Two-View and Single-View Tomosynthesis versus Full-Field Digital Mammography: High-Resolution X-Ray Imaging Observer Study. Radiology 2012;262(3):788–96.

32. Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. Radiology 2013;269(3):694–700.

33. Andersson I, Ikeda DM, Zackrisson S, Ruschin M, Svahn T, Timberg P, et al. Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. Eur Radiol 2008;18(12):2817–25.

34. Smith A. Synthesized 2D Mammographic Imaging. :13.

35. Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer: a prospective multicentre cohort study (MARIBS). The Lancet 2005;365(9473):1769–78.

36. Bruno MA, Walker EA, Abujudeh HH. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. RadioGraphics 2015;35(6):1668–76.

37. Freer TW, Ulissey MJ. Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center. Radiology 2001;220(3):781–6.

38. Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004;96(3):185–90.

39. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. Radiology 2005;236(2):451–7.

40. Cupples TE, Cunningham JE, Reynolds JC. Impact of computer-aided detection in a regional screening mammography program. AJR Am J Roentgenol 2005;185(4):944–50.

41. Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection--prospective evaluation. Radiology 2006;239(2):375–83.

42. Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. AJR Am J Roentgenol 2006;187(1):20–8.

43. Destounis SV, DiNitto P, Logan-Young W, Bonaccio E, Zuley ML, Willison KM. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. Radiology 2004;232(2):578–84.

44. Butler SA, Gabbay RJ, Kass DA, Siedler DE, O'shaughnessy KF, Castellino RA. Computer-aided detection in diagnostic mammography: detection of clinically unsuspected cancers. AJR Am J Roentgenol 2004;183(5):1511–5.

45. Warren Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215(2):554–62.

46. Erickson BJ, Bartholmai B. Computer-aided detection and diagnosis at the start of the third millennium. J Digit Imaging 2002;15(2):59–68.

47. Summers RM. Road maps for advancement of radiologic computer-aided detection in the 21st century. Radiology 2003;229(1):11–3.

48. Abe H, Macmahon H, Shiraishi J, Li Q, Engelmann R, Doi K. Computer-aided diagnosis in chest radiology. Semin Ultrasound CT MR 2004;25(5):432–7.

49. Computer-aided diagnosis in thoracic CT. - PubMed - NCBI [Internet]. [cited 2018 Sep 17];Available from: https://www.ncbi.nlm.nih.gov/pubmed/16274004

50. Doi K, Giger ML, MacMahon H, Hoffmann KR, Nishikawa RM, Schmidt RA, et al. Computer-aided diagnosis: development of automated schemes for quantitative analysis of radiographic images. Semin Ultrasound CT MR 1992;13(2):140–52.

51. Doi null, Giger null, Nishikawa null, Schmidt null. Computer-Aided Diagnosis of Breast Cancer on Mammograms. Breast Cancer Tokyo Jpn 1997;4(4):228–33.

52. Health C for D and R. Medical Devices - Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions - Guidance for Industry and Food and Drug Administration Staff [Internet]. [cited 2018 Sep 19];Available from: https://www.fda.gov/MedicalDevices/ucm187249.htm

53. Costaridou L. Medical Image Analysis Methods. CRC Press; 2005.

54. Giger ML. Computer-aided Diagnosis in Diagnostic Mammography and Multimodality Breast Imaging. :14.

55. Vachon CM, Pankratz VS, Scott CG, Haeberle L, Ziv E, Jensen MR, et al. The Contributions of Breast Density and Common Genetic Variation to Breast Cancer Risk. JNCI J Natl Cancer Inst [Internet] 2015 [cited 2018 Jul 9];107(5). Available from: https://academic.oup.com/jnci/article/107/5/dju397/890191

56. Saftlas AF, Hoover RN, Brinton LA, Szklo M, Olson DR, Salane M, et al. Mammographic densities and risk of breast cancer. Cancer 1991;67(11):2833–8.

