THE UNIVERSITY OF CHICAGO

MACHINE LEARNING OF BREAST DCE-MRI IN ASSESSING BACKGROUND PARENCHYMAL ENHANCEMENT FOR 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

LINDSAY NICOLE DOUGLAS

CHICAGO, ILLINOIS

JUNE 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful for the support that I have received from many individuals during my time at the University of Chicago. As I reflect on the past five years, I recognize that I could not have made it to where I am now without their contributions.

First and foremost, I want to thank Maryellen Giger for being an excellent advisor that guided me throughout the projects we completed for this work. She has also been a wonderful mentor; her compassion, wisdom, generosity, and patience have encouraged me both personally and professionally.

I would like to express my sincerest gratitude to Emily Marshall, Ingrid Reiser, and Zheng Feng Lu for being incredible role models for me since my first exposure (no pun intended) to imaging medical physics. To Sam Armato, Hania Al-Hallaq, and Patrick La Riviere, thank you for support in- and outside of the classroom. I appreciate the valuable contributions and feedback that members of my thesis committee, including Deepa Sheth and Hiroyuki Abe, provided me with over the course of my dissertation work.

I am especially grateful to Natalie Baughan, Hadley DeBrosse, Mena Shenouda, Mira Liu, and Linnea Kremer for being wonderful people that are a pleasure to have not just as peers, but as friends. Thank you to many of the previous students who have steered me in the right direction, whether that was in classes, for research, or to the best coffee spots on campus: Inna Gertsenshteyn, Jordan Fuhrman, Isabelle Hu, Brittany Broder, and Jennie Crosby. There are many others in the Graduate Program in Medical Physics that I have to thank. To Ben Preusser, Jon George, and Andrew McVea, I am glad we could work through all of the courses together in my first year. To current and past students that I have been able to get to know in the past few years: Joseph Cozzi,

Julian Bertini, Gia Jadick, Geneva Schlafly, Lucas Berens, and others, I wish I could have interacted with you even more, and I can't wait to see where you go in the future.

As a member of the Giger Lab, I have also had the pleasure of working with many other individuals who have provided me with excellent suggestions and insight throughout the years. I would therefore like to thank the following people: Hui Li, Li Lan, Heather Whitney, Madeleine Durkee, Sasha Edwards, John Papaioannou, Karen Drukker, and Chun Wai Chan. I am also grateful for the experiences I was able to have mentoring the following undergraduate and high school students during summer research experiences: Roma Bhattacharjee, Trisha Mondal, Esther Lee, Maya Ballard, and Catherine Collins.

There are a few notable individuals that I have to thank for inspiring me to pursue medical physics before I came to study at UChicago. To Nathan Dorsch, thank you for being such an amazing teacher and for sparking my interest in physics; you changed my life. To Scott Holmstrom, Jerry McCoy, and George Miller, thank you for always encouraging me to think like a physicist.

Last, but certainly not least, I want to thank my family and friends for their never-ending support. You've always been there when I needed you, and you each have a unique ability to pick me up when I am down. Even though we may have felt like we were going insane at times during the pandemic, we got through it together. I never would have made it this far without you. Thank you, Mom, Dad, Sara, Jacob, Auntie, and Candace for everything (and thanks to Teddy, Alice, and Toews, for loving me as long as I keep petting you!).

ABSTRACT

To enhance breast cancer screening practices, artificial intelligence (AI) systems have been developed to aid radiologists in a variety of tasks. Machine learning (ML) techniques for computer-aided diagnosis are based on human-engineered or deep learning methods, and they depend on accurate segmentation for useful feature extraction. As the use of dynamic contrast-enhanced (DCE) magnetic resonance imaging (MRI) has increased in breast imaging, particularly for high-risk screening, the potential for AI to provide significant clinical benefit has grown. There is need for a deeper understanding of how breast MRI can be used for diagnosis and risk assessment in order to develop robust, generalizable AI systems for quantifying clinically valuable breast characteristics.

This dissertation presents novel methods for computerized assessment of background parenchymal enhancement (BPE), a known risk factor for breast cancer, from breast DCE-MRI. In Chapter 1, we introduce the background of breast cancer screening with a focus on AI applications to motivate the subsequent chapters. In Chapter 2, we investigate segmentation techniques for lesions and breast regions. In Chapter 3, we develop an ML technique for computer BPE scoring that includes electronic lesion removal. In Chapter 4, we perform an independent evaluation of the BPE scoring algorithm applied to high-risk patients. Ultimately, the results of this work have the potential to encourage future incorporation of quantitative image analysis into the clinical workflow for radiologists and therefore improve patient care.

Segmentation of lesions and breasts: Methods for segmentation of breast lesions and breasts from DCE-MRI were investigated using a dataset of patients diagnosed with cancerous or benign mass- or nonmass-enhancing lesions. Lesion segmentation performances of U-Net convolutional

neural networks were compared to the fuzzy c-means (FCM) clustering algorithm and to radiologist delineations. Breast segmentation was performed on post-contrast subtraction maximum intensity projection images. Results suggest that using a 2D U-Net on post-contrast subtraction DCE-MRIs is feasible and could be an effective alternative to FCM or 3D U-Net for lesion segmentation.

Computerized assessment of BPE: An automatic computer BPE scoring method that includes electronic lesion removal was developed using a dataset of DCE-MRIs that had radiologist BPE ratings available from prior clinical review. Qualitative, radiologist-reported BPE ratings and quantitative, computer BPE scores were evaluated for different breast regions, and the effect of varying image types and magnet strengths was investigated. A statistically significant correlation was found between the radiologist and computer BPEs. Results demonstrated promising performances of the computerized method for classifying BPE levels across various viewing projections and DCE timepoints.

BPE scoring on a high-risk dataset: The role of BPE in predicting breast cancer was explored for a dataset of high-risk screening DCE-MRIs. An independent validation of the BPE scoring algorithm reproduced findings from the initial dataset on an independent dataset. In addition, results found a statistically significant difference between the computer BPE scores of patients that developed cancer and those of non-cancer patients with low BPE. Future investigations involving enriched datasets would expand the understanding of the role that computer BPE scores can have in predicting cancer.

CHAPTER 1

INTRODUCTION

Technological advancements in imaging modalities and artificial intelligence (AI) systems have contributed to the evolution of breast cancer screening practices over the past few decades. AI has been developed to enhance the efficiency of interpretation tasks associated with breast cancer screening for decades, and its potential is greater now since newer acquisition systems yield 3D and 4D images. In particular, magnetic resonance imaging (MRI) has a growing role in the diagnosis and management of breast cancer with the potential to avoid intrusive biopsies, provide tumor staging, and monitor treatment response. Machine learning techniques for computer-aided diagnosis of breast cancer have been developed based on features extracted from dynamic contrast-enhanced (DCE) MRI, which depend on the accurate segmentation of lesions and other breast regions. Ultimately, to improve risk assessment and cancer diagnosis, machine learning may be used to yield quantitative values for clinically significant breast characteristics, such as background parenchymal enhancement.

1.1 Breast Cancer Screening

Globally, female breast cancer is the most commonly diagnosed cancer, and it is the greatest contributor to cancer death in women. [1] Since the peak of breast cancer mortality in 1989, there has been a 42% decrease in mortality in the United States. [2] In the late 1990s, the annual decline in mortality rate was more than 3%, although in recent years, the decline has slowed to 1% annually, possibly due to plateauing mammography rates and a slight increase in incidence rates. [2] Screening mammography has played an important role in reducing breast cancer-related mortality by increasing cancer detection rates at earlier stages. As a result, cancer screening can

enable less invasive and more effective treatment. [3, 4] However, limited contrast and overlapping tissue in the 2D projection images from mammograms is not ideal for those with dense breasts, contributing to overdiagnosis and overtreatment. [4–6] To address the need for more effective screening, additional imaging modalities have been reassessed for their role in augmenting screening mammography. [7–9]

Digital breast tomosynthesis (DBT) has been shown to have greater cancer detection rates compared to 2D mammography, as it can reduce false positives resulting from overlapping normal tissues. [3] In addition, the non-ionizing radiation imaging modalities, whole-breast ultrasound (US) and 3D MRI, have demonstrated sensitivity benefits over mammography especially in detecting mammographically occult disease. [6, 10] Whole-breast US has demonstrated benefits in patients with dense breasts, although it has an increased risk of false positives and limited ability for screening in the general population. [3, 6] MRI also offers the benefits of 3D resolution along with temporal information from DCE-MRI, and is beneficial for use in women with dense breasts and above-average risk. [6, 11] Features extracted from MRI, including lesion size, shape, and texture, can serve as strong indicators for use in diagnosis. [12] Additionally, MRI has been shown to detect approximately 10 per 1000 cancers that were otherwise undetected by mammography or US. [13] Therefore, DCE-MRI is being used as a supplemental screening modality in patients who have qualified as high-risk based on a series of known risk factors, such as hormonal and reproductive status, genetic mutation status (e.g., *BRCA1*, *BRCA2*, or *PALB2*), personal history of breast disease, radiation exposure, or family history of breast and ovarian cancers. [14–18]

While specific guidelines vary around the world, the World Health Organization recommends mammography screening every 2 years for average-risk women aged 50-69-years-

old. [4] The American Cancer Society recommends annual screening mammography or DBT starting at 40 years old for average-risk women and recommends annual MRI as an adjunct to screening mammography or DBT starting at 30 years old for high-risk women. [3, 10] The United States Food and Drug Administration (FDA) has established federal regulations to maintain the consistency and quality of breast cancer screening practices. For instance, under the Mammography Quality Standards Act (MQSA), FDA has developed standards for accreditation, certification, and inspection of mammography facilities. [19] Recently, in order to improve individualized breast health care, the FDA issued a final rule to update the MQSA regulations and require facilities to inform patients of their breast density, a known risk factor for breast cancer. [20] Also, the American College of Radiology (ACR) has designed the Breast Imaging Reporting & Data System (BI-RADS©) atlas to standardize radiologist reports; it includes guidelines for imaging terminology, assessment structures, and management recommendations for mammography, US, and MRI. [21]

1.2 Breast Magnetic Resonance Imaging

MRI has had an integral role in improving breast cancer diagnoses and potentially reducing biopsies, in tumor staging, and in monitoring treatment response. [12, 22, 23] Routine breast MRI is performed using a 1.5 or 3.0 Tesla (T) magnet and a dedicated multichannel breast coil. [6] Although a 3.0T magnet may be more susceptible than 1.5T to imaging artifacts, using a higher magnet strength has the benefit of improved spatial resolution, signal-to-noise ratios, and faster data collection, ultimately providing better image quality. [24] A typical protocol can involve multiple acquisitions, including T2-weighted, diffusion-weighted, and T1-weighted sequences (Figure 1.1). T2 weighting is fluid-sensitive and therefore useful for visualization of cysts and

edema; these images show detail on normal tissue structure and lesion morphology. [23] Diffusion-weighted imaging quantifies the random movement of water in tissue, so its signal is sensitive to the cell density and tissue microstructure. [23] T1 weighting is used to highlight anatomic detail and functional information using a dynamic contrast-enhancement series. [6, 23]

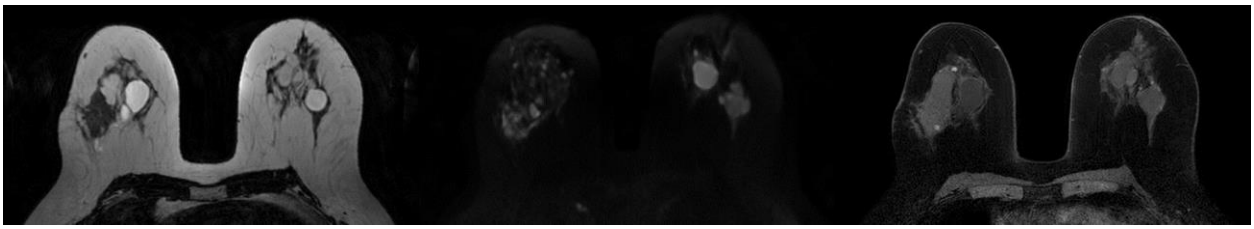

Figure 1.1: Examples of T2-weighted (left), diffusion-weighted (center), and pre-contrast T1-weighted (right) axial images acquired from a single patient.