57. Oza AM, Boyd NF. Mammographic parenchymal patterns: a marker of breast cancer risk. Epidemiol Rev 1993;15(1):196–208.

58. Mendel KR, Li H, Giger ML. Quantitative breast MRI radiomics for cancer risk assessment and the monitoring of high-risk populations [Internet]. International Society for Optics and Photonics; 2016 [cited 2017 Oct 25]. page 97851W.Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9785/97851W/Quantitative-breast-MRI-radiomics-for-cancer-risk-assessment-and-the/10.1117/12.2217775.short

59. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol 2006;15(6):1159–69.

60. Li H, Giger ML, Olopade OI, Chinander MR. Power Spectral Analysis of Mammographic Parenchymal Patterns for Breast Cancer Risk Assessment. J Digit Imaging 2008;21(2):145–52.

61. Li H, Giger ML, Huo Z, Olopade OI, Lan L, Weber BL, et al. Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location. Med Phys 2004;31(3):549–55.

62. Kerlikowske K, Ichikawa L, Miglioretti DL, Buist DSM, Vacek PM, Smith-Bindman R, et al. Longitudinal Measurement of Clinical Mammographic Breast Density to Improve Estimation of Breast Cancer Risk. JNCI J Natl Cancer Inst 2007;99(5):386–95.

63. Giger ML, Karssemeijer N, Schnabel JA. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu Rev Biomed Eng 2013;15:327–57.

64. Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst 1995;87(9):670–5.

65. Li H, Giger ML, Olopade OI, Lan L. Fractal Analysis of Mammographic Parenchymal Patterns in Breast Cancer Risk Assessment. Acad Radiol 2007;14(5):513–21.

66. Li H, Giger ML, Lan L, Janardanan J, Sennett CA. Comparative analysis of image-based phenotypes of mammographic density and parenchymal patterns in distinguishing between BRCA1/2 cases, unilateral cancer cases, and controls. J Med Imaging Bellingham Wash 2014;1(3):031009.

67. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic Density and the Risk and Detection of Breast Cancer. N Engl J Med 2007;356(3):227–36.

68. Wang J, Kato F, Oyama-Manabe N, Li R, Cui Y, Tha KK, et al. Identifying Triple-Negative Breast Cancer Using Background Parenchymal Enhancement Heterogeneity on Dynamic Contrast-Enhanced MRI: A Pilot Radiomics Study. PLOS ONE 2015;10(11):e0143308.

69. Keller BM, Conant EF, Oh H, Kontos D. Breast Cancer Risk Prediction via Area and Volumetric Estimates of Breast Density [Internet]. In: Breast Imaging. Springer, Berlin, Heidelberg; 2012 [cited 2018 Jul 9]. page 236–43.Available from: https://link.springer.com/chapter/10.1007/978-3-642-31271-7_31

70. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern 1973;SMC-3(6):610–21.

71. Saslow D, Boetes C, Burke W, Harms S, Leach MO, Lehman CD, et al. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. CA Cancer J Clin 2007;57(2):75–89.

72. Kelsey JL, Gammon MD, John EM. Reproductive factors and breast cancer. Epidemiol Rev 1993;15(1):36–47.

73. ABC of breast diseases: Breast cancer—epidemiology, risk factors, and genetics [Internet]. [cited 2018 Dec 19];Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1118507/

74. King M-C, Marks JH, Mandell JB. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. Science 2003;302(5645):643–6.

75. A Meta-analysis of Alcohol Consumption in Relation to Risk of Breast Cancer | JAMA | JAMA Network [Internet]. [cited 2018 Dec 19];Available from: https://jamanetwork.com/journals/jama/article-abstract/373223

76. Burnside ES, Sickles EA, Bassett LW, Rubin DL, Lee CH, Ikeda DM, et al. The ACR BI-RADS® Experience: Learning From History. J Am Coll Radiol JACR 2009;6(12):851–60.

77. Wolfe J. Breast patterns as an index of risk for developing breast cancer. Am J Roentgenol 1976;126(6):1130–7.

78. Mammographic signs as risk factors for breast cancer | British Journal of Cancer [Internet]. [cited 2018 Dec 27];Available from: https://www.nature.com/articles/bjc198232

79. Carlile T, Kopecky KJ, Thompson DJ, Whitehead JR, Gilbert FI, Present AJ, et al. Breast cancer prediction and the Wolfe classification of mammograms. JAMA 1985;254(8):1050–3.

80. Whitehead J, Carlile T, Kopecky KJ, Thompson DJ, Gilbert FI, Present AJ, et al. Wolfe mammographic parenchymal patterns. A study of the masking hypothesis of Egan and Mosteller. Cancer 1985;56(6):1280–6.

81. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2015;278(2):563–77.

82. Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev 2016;1(2):207–26.

83. Li H, Giger ML, Olopade OI, Margolis A, Lan L, Chinander MR. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. Acad Radiol 2005;12(7):863–73.

84. Chinander MR, Giger ML, Martell JM, Jiang C, Favus MJ. Computerized radiographic texture measures for characterizing bone strength: A simulated clinical setup using femoral neck specimens. Med Phys 1999;26(11):2295–300.

85. Huo Z, Giger ML, Olopade OI, Wolverton DE, Weber BL, Metz CE, et al. Computerized Analysis of Digitized Mammograms of BRCA1 and BRCA2 Gene Mutation Carriers. Radiology 2002;225(2):519–26.

86. Mandelbrot BB. The Fractal Geometry of Nature. 1997; 1983.

87. Sarkar N, Chaudhuri BB. An efficient differential box-counting approach to compute fractal dimension of image. IEEE Trans Syst Man Cybern 1994;24(1):115–20.

88. Burgess AE. Mammographic structure: data preparation and spatial statistics analysis [Internet]. In: Medical Imaging 1999: Image Processing. International Society for Optics and Photonics; 1999 [cited 2018 Nov 12]. page 642–54.Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/3661/0000/Mammographic-structure-data-preparation-and-spatial-statistics-analysis/10.1117/12.348620.short

89. Heine JJ, Velthuizen RP. Spectral analysis of full field digital mammography data. Med Phys 2002;29(5):647–61.

90. Sonka M, Hlavac V, Boyle R. Image Processing, Analysis, and Machine Vision. Cengage Learning; 2014.

91. Chen W, Giger ML, Li H, Bick U, Newstead GM. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. Magn Reson Med 2007;58(3):562–71.

92. Bhooshan N, Giger ML, Jansen SA, Li H, Lan L, Newstead GM. Cancerous Breast Lesions on Dynamic Contrast-enhanced MR Images: Computerized Characterization for Image-based Prognostic Markers. Radiology 2010;254(3):680–90.

93. Mazurowski MA, Zhang J, Grimm LJ, Yoon SC, Silber JI. Radiogenomic analysis of breast cancer: luminal B molecular subtype is associated with enhancement dynamics at MR imaging. Radiology 2014;273(2):365–72.

94. Agner SC, Rosen MA, Englander S, Tomaszewski JE, Feldman MD, Zhang P, et al. Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: a feasibility study. Radiology 2014;272(1):91–9.

95. Burnside ES, Drukker K, Li H, Bonaccio E, Zuley M, Ganott M, et al. Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage. Cancer 2016;122(5):748–57.

96. Yaffe MJ, Bloomquist AK, Mawdsley GE, Pisano ED, Hendrick RE, Fajardo LL, et al. Quality control for digital mammography. Med Phys 2006;33(3):737–52.

97. Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the mammography quality standards act (MQSA). Am J Roentgenol 1995;165(1):19–25.

98. Rangayyan RM, Ayres FJ, Leo Desautels JE. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. J Frankl Inst 2007;344(3):312–48.