Conventionally, DCE-MRI includes a native T1-weighted acquisition followed by the repeated sequence in 60-90 second intervals after the administration of contrast material (usually, gadolinium). [6, 23] Figure 1.2 shows an example of the bright signal produced as the contrast is taken up by the highly vascularized tumor more than the surrounding fibroglandular tissue. Usually, the pre-contrast image is subtracted from the post-contrast images to visualize the uptake of contrast above the baseline. In recent years, the use of 3.0T instead of 1.5T MRI has increased, resulting in potential changes in perceived signal intensity. The shift from 1.5T to 3.0T increases the T1 relaxation time for fat and glandular tissue more than it does for gadolinium, resulting in a greater relative difference in signal between enhancing and nonenhancing tissues, thus making enhancing tissue more conspicuous. [25, 26]

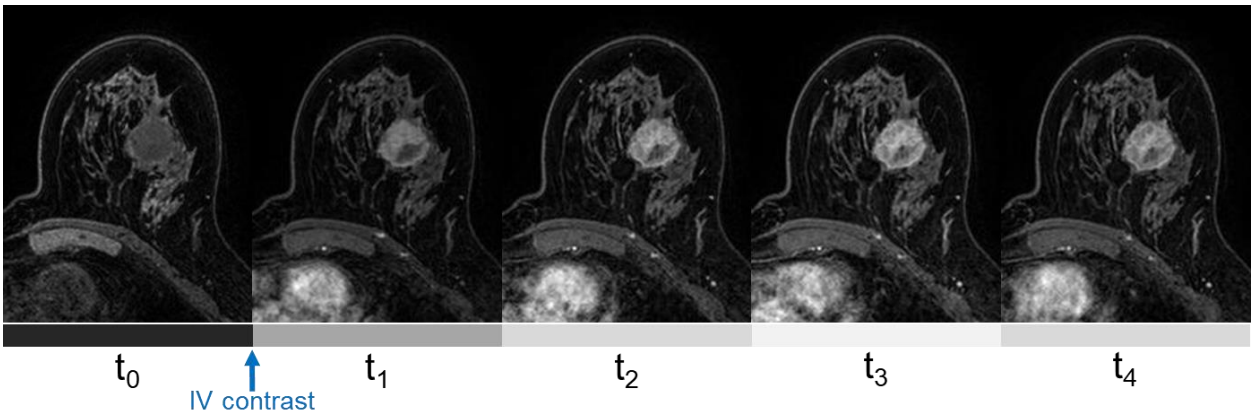

Figure 1.2: Pattern of contrast uptake in a malignant tumor over the course of a typical dynamic contrast-enhanced MRI series. Post-contrast images ($t_{1\text{-}5}$) were acquired at 65 second intervals following the pre-contrast image (t_0).

The uptake and washout patterns of the contrast agent can provide functional information about lesions and the surrounding fibroglandular tissue. [6, 22, 23] For instance, lesions with rapid initial enhancement followed by washout are more likely malignant, while a slow initial rise followed by a persistent enhancement is a typical pattern of benign lesions (Figure 1.3). [27] Normal fibroglandular tissue tends to exhibit a slow early and persistent delayed uptake of contrast. This contrast uptake and resulting imaging appearance is referred to as background parenchymal enhancement (BPE). [28] Based on the visually perceived volume and intensity of enhancement in normal fibroglandular breast tissue after contrast injection for DCE-MRI, radiologists qualitatively rate BPE as minimal, mild, moderate, or marked according to BI-RADS. [21, 29, 30]

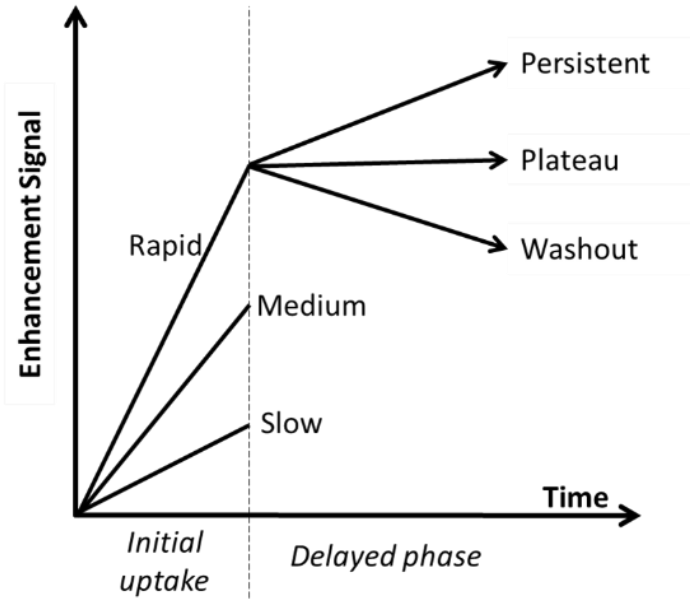

Figure 1.3: Kinetic curve assessment includes interpretation of the initial and delayed phases of contrast uptake in lesions over the course of a DCE-MRI series.

To maintain the efficiency and throughput and to increase the performance of screening MRI, abbreviated and ultrafast protocols have been developed. [9, 11, 31] The abbreviated MRI (AB-MR) sequence was introduced as a shortened alternative to the standard DCE protocol; the acquisition time is merely 3 minutes, compared to the 17 minutes necessary for a full protocol. [32, 33] AB-MR includes one pre- and one post-contrast T1-weighted acquisition, and it has been shown to have diagnostic accuracy similar to full protocols. [11, 23] Imaging at just one post-contrast timepoint, where the greatest divergence between tumor and background enhancement occurs, maximizes tumor conspicuity but loses delayed-phase contrast kinetic information. [11] In order to image the initial contrast uptake that occurs prior to the first post-contrast acquisition for the full or abbreviated protocols, ultrafast MRI has been designed to capture the early inflow of contrast into lesions at a high temporal resolution by sequentially imaging every 6-7 seconds or shorter within the first minute after contrast injection. [11, 23] Malignancies have more significant

early contrast uptake than healthy tissues, and the early wash-in kinetics have been shown to have great discriminating power between benign and malignant lesions. [11, 23, 27] Ultrafast MRI offers a variety of new diagnostic parameters that have the potential to serve as radiomic features with prognostic value, such as maximum slope and initial enhancement rate. [11]

1.3 Artificial Intelligence for Breast Cancer Screening and Diagnosis

Beyond the image quality, the benefit of medical imaging exams also relies on the quality of the interpretation. While there are many imaging modalities used to provide radiologists with an abundance of data for each patient, human interpretation is inherently limited by structural noise, incomplete visual search patterns, suboptimal image quality, or fatigue. [9, 34–36] To effectively interpret DBT, US, or MRI data, additional expertise may be required for detection, diagnoses, and patient management. Artificial intelligence methods, including detection, diagnosis, or segmentation tasks, have been developed to support radiologists in their interpretation decision-making process. In the future, AI has the potential to improve the efficiency of image interpretation, and in some implementations, AI may outperform humans. Additionally, AI algorithms may be able to integrate a variety of data inputs, i.e., originating from patient images, genetic information, demographics, or medical history. [8] These systems would potentially make a unique impact on patients at high risk for breast cancer by enabling a more personalized assessment of risk.

Machine learning is a subset of AI that includes conventional and deep learning methods; both methods involve programs designed to identify patterns and make predictions without direct human intervention. Conventional methods have used human-engineered radiomic features to characterize and, further, to classify breast lesions. [37, 38] Most computer-aided diagnosis

systems for breast screening fall into the categories of human-engineered or deep-learning-based AI, utilizing radiomic and/or deep network extracted features to perform a task. Computer-aided diagnosis systems can be further divided into categories based upon the specific clinical task, including computer-aided detection (CADe), computer-aided diagnosis (CADx), triage (CADt) or rule-out.

1.3.1 History of AI in Breast Cancer Screening

While the first applications of AI to breast screening focused on mammography, many of the techniques were further customized and translated to other screening modalities. As DBT emerged in the field as a promising three-dimensional alternative to standard mammography, AI techniques were rapidly extended to DBT imaging. [39–41] Additionally, CADe and CADx methods for both 2D and 3D breast US were developed in the 2000s. [42, 43] As MRI became an adjunct imaging modality for screening women with dense breasts, AI systems were developed for DCE-MRI. In 1998, a method for automated extraction of human-engineered radiomic features based on size, shape, and kinetics of radiologist-delineated masses was successful in distinguishing benign from malignant lesions. [44] Soon after, techniques for three-dimensional computerized lesion segmentation from DCE-MRI were introduced. [45] In the early 2000s, texture analysis, morphology, and kinetics were incorporated in the development of automatic methods for lesion classification. [46–50] Some of the early commercial software systems offered interactive tools for the assessment of DCE-MRI that could be integrated in the clinical workflow, providing decision support while reducing evaluation time and observer variability. [48, 51] The first commercial CADx system was QuantX (Quantitative Insights, now Qlarity Imaging, Chicago, IL),

approved by the FDA in 2017 as a second reader to be used after a radiologist's initial review of a suspect lesion on DCE-MRI. [12]

1.3.2 Current State of AI in Breast Cancer Screening

Within the last 10 years, AI has been a dominant force in breast cancer screening research. AI is being implemented for a range of uses: as a second reader, as a concurrent reader, as a primary reader in rule out, and as a triage system for the prioritization of cases for reader order. [52–54] Note that at present, AI systems are serving primarily as an aid to the reader and are not intended to replace the breast radiologist, although future efforts may be directed towards developing methods to function autonomously. Figure 1.4 demonstrates the process for developing AI to serve in the different roles during the screening workflow.

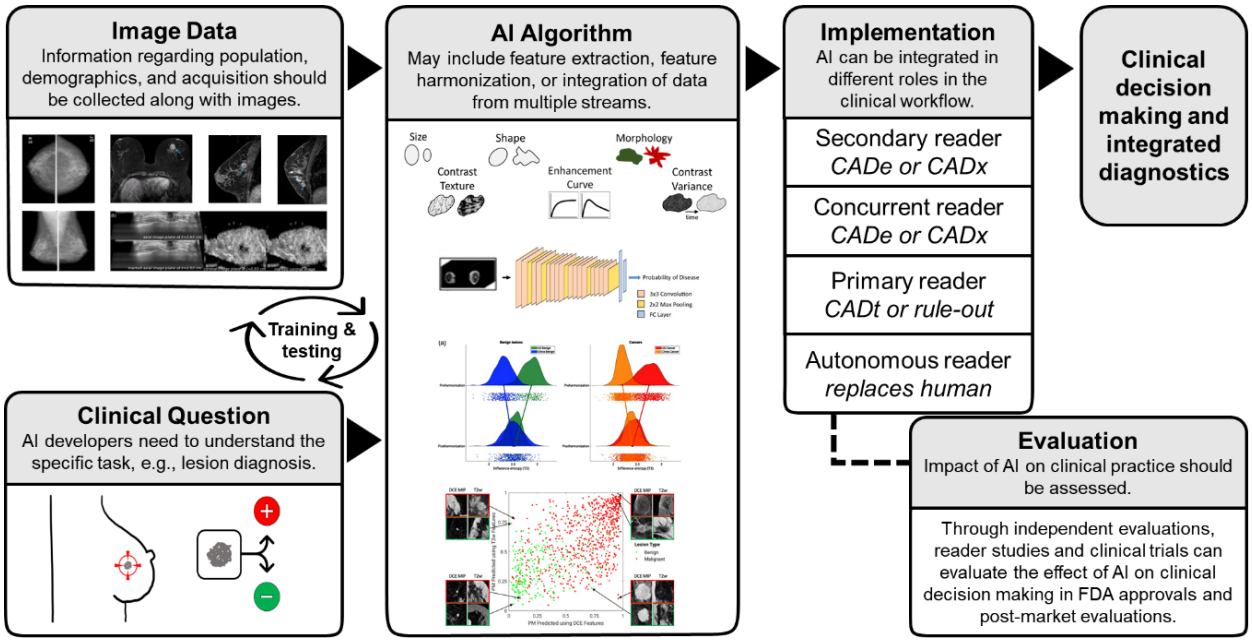

Figure 1.4: Schematic illustrating the components in developing an artificial intelligence (AI) algorithm for breast cancer screening. Specific algorithms will be trained and tested for unique tasks that are based on the dataset and clinical questions. The systems can serve different roles for the end user in computer-aided detection (CADe), computer-aided diagnosis (CADx), triage (CADt), or rule-out tasks. The impact of AI on the efficacy and efficiency of the clinical interpretation and workflow may be quantified with reader studies before approval by the Food and Drug Administration (FDA). (Source [55])

Methods for the development of human-engineered techniques and deep learning algorithms for screening modalities in the past decade have shown a variety of promising advancements. [38, 54, 56, 57] In mammography and DBT, human-engineered techniques have been expanded to include a wider selection of complex image features, and deep learning algorithms for detection and classification have been developed for faster implementation. [57, 58] Studies have shown that AI-assisted methods can maintain the accuracy of diagnosis while increasing the efficiency of interpretation for automated 3D breast ultrasound examinations. [59, 60] A number of AI techniques have also been developed to automatically detect and classify lesions based on the dynamic and morphological information contained in several MRI sequences. [61–63] Both human-engineered and deep learning AI techniques have each been shown to perform well in breast lesion classification tasks; a number of publications have cited significant improvements in algorithm performances when merging human-engineered radiomic and deep learning algorithms into the machine learning decision across mammography, US, and DCE-MRI, even with modestly sized datasets. [7, 64, 65]

Additionally, modifying the image format input to deep networks in order to more efficiently incorporate volumetric and temporal information, such as post-contrast subtraction maximum intensity projection (MIP) images, has been shown to further improve performance in breast tumor classification tasks. [66] Further performance improvements have been reported by effectively fusing image data from multiparametric breast MRIs (DCE-MRI, T2w, diffusion-weighted imaging), through either human-engineered or deep learning methods, at the pixel level, the feature level, or the classifier output level. [67, 68] Basically, effective development of an AI algorithm requires knowledge of the image acquisition process and the various formats of image

presentation/reconstruction as well as the architecture of the radiomics/deep network itself. As of April 2023, 22 AI algorithms for breast imaging have been cleared by the FDA. [69] Of the cleared algorithms, 10 are for the purpose of breast density assessment, while 12 are intended to analyze breast lesion characteristics. While the majority of cleared algorithms are for use on mammography or DBT imaging, two are based on breast US and one on breast screening MRI.