99. Ganesan K, Acharya UR, Chua CK, Min LC, Abraham KT, Ng K. Computer-Aided Breast Cancer Detection Using Mammograms: A Review. IEEE Rev Biomed Eng 2013;6:77–98.

100. Brisson J, Diorio C, Mâsse B. Wolfe's Parenchymal Pattern and Percentage of the Breast with Mammographic Densities: Redundant or Complementary Classifications? Cancer Epidemiol Prev Biomark 2003;12(8):728–32.

101. Li H, Giger ML, Sun C, Ponsukcharoen U, Huo D, Lan L, et al. Pilot study demonstrating potential association between breast cancer image-based risk phenotypes and genomic biomarkers. Med Phys 2014;41(3):031917.

102. Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment | Breast Cancer Research | Full Text [Internet]. [cited 2018 Dec 31];Available from: https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-016-0755-8

103. Taylor P, Hajnal S, Dilhuydy M-H, Barreau B. Measuring image texture to separate "difficult" from "easy" mammograms. Br J Radiol 1994;67(797):456–63.

104. Huo Z, Giger ML, Wolverton DE, Zhong W, Cumming S, Olopade OI. Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection. Med Phys 2000;27(1):4–12.

105. Efron B. Better Bootstrap Confidence Intervals. J Am Stat Assoc 1987;82(397):171–85.

106. Giger ML. Computerized analysis of images in the detection and diagnosis of breast cancer. Semin Ultrasound CT MR 2004;25(5):411–8.

107. Ghosh K, Hartmann LC, Reynolds C, Visscher DW, Brandt KR, Vierkant RA, et al. Association Between Mammographic Density and Age-Related Lobular Involution of the Breast. J Clin Oncol 2010;28(13):2207–12.

108. National Center for Health Statistics (US). Health, United States, 2016: With Chartbook on Long-term Trends in Health [Internet]. Hyattsville (MD): National Center for Health Statistics (US); 2017 [cited 2018 May 29]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK453378/

109. Mendel KR, Li H, Lan L, Cahill CM, Rael V, Abe H, et al. Quantitative texture analysis: robustness of radiomics across two digital mammography manufacturers' systems. J Med Imaging Bellingham Wash 2018;5(1):011002.

110. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50(11):757–65.

111. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS ONE [Internet] 2017 [cited 2018 May 29];12(9). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5608195/

112. Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 2017;44(3):1050–62.

113. Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Med Phys 2015;42(12):6784–97.

114. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation | SpringerLink [Internet]. [cited 2018 May 31];Available from: https://link.springer.com/article/10.1007/s11307-016-0940-2

115. Drukker K, Pesce L, Giger M. Repeatability in computer-aided diagnosis: Application to breast cancer diagnosis on sonography. Med Phys 2010;37(6Part1):2659–69.

116. Hunter LA, Krafft S, Stingo F, Choi H, Martel MK, Kry SF, et al. High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. Med Phys 2013;40(12):121916.

117. Zhao B, Tan Y, Tsai W-Y, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep 2016;6:23428.

118. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res 2015;24(1):27–67.

119. Obuchowski NA, Barnhart HX, Buckler AJ, Pennello G, Wang X-F, Kalpathy-Cramer J, et al. Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. Stat Methods Med Res 2015;24(1):107–40.

120. Kessler LG, Barnhart HX, Buckler AJ, Choudhury KR, Kondratovich MV, Toledano A, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res 2015;24(1):9–26.

121. Ryan TH. Significance tests for multiple comparison of proportions, variances, and other statistics. Psychol Bull 1960;57(4):318–28.

122. Jr JHW. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 1963;58(301):236–44.

123. Soper HE, Young AW, Cave BM, Lee A, Pearson K. ON THE DISTRIBUTION OF THE CORRELATION COEFFICIENT IN SMALL SAMPLES. APPENDIX II TO THE PAPERS OF “STUDENT” AND R. A. FISHER. A COOPERATIVE STUDY. Biometrika 1917;11(4):328–413.

124. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. Bioinforma Oxf Engl 2005;21(8):1509–15.

125. Darling DA. The Kolmogorov-Smirnov, Cramer-von Mises Tests. Ann Math Stat 1957;28(4):823–38.

126. Draper NR. Applied regression analysis. 3rd ed. New York: Wiley; 1998.

127. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med 2018;jnumed.117.199935.

128. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8(1):118–27.

129. Kendall MG. A NEW MEASURE OF RANK CORRELATION. Biometrika 1938;30(1–2):81–93.

130. Sen PK. Estimates of the Regression Coefficient Based on Kendall's Tau. J Am Stat Assoc 1968;63(324):1379–89.

131. Metz CE, Herman BA, Roe CA. Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets. Med Decis Making 1998;18(1):110–21.

132. Burgess AE, Judy PF. Signal detection in power-law noise: effect of spectrum exponents. JOSA A 2007;24(12):B52–60.

133. Human observer detection experiments with mammograms and power-law noise - Burgess - 2001 - Medical Physics - Wiley Online Library [Internet]. [cited 2018 Nov 12];Available from: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1355308

134. Soille P, Rivest J-F. On the Validity of Fractal Dimension Measurements in Image Analysis. J Vis Commun Image Represent 1996;7(3):217–29.

135. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of Radiomic Features in [11C]Choline and [18F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization. Mol Imaging Biol 2016;18(6):935–45.

136. Miles KA, Ganeshan B, Hayball MP. CT texture analysis using the filtration-histogram method: what do the measurements mean? Cancer Imaging 2013;13(3):400–6.

137. Ganeshan B, Burnand K, Young R, Chatwin C, Miles K. Dynamic contrast-enhanced texture analysis of the liver: initial assessment in colorectal cancer. Invest Radiol 2011;46(3):160–8.

138. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol J Eur Soc Ther Radiol Oncol 2015;114(3):345–50.

139. Assessment of Response to Tyrosine Kinase Inhibitors in Metastatic Renal Cell Cancer: CT Texture as a Predictive Biomarker | Radiology [Internet]. [cited 2018 Nov 5];Available from: https://pubs.rsna.org/doi/10.1148/radiol.11110264

140. Ghetti C, Borrini A, Ortenzia O, Rossi R, Ordóñez PL. Physical characteristics of GE Senographe Essential and DS digital mammography detectors. Med Phys 2008;35(2):456–63.

141. Gingold EL, Wu X, Barnes GT. Contrast and dose with Mo-Mo, Mo-Rh, and Rh-Rh target-filter combinations in mammography. Radiology 1995;195(3):639–44.

142. (6) Comparison of Acquisition Parameters and Breast Dose in Digital Mammography and Screen-Film Mammography in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial | Request PDF [Internet]. ResearchGate [cited 2018 Sep 10];Available from: https://www.researchgate.net/publication/41111702_Comparison_of_Acquisition_Parameters_and_Breast_Dose_in_Digital_Mammography_and_Screen-Film_Mammography_in_the_American_College_of_Radiology_Imaging_Network_Digital_Mammographic_Imaging_Screening_Tria

143. Reiser I, Nishikawa RM, Giger ML, Wu T, Rafferty EA, Moore R, et al. Computerized mass detection for digital breast tomosynthesis directly from the projection images. Med Phys 2006;33(2):482–91.

144. Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis - ScienceDirect [Internet]. [cited 2017 Oct 6];Available from: http://www.sciencedirect.com/science/article/pii/S1532046411000724

145. van Schie G, Wallis MG, Leifland K, Danielsson M, Karssemeijer N. Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2D mammograms. Med Phys 2013;40(4):n/a-n/a.

146. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

147. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks [Internet]. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012 [cited 2017 Oct 23]. page 1097–1105.Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

148. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 2016;35(5):1285–98.

149. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys 2017;44(10):5162–71.

150. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging 2016;3(3):034501.

151. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). 2015. page 294–7.

152. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. page 248–55.

153. A Survey on Transfer Learning - IEEE Journals & Magazine [Internet]. [cited 2017 Oct 23];Available from: http://ieeexplore.ieee.org/abstract/document/5288526/

154. Samala Ravi K., Chan Heang-Ping, Hadjiiski Lubomir, Helvie Mark A., Wei Jun, Cha Kenny. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Med Phys 2016;43(12):6654–66.

155. Samala RK, Chan H-P, Hadjiiski LM, Cha K, Helvie MA. Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis [Internet]. In: Medical Imaging 2016: Computer-Aided Diagnosis. International Society for Optics and Photonics; 2016 [cited 2018 May 10]. page 97850Y.Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9785/97850Y/Deep-learning-convolution-neural-network-for-computer-aided-detection-of/10.1117/12.2217092.short

156. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Richter C, Cha K. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. Phys Med Biol 2018;63(9):095005.

157. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs [Internet] 2014 [cited 2017 Dec 5];Available from: http://arxiv.org/abs/1409.1556

158. Zheng L, Zhao Y, Wang S, Wang J, Tian Q. Good Practice in CNN Feature Transfer. ArXiv160400133 Cs [Internet] 2016 [cited 2018 Jun 4];Available from: http://arxiv.org/abs/1604.00133

159. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.

160. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

161. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8(4):283–98.

162. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998;20(3):226–39.

163. Gajdos C, Ian Tartter P, Bleiweiss IJ, Hermann G, de Csepel J, Estabrook A, et al. Mammographic Appearance of Nonpalpable Breast Cancer Reflects Pathologic Characteristics. Ann Surg 2002;235(2):246–51.

164. Friedman M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. Ann Math Stat 1940;11(1):86–92.

165. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839–43.

166. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat 1979;6(2):65–70.

167. Chan H-P, Wei J, Sahiner B, Rafferty EA, Wu T, Roubidoux MA, et al. Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience. Radiology 2005;237(3):1075–80.

168. Michell MJ, Iqbal A, Wasan RK, Evans DR, Peacock C, Lawinski CP, et al. A comparison of the accuracy of film-screen mammography, full-field digital mammography, and digital breast tomosynthesis. Clin Radiol 2012;67(10):976–81.

169. Morra L, Sacchetto D, Durando M, Agliozzo S, Carbonaro LA, Delsanto S, et al. Breast Cancer: Computer-aided Detection with Digital Breast Tomosynthesis. Radiology 2015;277(1):56–63.

170. Park JM, Franken EA, Garg M, Fajardo LL, Niklason LT. Breast tomosynthesis: present considerations and future applications. Radiogr Rev Publ Radiol Soc N Am Inc 2007;27 Suppl 1:S231-240.

171. Peppard HR, Nicholson BE, Rochman CM, Merchant JK, Mayo RC, Harvey JA. Digital Breast Tomosynthesis in the Diagnostic Setting: Indications and Clinical Applications. RadioGraphics 2015;35(4):975–90.

172. Baker JA, Lo JY. Breast tomosynthesis: state-of-the-art and review of the literature. Acad Radiol 2011;18(10):1298–310.

173. Kim DH, Kim ST, Chang JM, Ro YM. Latent feature representation with depth directional long-term recurrent learning for breast masses in digital breast tomosynthesis. Phys Med Biol 2017;62(3):1009.

174. Smith RA, Cokkinides V, Brawley OW. Cancer screening in the United States, 2009: a review of current American Cancer Society guidelines and issues in cancer screening. CA Cancer J Clin 2009;59(1):27–41.

175. Qaseem A, Snow V, Sherif K, Aronson M, Weiss KB, Owens DK, et al. Screening Mammography for Women 40 to 49 Years of Age: A Clinical Practice Guideline from the American College of Physicians. Ann Intern Med 2007;146(7):511.

176. Oeffinger KC, Fontham ETH, Etzioni R, Herzig A, Michaelson JS, Shih Y-CT, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA 2015;314(15):1599–614.