1.3.3 Future of AI in Breast Cancer Screening

The next generation of AI in breast cancer screening is expected to further increase the efficiency and efficacy of medical image interpretation across all modalities. One aspect of this goal is to extend AI from a second or concurrent reader (CADe, CADx) to an autonomous or partially autonomous reader. [57] Recent studies have shown software that approach or exceed the performance of radiologists. [70, 71] For example, McKinney et al. showed an AI detection system for screening mammography capable of outperforming six radiologists, with an average absolute margin in the area under the receiver operating characteristic curve (AUC-ROC) of 11.5% between the AI system and the radiologists. [70] However, limitations and challenges exist between the current state of AI and clinically applicable autonomous reading. For instance, many reports on the diagnostic accuracy of AI exist, but there is a lack of evidence on the perception and implementation of AI in actual clinical practice. [72]

While the majority of AI research has focused on single interpretation tasks such as detection or diagnosis, a large area in which AI may have an impact on future breast cancer screening workflow is through the application of AI to streamline ancillary tasks. For example, AI may preprocess images or assist in the generation of standardized reporting documents. [37] Image pre-processing may include image reconstruction, artifact correction, noise reduction, and user-

preferred arrangement (hanging) of images. Additionally, computer-aided triage and CAD rule-out software could streamline clinical workflow and reduce radiologist workloads. It is important to note that according to the FDA Code of Federal Regulations (CFR) Title 21, computer-aided triage refers to software used to prioritize images and not to remove any from a given imaging queue. [73] In contrast, CAD rule-out would potentially remove a subset of cases from a screening queue if deemed to be below a pre-determined risk threshold at which human reading in addition to computer reading is not necessary. The rule-out software would act as a truly autonomous reader for the subset of cases removed from a screening queue, and although the software is currently ineligible under CFR, simulation studies have demonstrated the potential for rule-out to improve sensitivity and efficiency without an impact on sensitivity. [74–76]

1.3.4 Challenges for AI in Breast Cancer Screening and Diagnosis

As AI algorithms have been developed with increasing complexity, various challenges have been considered including explainability and interpretability, robustness and repeatability, generalizability, and ethical implementations. Potential users of AI systems may serve a variety of roles, including clinicians, researchers, or regulators, and therefore have unique interests in the system outputs. [77, 78] The output of AI systems may not be trusted unless they are well understood, thus, algorithms need to be explainable, interpretable, and user-friendly. [35, 79] The robustness and repeatability of an algorithm should also be considered during its development. For conventional methods that involve extraction of human-engineered radiomic features, issues regarding the robustness of a model to perform on a different system have often been reported. [80, 81] Additionally, an investigation of case-based repeatability within one system found that the confidence of the computer output varied between cases that were clearly benign, clearly

malignant, or nondescript; this performance is similar to radiologists in that obvious malignant cases and obvious benign cases are easier to diagnose than confusing cases. [82] Another major challenge that AI algorithms are susceptible to is generalizability. To combat the restriction that small and potentially biased datasets have on training new AI algorithms, a few collaborative efforts have been established to produce large, well-curated datasets. [83, 84] Although developers may need to design their algorithms for a specific task on a single system in order to produce useful results, standardized training, testing, and evaluation methods should be used to ensure generalizability among populations and imaging systems. [77, 79, 85] Finally, the ethical use of AI in the clinic needs to be considered. Algorithms that are cleared for clinical use (Section 1.3.2) must be used as intended by the developers, and further, the impact of AI systems on the clinical workflow should be understood prior to implementation. [72]

1.4 Organization of Primary Research Aims

In this dissertation, we aim to address some of the challenges of AI in breast cancer screening, and we explore the role that machine learning can have in breast DCE-MRI. More specifically, we focus on the need for accurate lesion segmentation and a robust method for assessing background parenchymal enhancement (BPE). In Chapter 1, we introduced the background of breast cancer screening with a focus on AI applications to motivate Chapters 2 – 4. In the subsequent chapters, each of these primary aims are then discussed:

- Investigation of machine intelligence for segmentation of lesions and breast regions in breast DCE-MRI (Chapter 2)
- Development of a machine learning technique for computer BPE scoring from breast DCE-MRI (Chapter 3)

- Independent evaluation of the BPE scoring method applied to high-risk screening patients in the task of classifying cancer vs. non-cancer diagnoses, using current and prior MRIs (Chapter 4)

The progression of the dissertation through the primary aims is demonstrated in Figure 1.5, including details on the datasets used for each chapter (further dataset information is provided in Sections 2.3, 3.2, and 4.2).

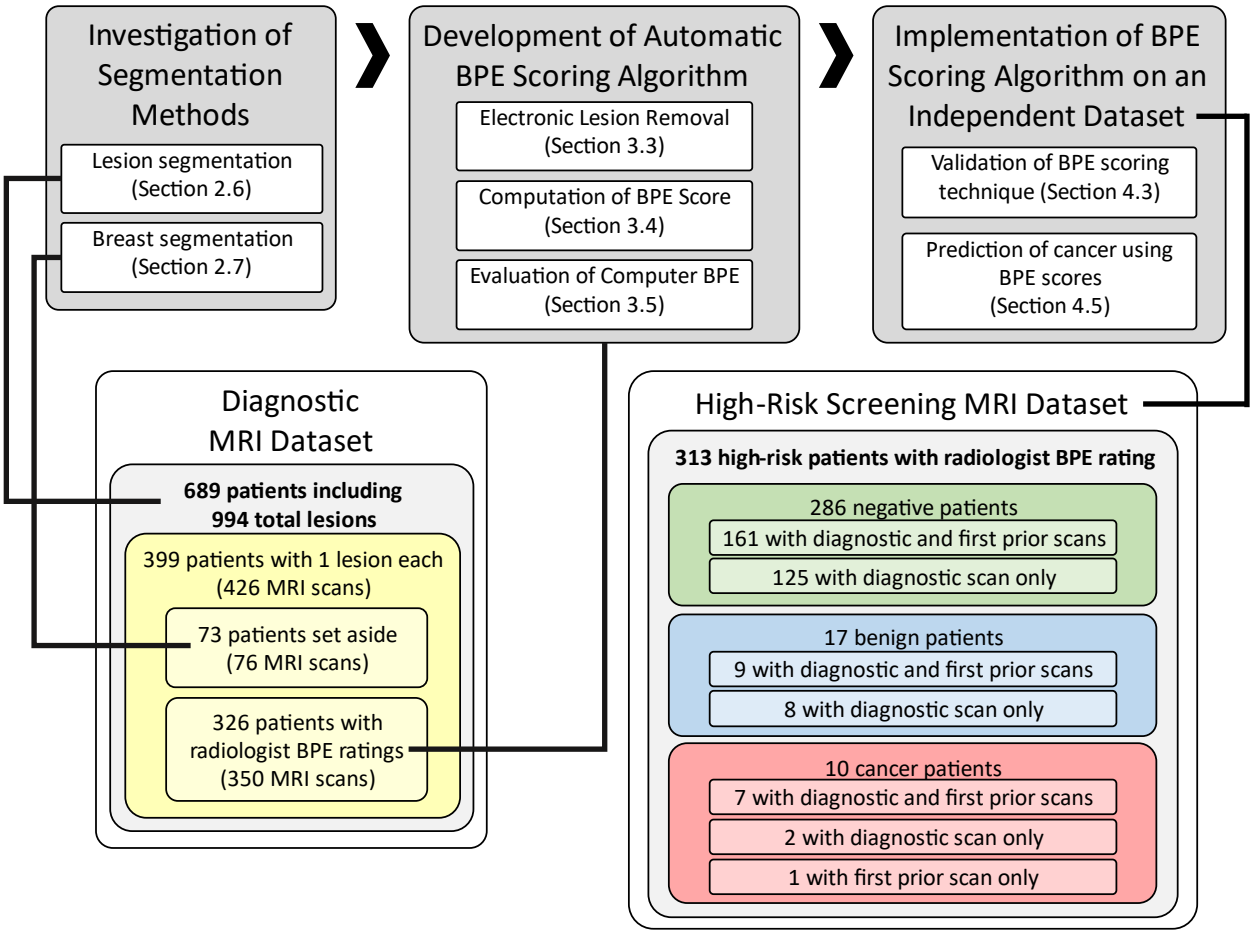

Figure 1.5: Progression between the major components of each of the primary aims completed in this dissertation. Segmentation methods investigated in Chapter 2 were incorporated into the development of the BPE scoring algorithm developed in Chapter 3. Then, the BPE scoring algorithm was implemented on an independent dataset in Chapter 4.

CHAPTER 2

SEGMENTATION OF LESIONS AND BREASTS FROM DCE-MRI

2.1 Introduction

Precise lesion segmentation is required to extract relevant tumor features to be used in computer-aided diagnosis (CADx) systems, and the distinction between enhancement in lesion and non-lesion tissue is necessary for quantitative evaluation of the surrounding fibroglandular tissue (FGT). Segmentation algorithms have been incorporated into many studies that have been designed to identify breasts, lesions, and FGT from breast magnetic resonance imaging (MRI), and they continue to be refined. [86–90] Depending on the subsequent tasks performed using segmented regions, a general location or detailed shape may be sufficient. [88] A variety of methods that have flexible training parameters have been developed for segmentation, such as the fuzzy c-means (FCM) clustering algorithm and the U-Net convolutional neural network. [91, 92] The FCM clustering algorithm has been well-established for breast lesion segmentation on dynamic contrast-enhanced (DCE) MRI; it analyzes the contrast uptake over time and yields volumetric segmentations (Figure 2.1). [91]

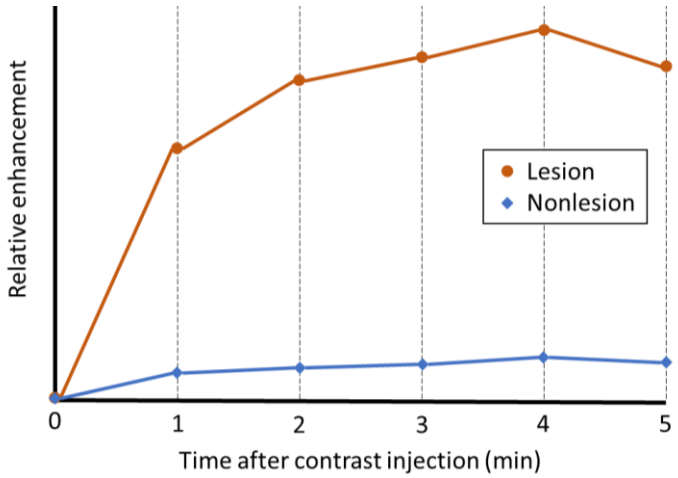

Figure 2.1: Based on the difference in contrast uptake curves, the fuzzy c-means algorithm clusters nonlesion and lesion voxels within a region of interest selected by the user.

The U-Net is a deep learning convolutional neural network (CNN) that produces segmentations based on an image input (Figure 2.2). [92] In 2015, Ronneberger, et. al. originally developed the U-Net to produce accurate biomedical segmentations using only a few training images. [92] The first half of the U-Net, the contracting path, resembles a typical CNN that contains a series of convolutions followed by max pooling operations, and the second half of the U-Net, the expanding path, contains upsampling through a series of transposed convolutions (Figure 2.2). [92] The contracting path captures the image context in high-resolution feature maps. Skip connections pass the high-resolution context features into the corresponding expanding path's feature maps to enable precise localization of the predicted pixel class label (i.e., object or background). [92] Implementation of the U-Net model is relatively simple, as it includes just a few input parameters, including: the image size, batch size, number of epochs, and threshold for the binary segmentation.

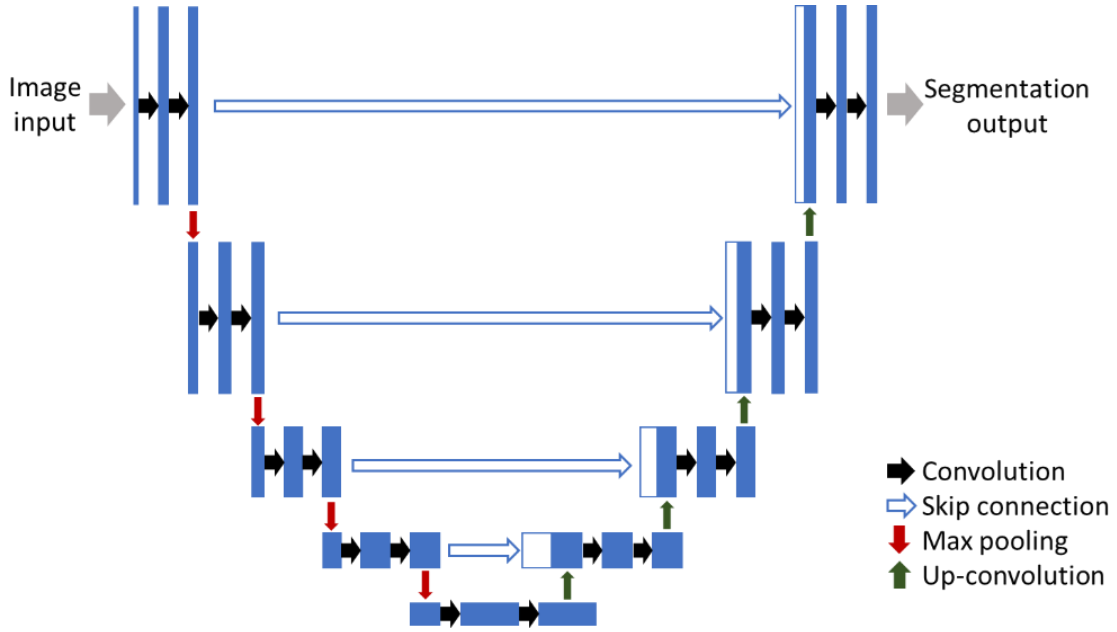

Figure 2.2: The standard U-Net architecture includes a contracting path and symmetric expanding path to produce binary segmentation maps from input images. The first half involves a series of convolutional and pooling layers to encode the image context, and the second half uses transposed convolutions to decode the localization of structures in the image.

Without the need for information from an entire dynamic time series, as required for FCM segmentation, the U-Net offers the benefit of producing accurate segmentations from a single timepoint or from post-contrast subtraction images created from a variety of imaging sequences, including regular and abbreviated DCE-MRI acquisitions. [32] The U-Net architecture has flexibility to be designed to accept either 2D image slices or 3D image volumes. [93] For the former, "quasi-3D" lesion segmentations may be obtained from stacking slice-by-slice segmentations, however, the lack of vertical (out-of-slice) continuity obtained by this "quasi-3D" U-Net may be a potential source of error that a fully 3D U-Net avoids.

In this chapter, we investigated the potential of using U-Nets in breast lesion segmentation on DCE-MRI by comparing the performances of 2D and 3D U-Nets relative to FCM clustering and radiologist delineations. The U-Nets used in our investigations were trained to segment masses and nonmass enhancing lesions from either first or second post-contrast subtraction images, i.e., subtraction images between the first or second post-contrast image and the pre-contrast image. Additionally, we trained a 2D U-Net for whole breast segmentation from maximum intensity projections of the second post-contrast subtraction MRIs.

2.2 Methods

The viability of using U-Nets in breast lesion segmentation on DCE-MRI was assessed by comparing the performances of 2D and 3D U-Nets in four evaluations. First, in Comparison A, quasi-3D and 3D U-Nets were compared to FCM, which served as a surrogate reference standard. Second, in Comparison B, the 2D U-Net, 3D U-Net, and FCM segmentations were compared to 2D radiologist delineations on lesion center slices for a subset of 71 lesions. Next, in Comparison C, segmentations from first post-contrast subtraction images were compared to second post-

contrast subtraction images for quasi-3D and 3D U-Nets. Lastly, in Comparison D, the segmentation performance of each method was evaluated for mass versus nonmass enhancing lesions.

2.3 Dataset

The dataset consisted of DCE-MRIs of 994 unique breast lesions (724 malignant and 270 benign) from 689 patients (each with 1-5 diagnosed lesions) aged 23-89 years. The deidentified data were retrospectively collected at the University of Chicago over a span of 9 years (from 2005 to 2013) under Health Insurance Portability and Accountability Act (HIPAA)-compliant Institutional Review Board (IRB)-approved (11606B) protocols. Routine bilateral breast MRI was performed using a Philips Achieva scanner with either 1.5T (N = 473) or 3.0T (N = 216) magnet strength. The breast DCE-MRI protocol included a fat-saturated 3D T1 weighted spoiled gradient-echo sequence that was used to acquire pre- and post-contrast images with a temporal resolution of 60-75 seconds (TE = 2.2-2.8 ms, TR = 4.5-7.5 ms, flip angle = 10-20°, in-plane resolution = 0.5-1.0 mm, FOV = 28.0-44.1 cm, matrix = 320–552 x 256–525, slice thickness = 1-3.5 mm, interslice gap = 0.8-2.5mm). Table 2.1 contains the clinical characteristics of the data obtained from pathology and radiology reports, including pathological truth (benign or malignant) and lesion type (mass or nonmass enhancement). A subset of 71 lesions were manually selected for radiologist delineations so that the distribution of pathological truth and lesion type within the subset remained similar to the overall group distribution (Table 2.1). Table 2.2 presents size distributions of the lesions.

Table 2.1: Summary of the DCE-MRI dataset by lesion type. Lesions were categorized by pathological truth and enhancement type. Lesions that were not marked as either mass or nonmass enhancing were not selected for the radiologist-delineated subset and were labeled "unknown."

	Enhancement type	Pathological 'truth'	
		Benign	Malignant
All lesions (N = 994)	**Mass**	170	517
	Nonmass	49	175
	Unknown	51	32
Subset of lesions outlined by radiologist (N = 71)	**Mass**	14	40
	Nonmass	7	10
	Unknown	0	0

Table 2.2: Summary of the DCE-MRI dataset by lesion size. Lesions were categorized by effective diameter (mm), defined by $2*\sqrt{(A/\pi)}$, where A is the area of the lesion in the center slice of the FCM segmentation in mm^2.

Effective lesion diameter (mm):	<5	5–9	10–14	15–19	>20	All
All lesions	64	344	252	125	209	994
Subset of lesions outlined by radiologist	2	11	32	15	11	71

2.4 Establishment of Reference Standards and Preprocessing

Each lesion had previously been segmented using a well-established, in-house, automated 3D FCM approach that yielded, as a surrogate reference standard, a 3D binary lesion segmentation. [91] FCM segmentation was performed within a region defined by a human operator's selection of a rectangular bounding box about the lesion in a middle slice along with an indication of the first and last slices in which the lesion appeared. [91] The bounding-box volume of interest (VOI) for the FCM segmentation of each lesion was also used as input for subsequent U-Net segmentations of post-contrast subtraction images. Second post-contrast subtraction images were primarily used as inputs for the U-Net, however, first post-contrast subtraction images were introduced for evaluation in Comparisons C and D of this study.

In addition, an expert radiologist (7 years of experience in breast imaging) manually delineated the lesion within the center slice of the second post-contrast subtraction VOI for the subset of 71 lesions. Here, the radiologist segmentations were used as the reference standard for Comparison B of this study. Since radiologist segmentations were only available for a limited set of center slices and FCM segmentations are used in an FDA-approved clinical breast MRI workstation, [12] FCM segmentations served as a reasonable surrogate reference standard to train the U-Net architectures.

2.5 U-Net Architectures

Two different U-Net architectures were evaluated in this study. The first was a 2D U-Net. [92] We found that the top and bottom slices of lesions were most difficult to segment, so those two slices were excluded from each lesion in training (though they remained in the test set lesions). The image slices of each lesion's VOI were resized, by interpolation with a preserved pixel value range, to 256 x 256 pixels prior to input into the 2D U-Net. The 256 x 256 pixels probability map outputs, with values ranging from 0 to 1, were converted to binary segmentation images based on a threshold of 0.25. The 2D U-Net only processed one image slice at a time, so "quasi-3D" lesion segmentations were produced by stacking the 2D slice-by-slice segmentations obtained by the 2D U-Net to form a 3D volume (Figure 2.3). In this work, "quasi-3D U-Net" refers to the volumetric segmentation produced by the 2D U-Net architecture.

The second architecture evaluated in this study was a 3D U-Net. [93] This network is similar to the structure of the 2D U-Net, but it was modified with the added third dimension. Prior to input into the 3D U-Net, the lesion VOIs were resized, by interpolation with a preserved pixel value range, to 256 x 256 x N voxels (N is the number of slices in the lesion). The network

produced 256 x 256 x N voxel probability map outputs, with values ranging from 0 to 1, which were converted to binary segmentation volumes based on a threshold of 0.23. The threshold for the binary conversion was selected from a range of values between 0.14 and 0.30 to produce the greatest mean Dice similarity coefficient calculated from the resulting segmentations and their references, hence the slight difference in threshold used for 2D U-Net and 3D U-Net.

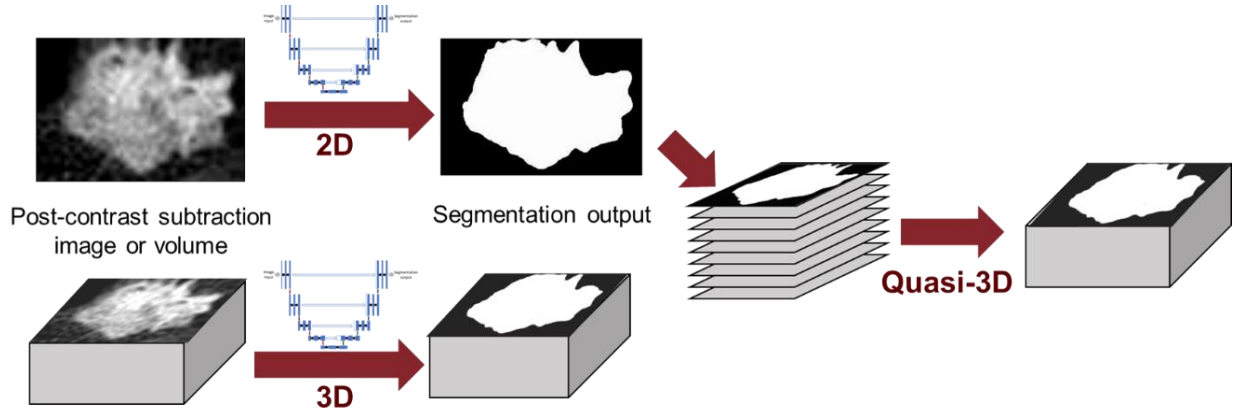

Figure 2.3: Representation of the U-Net segmentations investigated for breast lesion segmentation from DCE-MRI regions of interest. The standard 2D U-Net produced 2D segmentations that were stacked to create quasi-3D segmentations, and the 3D U-Net architecture was modified for 3D segmentations from volumetric inputs.

2.6 Training and Statistical Analysis of Segmentation Performances

Five-fold cross-validation by lesion (N = 994 lesions) was conducted to train and evaluate the U-Net models. The folds were partitioned such that each fold contained a similar distribution of lesion types based on pathological truth (malignant or benign), lesion enhancement type, and lesion size. Since adjacent slices within the same lesion VOI often are very similar in appearance, all slices belonging to a given lesion were always allocated to the same fold. Training and test folds were allocated by lesion, i.e., not by slice or patient.

Dice similarity coefficient (DSC) and Hausdorff distance (HD) were used to evaluate the performances of different segmentation methods relative to the specific reference standard (Figure

2.4). Note that greater segmentation performance is indicated by higher DSCs and lower HDs. Predictions from the quasi-3D and 3D U-Nets were resized to their original lesion VOI dimensions before DSCs and HDs were calculated between the predictions and the reference standards. HDs were calculated for each slice, and the median HD for each lesion was used for 3D performance comparisons. To assess the statistical significance of difference in performance, the Wilcoxon signed-rank test was used for matched cases in Comparisons A, B, and C, and the Mann-Whitney U-test was used in Comparison D for analysis of unmatched cases. [94–97] The Bonferroni correction was used to correct p-values for multiple comparisons in Comparisons B, C, and D. [98]

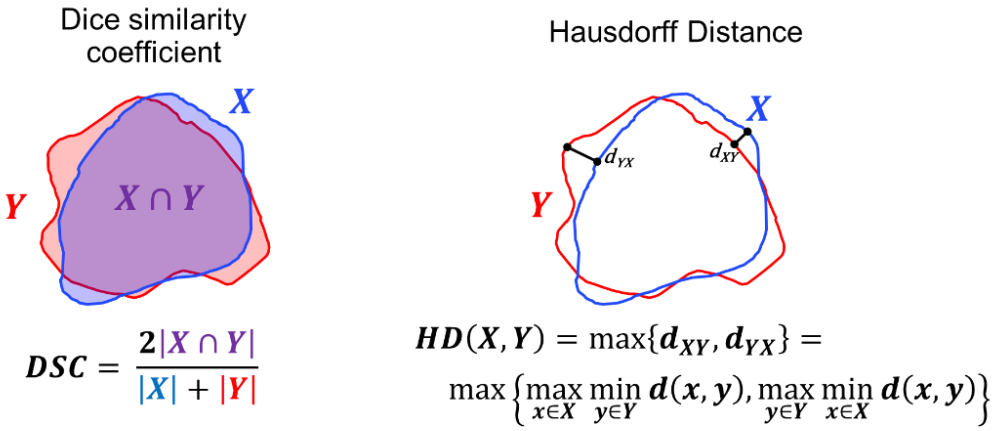

Figure 2.4: Dice similarity coefficient (DSC) is a measure of how well two regions, X and Y, overlap, and Hausdorff distance (HD) is a measure of how well the margins of two regions agree. DSC and HD were used to evaluate the performances of the different segmentation methods relative to the specific reference standard.