177. Yankaskas BC, May RC, Matuszewski J, Bowling JM, Jarman MP, Schroeder BF. Effect of Observing Change from Comparison Mammograms on Performance of Screening Mammography in a Large Community-based Population. Radiology 2011;261(3):762–70.

178. Santeramo R, Withey S, Montana G. Longitudinal Detection of Radiological Abnormalities with Time-Modulated LSTM. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T,

Martel A, Maier-Hein L, et al., editors. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer International Publishing; 2018. page 326–33.

179. Shao Y, Chen Z, Ming S, Ye Q, Shu Z, Gong C, et al. Predicting the Development of Normal-Appearing White Matter With Radiomics in the Aging Brain: A Longitudinal Clinical Study. Front Aging Neurosci [Internet] 2018 [cited 2019 Apr 8];10. Available from: https://www.frontiersin.org/articles/10.3389/fnagi.2018.00393/full

180. Antropova N, Huynh B, Li H, Giger ML. Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks. J Med Imaging 2018;6(1):011002.

181. Díez Y, Oliver A, Llado X, Freixenet J, Marti J, Vilanova JC, et al. Revisiting Intensity-Based Image Registration Applied to Mammography. IEEE Trans Inf Technol Biomed 2011;15(5):716–25.

182. Mendel K, Li H, Tayob N, El-Zein R, Bedrosian I, Giger M. Temporal mammographic registration for evaluation of architecture changes in cancer risk assessment [Internet]. In: Medical Imaging 2019: Computer-Aided Diagnosis. International Society for Optics and Photonics; 2019 [cited 2019 Mar 20]. page 1095041.Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10950/1095041/Temporal-mammographic-registration-for-evaluation-of-architecture-changes-in-cancer/10.1117/12.2512792.short

183. Mendel KR, Li H, Lan L, Chan C-W, King LM, Tayob N, et al. Temporal assessment of radiomic features on clinical mammography in a high-risk population [Internet]. In: Medical Imaging 2018: Computer-Aided Diagnosis. International Society for Optics and Photonics; 2018 [cited 2019 Mar 20]. page 105753Q.Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10575/105753Q/Temporal-assessment-of-radiomic-features-on-clinical-mammography-in-a/10.1117/12.2293368.short

184. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. ArXiv151203385 Cs [Internet] 2015 [cited 2019 Jan 24];Available from: http://arxiv.org/abs/1512.03385

185. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training Recurrent Neural Networks. ArXiv12115063 Cs [Internet] 2012 [cited 2019 Jan 24];Available from: http://arxiv.org/abs/1211.5063

186. Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. Int J Uncertain Fuzziness Knowl-Based Syst 1998;6(2):107–116.

187. Salazar AJ, Romero JA, Bernal OA, Moreno AP, Velasco SC, Díaz XA. Noninferiority and Equivalence Evaluation of Clinical Performance among Computed Radiography, Film, and Digitized Film for Telemammography Services [Internet]. Int. J. Telemed. Appl.2016

[cited 2019 Mar 21];Available from: https://www.hindawi.com/journals/ijta/2016/3642960/

188. Brasher PMA, Dobson G. Understanding non-inferiority trials: an introduction. Can J Anesth Can Anesth 2014;61(5):389–92.

189. Bland JM, Altman DG. The odds ratio. BMJ 2000;320(7247):1468.

190. Peng C-YJ, Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and Reporting. J Educ Res 2002;96(1):3–14.

191. Gierach GL, Li H, Loud JT, Greene MH, Chow CK, Lan L, et al. Relationships between computer-extracted mammographic texture pattern features and BRCA1/2mutation status: a cross-sectional study. Breast Cancer Res [Internet] 2014 [cited 2019 Feb 25];16(4). Available from: http://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-014-0424-8

192. Li H, Giger ML, Huynh BQ, Antropova NO. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. J Med Imaging 2017;4(4):041304.