2.6.1 Comparison A: Comparing Quasi-3D U-Net to 3D U-Net Using FCM as the Surrogate Reference Standard

The volumetric segmentations from quasi-3D and 3D U-Nets were compared and as previously noted, FCM segmentations served as the surrogate reference standard for the 994 lesions (Table 2.1). The Wilcoxon signed-rank test was used to assess statistically significant differences between quasi-3D and 3D U-Net segmentation performances (Figure 2.5).

2.6.2 Comparison B: Comparing FCM, Quasi-3D U-Net, and 3D U-Net Using Radiologist-Delineations as the Reference Standard

Next, FCM, quasi-3D U-Net, and 3D U-Net center slice segmentations were compared using the radiologist references available for the subset of 71 lesions. For each of the three segmentation methods, DSCs and HDs were calculated on the center slice with respect to the radiologist reference. Statistically significant differences between quasi-3D U-Net, 3D U-Net, and FCM segmentations were assessed using the Wilcoxon signed-rank test including a Bonferroni correction (Figure 2.6).

2.6.3 Comparison C: Comparing Segmentation Across Post-Contrast Timepoints (First vs. Second Post-Contrast)

The segmentations obtained in Comparison A using second post-contrast subtraction images as input were compared to those using the first post-contrast subtraction images; for the quasi-3D and 3D U-Nets as compared to the FCM reference standard. Wilcoxon signed-rank tests were used to assess statistical significance between the results after a Bonferroni correction.

2.6.7 Comparison D: Comparing Segmentation Across Lesion Enhancement Types (Mass vs. Nonmass Enhancement)

Finally, the segmentation performances on mass and nonmass enhancing lesions were compared. The segmentations resulting from the first and second post-contrast subtraction inputs to the quasi-3D and 3D U-Nets evaluated in Comparison C were compared based on lesion enhancement type. For each comparison, a Mann-Whitney U-test including a Bonferroni correction for statistical significance was used to compare the segmentation performances of the set of mass lesions to the set of nonmass lesions.

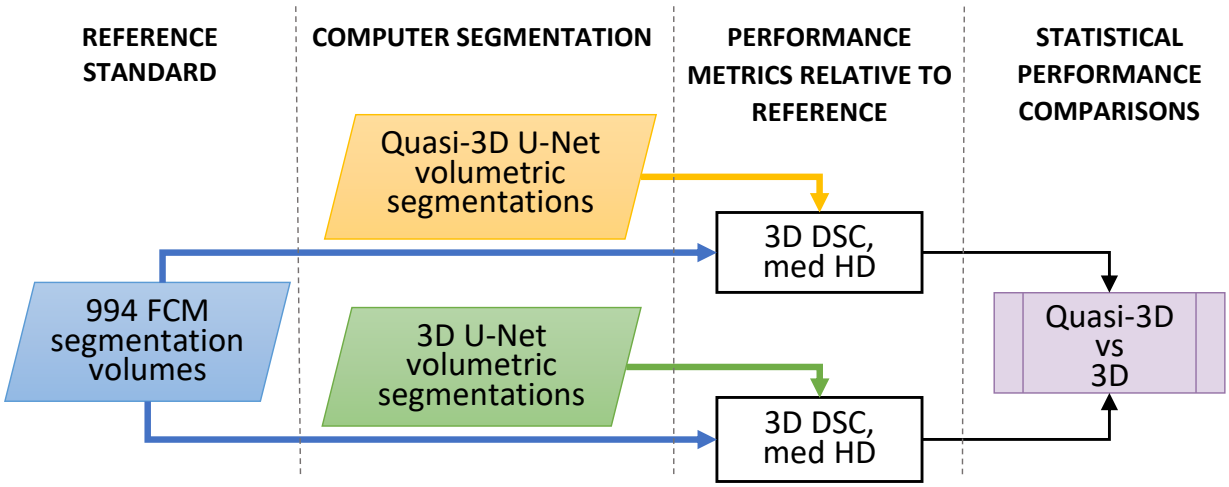

Figure 2.5: Flowchart of Comparison A of this study (N = 994). Fuzzy c-means (FCM) lesion segmentation volumes were used as reference standard to compare quasi-3D U-Net (2D architecture) and 3D U-Net segmentations in a by-lesion five-fold cross-validation process. DSC: Dice similarity coefficient; med HD: median Hausdorff distance of all slices in the lesion. Wilcoxon signed-rank tests were performed on the resulting paired data.

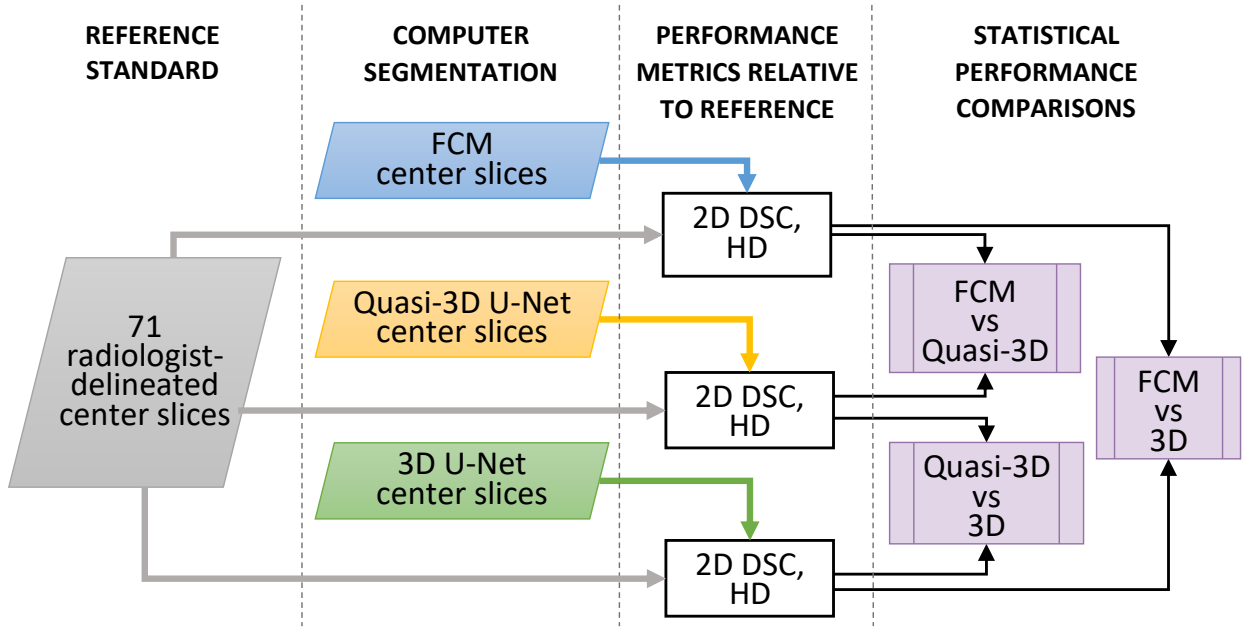

Figure 2.6: Flowchart of Comparison B of this study (N = 71). Radiologist segmentations were used as reference standard to compare fuzzy c-means (FCM), quasi-3D U-Net, and 3D U-Net center slice segmentations. DSC: Dice similarity coefficient; HD: Hausdorff distance. Wilcoxon signed-rank tests were performed on the resulting paired data.

2.7 Breast Segmentation

Beyond lesion segmentation, whole breast segmentation was investigated. From the original set of 689 patients, a new subset of 76 exams from 73 patients that were diagnosed with only one lesion was selected to contain a variety of lesion sizes and BPE levels (6 minimal, 18 mild, 26 moderate, 11 marked, and 15 unknown BPE). For this subset, an expert radiologist (7 years of experience in breast imaging) provided manual delineations of the breast margins on the maximum intensity projection (MIP) of the second post-contrast subtraction image volume. The radiologist-delineated breast margins were used as the reference standard for training a 2D U-Net convolutional neural network [92] for whole breast segmentation from the MIPs. Without a reference for the remaining test cases, visual assessment was used to ensure the binary mask sufficiently contained the entire breast region with minimal pixels from the chest wall.

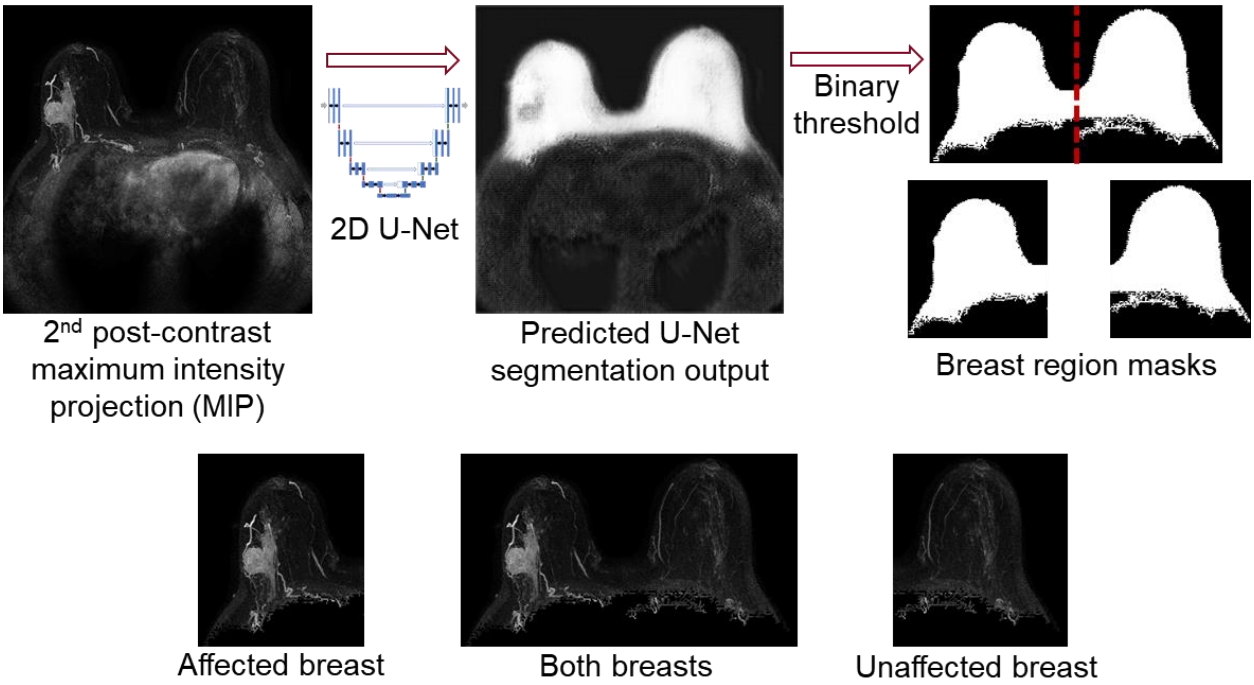

Figure 2.7: A 2D U-Net was trained for computerized breast segmentation on maximum intensity projections (MIP) of second post-contrast subtraction DCE-MRIs. A binary threshold was applied to the predicted U-Net output to generate breast region masks, and the individual breast regions were created by a vertical split at the center of the breast region containing both breasts.

To produce the breast region masks for use in our method for automatic background parenchymal enhancement (BPE) assessment (Chapter 3), a binary threshold was applied to the predicted U-Net outputs, followed by a post-processing step to identify the largest object from the mask as the region containing both breasts. The region containing both breasts was vertically split between the left and right sides to generate masks defining only the affected breast region and only the unaffected breast region. These breast masks were applied to the full post-contrast subtraction projection images to retain only the pixels belonging to both breasts, the affected breast, or the unaffected breast (Figure 2.7). The resulting breast regions were used for automatic scoring of BPE in Chapter 3.

2.8 Results

2.8.1 Comparison A: Comparing Quasi-3D U-Net to 3D U-Net Using FCM as the Surrogate Reference Standard

Segmentation performance was assessed by comparing the medians of DSC and HD (Table 2.3). Note that greater segmentation performance is indicated by higher DSCs and lower HDs. Of the 994 lesions in the dataset, the 3D U-Net failed to segment 6 lesions (from 3 unique patients) that were less than 9.1 mm in effective diameter ($2*\sqrt{(A/\pi)}$) and had an unknown enhancement type. Without prediction volumes available to compare to the reference standard, DSC was essentially zero and it was impossible to calculate HDs, therefore these lesions were excluded from HD statistical comparisons for the 3D U-Net. The results of the Wilcoxon signed-rank test show that the quasi-3D U-Net outperformed the 3D U-Net with statistical significance in terms of DSC ($p < 0.001$) and HD ($p < 0.001$) for lesion segmentation from second post-contrast subtraction VOIs.

Table 2.3: Comparison A: Summary statistics of the performance metrics of quasi-3D and 3D U-Nets as compared to fuzzy c-means (FCM) reference standards for volume segmentation. Minimum, maximum, and median values of DSC and HD metrics of all cases are shown in the table. U-Nets were trained and tested using five-fold cross validation by lesion. Parenthetical values represent 95% confidence intervals. Asterisks indicate statistically significantly greater performance after Bonferroni correction for two comparisons. DSC: Dice similarity coefficient; HD: Hausdorff distance. (N = 994)

Segmentation Method	Min DSC	Max DSC	Median DSC	Min HD	Max HD	Median HD
Quasi-3D U-Net	0.270	0.955	0.780* (0.774, 0.787)	0.737mm	73.6mm	7.30mm* (6.79, 7.72)
3D U-Net	0†	0.935†	0.721† (0.710, 0.732)	0.741mm††	98.2mm††	7.53mm†† (6.84, 7.97)

† Excluding six lesions with DSC = 0, minimum DSC = 0.035, maximum DSC = 0.935, and median DSC = 0.721 (0.710, 0.733).

†† Due to failed segmentation, six lesions were excluded from 3D U-Net HD results because HD could not be calculated for those lesions.

2.8.2 Comparison B: Comparing FCM, Quasi-3D U-Net and 3D U-Net Using Radiologist-Delineations as the Reference Standard

Based on the segmentation results for the subset of 71 lesions, we found that the center slices from each lesion segmentation produced by FCM, quasi-3D U-Net, and 3D U-Net were had good agreement (DSC, HD) with the radiologist-segmented reference standard (Table 2.4). The statistical comparisons of performance between each segmentation method's agreement with the reference standard is shown in Table 2.5. The results indicate that quasi-3D U-Net statistically significantly outperformed both 3D U-Net and FCM for lesion segmentation on second post-contrast subtraction center slices. We observed improved U-Net segmentation agreement with the radiologist reference as lesion size increased, and Figure 2.8 shows quasi-3D U-Net yielded greater DSC values than 3D U-Net, relative to radiologist delineations, across lesion sizes.

Table 2.4: Comparison B: Summary statistics of the performance metrics of fuzzy c-means (FCM), quasi-3D U-Net, and 3D U-Net, as compared to radiologist reference standard for center slice segmentation. Minimum, maximum, and median DSC and HD metrics of all cases are shown in the table. U-Nets were trained and tested using five-fold cross validation by lesion. Parenthetical values represent 95% confidence intervals. (N = 71) DSC: Dice similarity coefficient; HD: Hausdorff distance.

Segmentation Method	Min DSC	Max DSC	Median DSC	Min HD	Max HD	Median HD
FCM	0.209	0.961	0.832 (0.813, 0.859)	796mm	31.2mm	4.06mm (3.38, 4.80)
Quasi-3D U-Net	0.274	0.959	0.864 (0.845, 0.889)	0.796mm	28.5mm	3.28mm (2.97, 4.50)
3D U-Net	0.246	0.952	0.802 (0.766, 0.834)	1.13mm	26.6mm	4.17mm (3.04, 5.30)

Table 2.5: Comparison B: Statistical comparisons between the median performance metrics in Table 2.4 from fuzzy c-means (FCM), quasi-3D U-Net, and 3D U-Net center slice predictions using radiologist-delineations as the reference standard. U-Nets were trained and tested using five-fold cross validation by lesion. Raw, uncorrected p-values from the Wilcoxon signed-rank test are reported in the table; statistical significance was assessed after correcting for three comparisons. (N = 71) DSC: Dice similarity coefficient; HD: Hausdorff distance.

Segmentation	DSC comparisons	HD comparisons
FCM vs. Quasi-3D U-Net	Quasi-3D U-Net outperformed FCM (p = 2.92e-5)	Quasi-3D U-Net outperformed FCM (p = 6.28e-3)
Quasi-3D U-Net vs. 3D U-Net	Quasi-3D U-Net outperformed 3D U-Net (p = 4.13e-9)	Quasi-3D U-Net outperformed 3D U-Net (p = 0.014)
FCM vs. 3D U-Net	FCM outperformed 3D U-Net (p =1.87e-4)	Failed to reach statistical significance (p = 0.753)

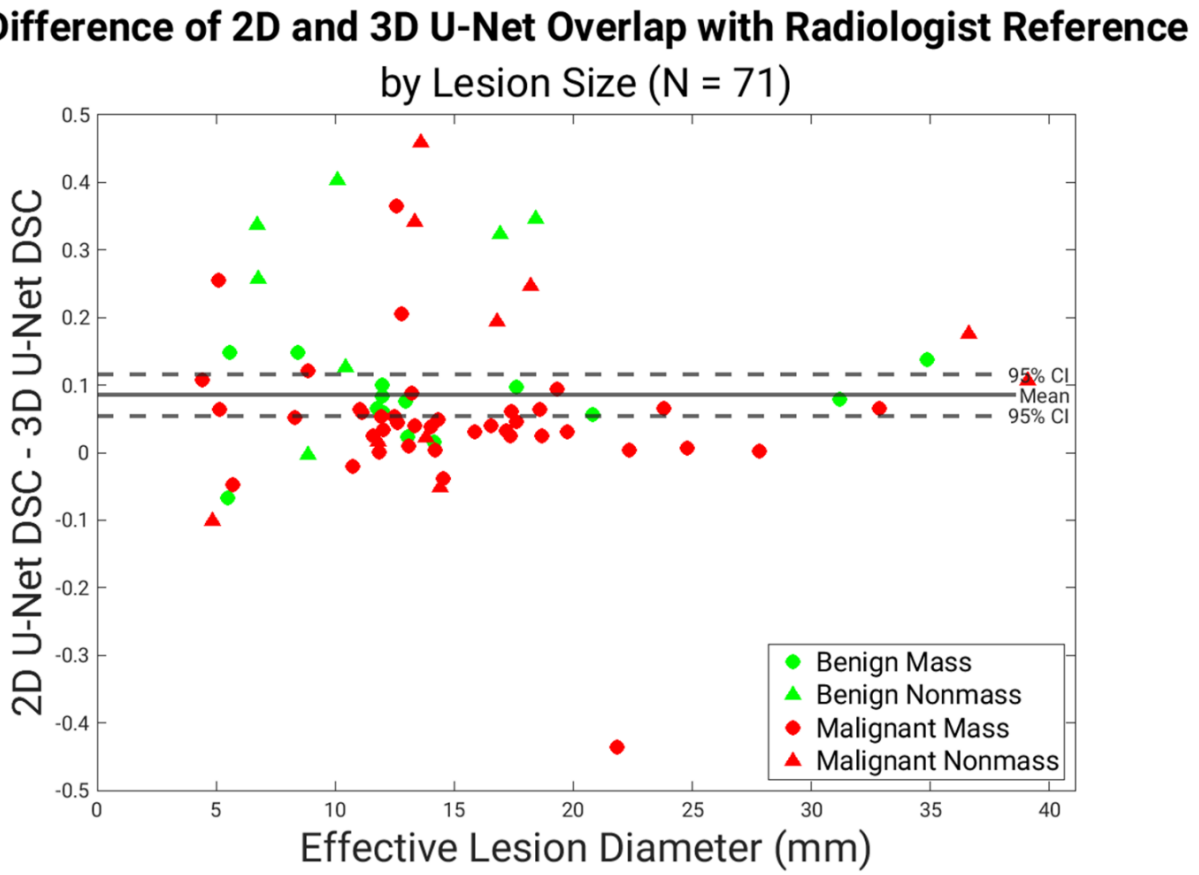

Figure 2.8: Difference in DSC calculated from the center slice of the quasi-3D (2D) U-Net or 3D U-Net and the radiologist reference, shown versus lesion size. The majority of lesions yielded greater agreement between the radiologist and the quasi-3D (2D) U-Net than with the 3D U-Net. DSC: Dice similarity coefficient.

2.8.3 Comparison C: Comparing Segmentation Across Post-Contrast Timepoints (First vs. Second Post-Contrast)

An example of the segmentations produced by the 2D U-Net, 3D U-Net, FCM, and radiologist for a mass and nonmass enhancing lesion are shown in Figure 2.9. In the second post-contrast subtraction images, more lesion enhancement is provided to the U-Net, which, as expected, tended to result in segmentations that that more closely resembled FCM than segmentations from the first post-contrast subtraction inputs. Also as expected, the radiologist delineations acquired on the

central slice of the second post-contrast subtraction image tended to resemble the center slice of the second post-contrast subtraction segmentation from the 2D U-Net.

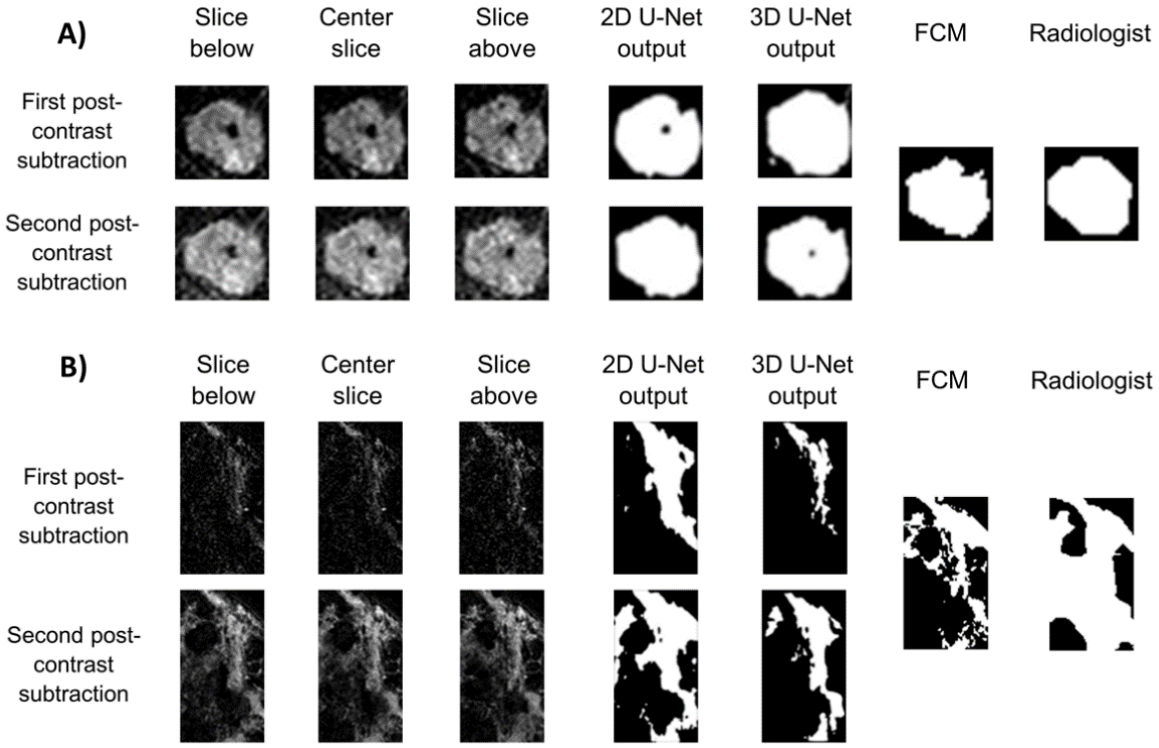

Figure 2.9: Example cases showing the center slice U-Net segmentations produced from the first or second post-contrast subtraction images for a A) mass enhancing lesion and B) nonmass enhancing lesion. The center slice fuzzy c-means (FCM) and radiologist references are also shown.

The performance metrics calculated for the segmentations produced by the U-Nets from first and second post-contrast subtraction inputs are included in Table 2.6. The statistical comparisons between the resulting DSC and HD metrics for each method are shown in Table 2.7 and Table 2.8. The results showed statistically significantly greater performance from the second post-contrast subtraction inputs than from the first post-contrast subtraction inputs using the quasi-3D and 3D U-Nets, except in the case of nonmass enhancing lesions using the 3D U-Net. The results from both the first and second post-contrast subtraction inputs supported the results found in Comparisons A and B; the quasi-3D U-Net statistically significantly outperformed the 3D U-

Net for the combined lesion types based on DSC (however, HD failed to show statistically significant differences between quasi-3D and 3D U-Net for the first post-contrast subtraction inputs).

2.8.4 Comparison D: Comparing Segmentation Across Lesion Types (Mass vs. Nonmass Enhancement)

The results in Table 2.6 – Table 2.8 demonstrated that, relative to the FCM reference standard, mass lesion segmentation statistically significantly outperformed nonmass enhancing lesion segmentation using first and second post-contrast subtraction image inputs to both the quasi-3D and 3D U-Nets. For nonmass enhancing lesions, quasi-3D U-Net always statistically significantly outperformed the 3D U-Net (as in Comparisons A and B). For mass lesions, the DSC results indicate that quasi-3D U-Net statistically significantly outperformed the 3D U-Net (as in Comparison B), however HD results from the first post-contrast subtraction inputs showed that the 3D U-Net statistically significantly outperformed quasi-3D U-Net.

Table 2.6: Comparisons C & D: Summary statistics of the performance metrics of quasi-3D U-Net and 3D U-Net, as compared to fuzzy c-means (FCM) surrogate reference standard. U-Nets were trained and tested using five-fold cross validation by lesion. (N_{mass} = 687; $N_{nonmass}$ = 224) DSC: Dice similarity coefficient; HD: Hausdorff distance.

Input	U-Net Model	Lesion Type	Median DSC	Median HD (mm)
1st post-contrast subtraction images	Quasi-3D	Mass	0.7492	6.9375
		Nonmass	0.6126	12.0575
	3D	Mass	0.7357	6.6667
		Nonmass	0.5858	12.9417
2nd post-contrast subtraction images	Quasi-3D	Mass	0.8059	6.8838
		Nonmass	0.6993	11.0459
	3D	Mass	0.7668	6.7734
		Nonmass	0.5458	15.1173

Table 2.7: Comparisons C & D: Statistical results for comparisons between input image type and lesion type using the Dice similarity coefficients of U-Net segmentations against fuzzy c-means reference. U-Nets were trained and tested using five-fold cross validation by lesion. Colors indicate the comparisons: Orange: mass lesions vs. nonmass lesions for a given U-Net and timepoint combination (Mann-Whitney U-test), Pink: Quasi-3D (2D) vs. 3D U-Net for a fixed lesion type and timepoint (Wilcoxon signed-rank test), Blue: first vs second post-contrast subtraction for a fixed lesion type and timepoint (Wilcoxon signed-rank test). Raw, uncorrected p-values are reported in the table; however, statistical significance (indicated by an asterisk) was assessed including a Bonferroni correction for 4 comparisons per sample. (N_{mass} = 687; $N_{nonmass}$ = 224)

Input			1st post-contrast subtraction images				2nd post-contrast subtraction images			
	U-Net		Quasi-3D		3D		Quasi-3D		3D	
		Lesions	Mass	Nonmass	Mass	Nonmass	Mass	Nonmass	Mass	Nonmass
1st post-contrast subtraction images	Quasi-3D	Mass	-	p < 0.001 *	p < 0.001 *	-	p < 0.001 *	-	-	-
		Nonmass		-	-	p < 0.001 *	-	p < 0.001 *	-	-
	3D	Mass			-	p < 0.001 *	-	-	p < 0.001 *	-
		Nonmass				-	-	-	-	p = 0.004
2nd post-contrast subtraction images	Quasi-3D	Mass					-	p < 0.001 *	p < 0.001 *	-
		Nonmass						-	-	p < 0.001 *
	3D	Mass							-	p < 0.001 *
		Nonmass								-

Table 2.8: Comparisons C & D: Statistical results for comparisons between input image type and lesion type using the Hausdorff distance of U-Net segmentations against fuzzy c-means reference. U-Nets were trained and tested using five-fold cross validation by lesion. Colors indicate the comparisons: Orange: mass lesions vs. nonmass lesions for a given U-Net and timepoint combination (Mann-Whitney U-test), Pink: Quasi-3D (2D) vs. 3D U-Net for a fixed lesion type and timepoint (Wilcoxon signed-rank test), Blue: first vs second post-contrast subtraction for a fixed lesion type and timepoint (Wilcoxon signed-rank test). Raw, uncorrected p-values are reported in the table; however, statistical significance (indicated by an asterisk) was assessed including a Bonferroni correction for 4 comparisons per sample. (N_{mass} = 687; $N_{nonmass}$ = 224)

Input			1st post-contrast subtraction images				2nd post-contrast subtraction images			
	U-Net		Quasi-3D		3D		Quasi-3D		3D	
		Lesions	Mass	Nonmass	Mass	Nonmass	Mass	Nonmass	Mass	Nonmass
1st post-contrast subtraction images	Quasi-3D	Mass	-	p < 0.001 *	p = 0.004	-	p < 0.001 *	-	-	-
		Nonmass		-	-	p < 0.001 *	-	p < 0.001 *	-	-

Table 2.8 (continued): Statistical results for comparisons between input image type and lesion type using the Hausdorff distance of U-Net segmentations against fuzzy c-means reference.

Input			1st post-contrast subtraction images				2nd post-contrast subtraction images			
	U-Net		Quasi-3D		3D		Quasi-3D		3D	
		Lesions	Mass	Nonmass	Mass	Nonmass	Mass	Nonmass	Mass	Nonmass
1st post-contrast subtraction images	3D	Mass			-	p < 0.001 *	-	-	p = 0.298	-
		Nonmass				-	-	-	-	p < 0.001 *
2nd post-contrast subtraction images	Quasi-3D	Mass					-	p < 0.001 *	p = 0.212	-
		Nonmass						-	-	p < 0.001 *
	3D	Mass							-	p < 0.001 *
		Nonmass								-

2.9 Discussion

This chapter explored the performance of volumetric segmentations obtained with a 2D U-Net (quasi-3D U-Net) and a 3D U-Net. Segmentation performance was assessed against a well-established FCM method, which served as a surrogate reference standard, or radiologist reference segmentation.

Results of the investigations found that there were statistically significant differences in performance between U-Net and FCM segmentation methods, relative to each other and to a radiologist reference segmentation. In the task of segmenting breast lesions from second post-contrast subtraction DCE-MRI VOIs, the quasi-3D U-Net statistically significantly outperformed the 3D U-Net in segmenting volumes, despite assumed advantages from vertical context (N = 988). Additionally, the comparison between center slices from FCM, quasi-3D U-Net, and 3D U-Net relative to the radiologist reference found that 2D U-Net (quasi-3D U-Net) statistically significantly outperformed FCM and 3D U-Net (N = 71). Relative to FCM volumes, U-Net

segmentations of second post-contrast subtraction inputs were statistically significantly greater than first post-contrast subtraction inputs, and segmentation of mass lesions statistically significantly outperformed nonmass lesion segmentation. Although improved segmentation results were found using second post-contrast subtraction inputs, the 2D U-Net (quasi-3D U-Net) statistically significantly outperformed the 3D U-Net for the first post-contrast subtraction inputs; this could provide a potential benefit to abbreviated MRI applications. Ultimately, the results of these investigations demonstrated that using a 2D U-Net to yield quasi-3D U-Net segmentation of breast lesions from post-contrast subtraction DCE-MRIs could be an effective alternative to FCM or 3D U-Net.

2.10 Additional Evaluations

Beyond the investigations using the standard U-Net architectures discussed in this chapter, we also performed a number of preliminary experiments using an Attention U-Net, which has been shown to outperform the standard U-Net in a number of medical image segmentation tasks since its development by Oktay et. al. in 2018 (Figure 2.10). [99] The attention blocks in the modified U-Net architecture were designed to focus the network on the image subject by suppressing the features originating from the background of the image. At each step of the decoding path, an attention gate combines the useful feature information from the next-lowest layer of the network with the useful spatial information from the symmetric layer of the encoding path (Figure 2.10). Weights are assigned to each of the pixels based on their relevance, and with each training epoch, the weights become more refined to emphasize the foreground pixels and suppress the background.

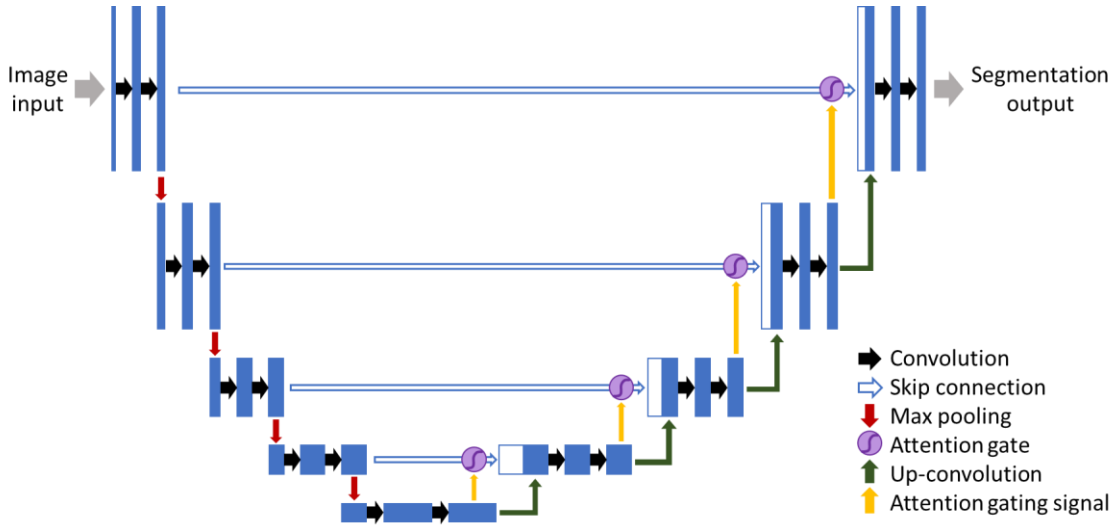

Figure 2.10: Attention U-Net involves attention gates that filter the image features through skip connections at each step of the encoding path. Each pixel weight corresponds to its relevance, which is based on the encoding path spatial information and decoding path feature information.

In our prior studies, we observed instances where the U-Net would overestimate the nonmass lesions and underestimate the mass lesions, so we predicted that attention gating would improve the segmentation predictions by focusing the network on the lesions and drawing attention away from the background tissue. Based on the FCM reference standard, we failed to show a statistically significant improvement in the overall segmentation performance using an attention-gated U-Net. Figure 2.11 shows examples of lesions that had improved or worsened segmentations using the attention-gated 3D U-Net versus the standard 3D U-Net. Even though we did not observe overall segmentation improvements, we evaluated the change in DSC based on the original segmentation performance (Figure 2.12). Although we did not evaluate statistical significance, it appeared that the attention-gated U-Net may offer an improvement for lesions with poor results (DSC < 0.5) from the original U-Net, whereas the lesions with reasonable results (DSC > 0.5) from the original U-Net would not benefit. We speculate that the attention U-Net would more likely provide benefits to segmentation tasks that involve larger regions of interest, unlike the bounding-box ROIs used in our study that contained limited background pixels.

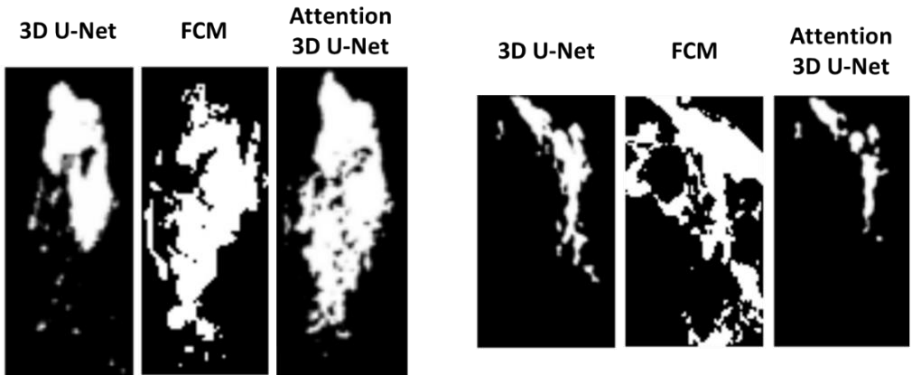

Figure 2.11: Nonmass enhancing lesions that had improved (left) or worsened (right) segmentation performance using the attention-gated 3D U-Net instead of the standard 3D U-Net. Compared to the fuzzy c-means reference, the lesion on the left had a dice similarity coefficient (DSC) increase from 0.25 to 0.63, and the lesion on the right had a DSC decrease from 0.5 to 0.37.

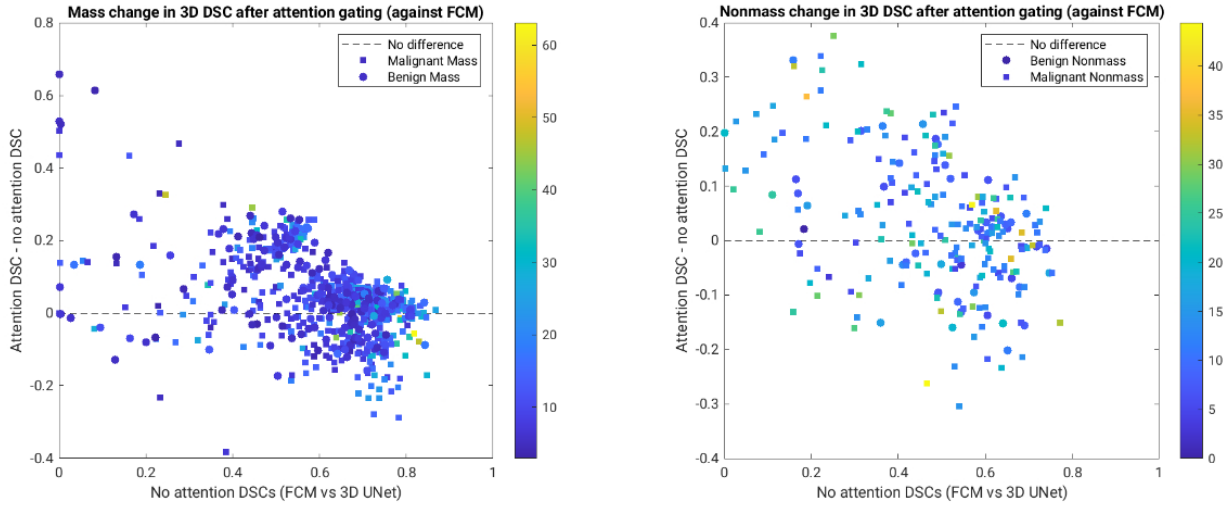

Figure 2.12: Change in dice similarity coefficient (DSC) between the fuzzy c-means (FCM) reference and the U-Net segmentations produced with and without attention-gating, for mass lesions (left) and nonmass lesions (right). For all lesions that had an original DSC of less than 0.5 using the 3D U-Net without attention, the median change in DSC after attention-gating was 0.113. For all lesions that had an original DSC of greater than 0.5 using the 3D U-Net without attention, the median change in DSC after attention-gating was 0.021. Colors indicate lesion size (mm).

In addition to the investigations using the attention-gated U-Net, we also did a preliminary evaluation of segmentation for abbreviated DCE-MRI sequences. For these experiments, we used the same first post-contrast subtraction images from the dataset of lesions used in sections 2.8.3 and 2.8.4. Although the images were acquired under a routine full DCE protocol, the first post-contrast image could be considered equivalent to the post-contrast image acquired during an

abbreviated sequence. The FCM surrogate reference standard was replaced by a "simulated-abbreviated" FCM reference, which was generated from the pre- and first post-contrast portions of the contrast uptake curve (Figure 2.1). We did not observe any notable changes in performance for the first post-contrast subtraction U-Net segmentations compared to the FCM reference versus the simulated-abbreviated-FCM reference. However, thorough statistical analysis was not performed, and the agreement between FCM and simulated-abbreviated-FCM was not evaluated.

2.11 Limitations and Future Work

There were several limitations of the work in this chapter. Firstly, the segmentation performances were evaluated within bounding-box VOIs; the inputs to the U-Net were based on the FCM volume dimensions. However, use of the bounding-box VOIs does mimic clinical practice where a radiologist may roughly indicate the region about a lesion as input to automatic characterization and computer-aided diagnosis. Also, there were a limited number of radiologist segmentations available, each acquired for the second post-contrast subtraction center slices (N = 71 lesions); this could have influenced the results of the comparisons performed in Comparison B. Additionally, there were no radiologist volume segmentations available for full lesion volumes, so FCM segmentations were used as surrogate reference standards. Finally, the 3D U-Net architecture may not be considered fully 3D since many lesions had too few slices to properly pool in the axial dimension. Future work may include an extended investigation of U-Net performance for breast lesion segmentation from abbreviated DCE-MRI sequences. Additionally, the performance of the U-Net for lesion segmentation from larger regions of interest, i.e., the entire breast MRI or a segmented breast region, should be evaluated; U-Nets trained with attention gating could be of particular benefit for those tasks.

CHAPTER 3

COMPUTERIZED ASSESSMENT OF BACKGROUND PARENCHYMAL ENHANCEMENT

3.1 Background Parenchymal Enhancement

Background parenchymal enhancement (BPE) is qualitatively defined according to the Breast Imaging Reporting & Data System (BI-RADS©) as minimal, mild, moderate, or marked BPE based on the visually perceived volume and intensity of enhancement in normal fibroglandular breast tissue after contrast injection for dynamic contrast-enhanced (DCE) magnetic resonance imaging (MRI) (Figure 3.1). [21, 29, 30] BPE is a significant predictor of breast cancer risk, with greater BPE increasing the odds of developing cancer. [29, 30, 100–102]

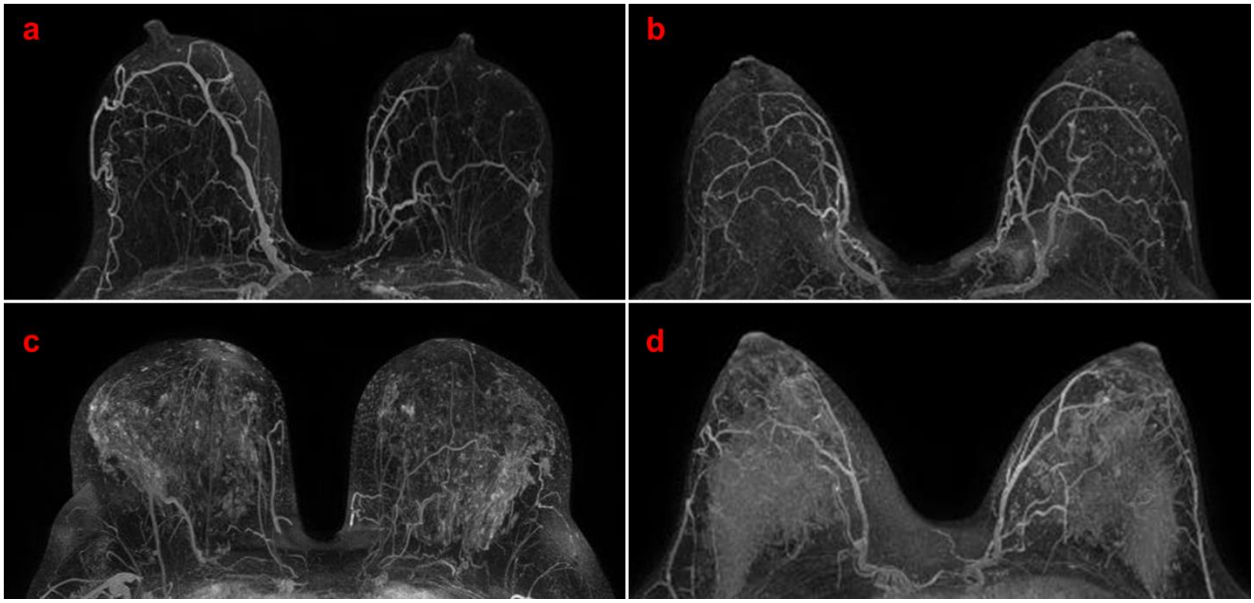

Figure 3.1: Examples of qualitative BPE assessments of minimal (a), mild (b), moderate (c), and marked (d) on second post-contrast subtraction maximum intensity projection images. Each image was acquired from a unique patient.

There is limited data assessing the physiology of BPE directly, so the exact biological mechanisms of it are still unknown. In terms of the distribution of the enhancement through the breast over the course of the dynamic contrast series, the enhancement often occurs initially at the

periphery of the fibroglandular tissue, due to the pattern of blood inflow from the internal and lateral thoracic arteries, which then feed into the retroareolar region, which enhances last. [28, 29] Normal fibroglandular tissue tends to exhibit a slow early and persistent delayed uptake of contrast, although in some cases of moderate or marked BPE, there is a rapid early contrast uptake. [28] A few studies measuring growth factors, microvessel density, and glandular components of the breast, suggest that BPE becomes elevated when there are greater concentrations of fibroglandular tissue that has greater vascularity, higher metabolic activity, and is sensitive to hormone level variations. [29]

Beyond ranging from minimal to marked in the overall degree, BPE can present with a bilateral, diffuse, symmetric, or asymmetric distribution; in some cases, it may present with a stippled, regional, or focal distribution (Figure 3.2). [28] Regardless of distribution, benign tissues typically present with more symmetric or bilateral areas of enhancement, whereas asymmetric enhancement may be due to malignancy or a response to breast cancer treatment. [28] In recent years, the use of 3.0T instead of 1.5T MRI has increased, which has provided improved contrast resolution, spatial resolution, and signal-to-noise ratio; this has improved image quality and characterization of BPE morphology. [26] However, it should be noted that in addition to the physiological factors already discussed, BPE appearance may also be influenced by MRI magnet strength due to its inherent dependence on T1-weighted relaxation times. [28] In many cases, tumor volumes can cause an overestimation of BPE by radiologists; the increased intensity of the tumor enhancement due to angiogenesis can inflate visual assessment of the BPE. Also, in cases with marked BPE, it can become difficult to differentiate between tumor and normal fibroglandular tissue, thus reducing sensitivity in breast cancer screening. [103] These effects have contributed to

the intra-observer variability in clinical BPE assessment that has been reported, thus necessitating a quantitative method for quantifying BPE. [101]

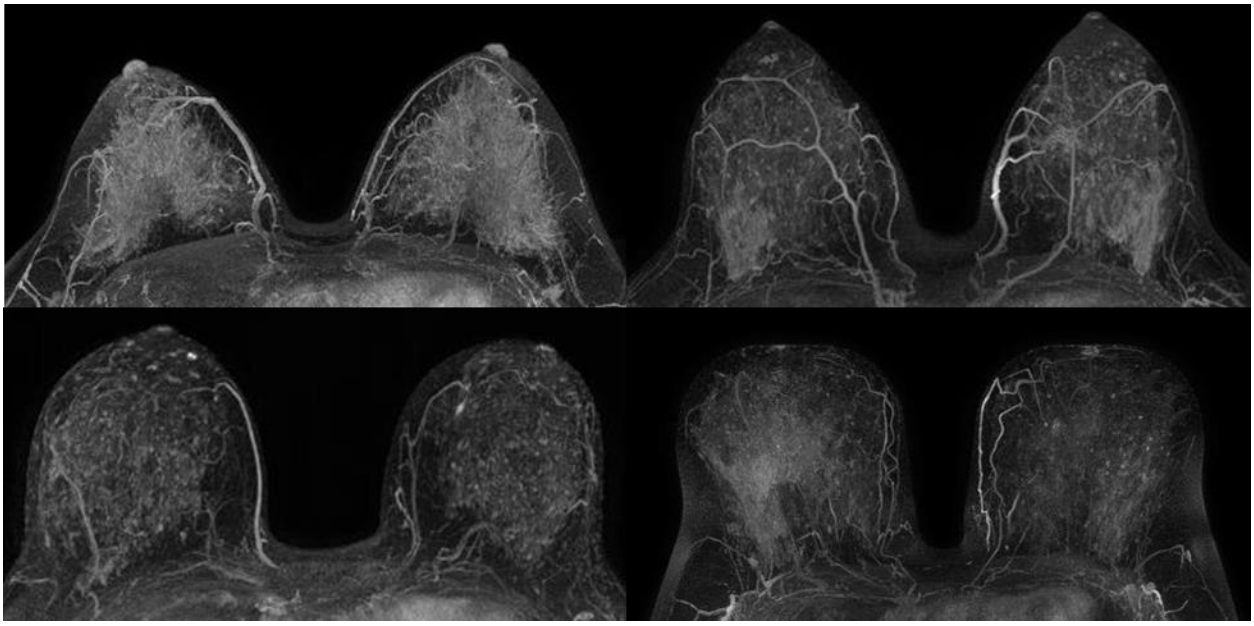

Figure 3.2: Examples of various distributions patterns and appearances of background parenchymal enhancement (BPE) in four unique patients. Radiologist reports for each case noted moderate BPE bilaterally.

BPE may be able to indicate physiologically active breast tissue and therefore serve as a biomarker for tissue prone to malignant transformation. [30, 102] A number of groups have developed quantitative measures for BPE, but a general consensus of the most useful value has yet to be reached. Human-engineered and deep-learned features for BPE have been calculated from both single MRI slices and MRI volumes, and while some incorporate FGT or breast segmentations, others rely on the entire image. [100, 104–106] One study that was based on a semi-automated segmentation algorithm achieved strong performance in distinguishing women who did and did-not develop breast cancer using a quantitative BPE value (AUC = 0.85, n = 95). [106] Additionally, another study found that the complexity of the BPE assessment caused only weak correlations between the investigators' quantitative values and the associated clinical ratings

(Spearman's $\rho = 0.65$, $n = 40$). [89] These studies demonstrate that further investigation is needed to develop a fully automated method for quantifying BPE.

Therefore, we have developed an automated method to segment breasts and electronically remove the influence of lesion enhancement on a computer BPE score. Our method was designed to mimic radiologist assessment of BPE, and it offers a more robust estimation of BPE levels from breast DCE-MR projection images. We investigated the performance of computer BPE scores from second post-contrast subtraction maximum intensity projections (MIPs) of both breasts, the affected breast, and the contralateral breast images created before and after the electronic removal of lesions. Additionally, we investigated the effect of various image parameters on the performance of computer BPE scores calculated from original and rescaled versions of maximum- or average-intensity projections (AIPs) of first- or second-post contrast subtraction DCE-MRI volumes. We speculated the transition from 1.5T to 3.0T MRI may influence radiologist perception of BPE and therefore the performance of machine-learning methods for BPE quantification, so this chapter also includes an additional evaluation of the influence of magnet strength on the performance of our BPE scoring method and its influence on radiologist perception of BPE.

3.2 Dataset

Out of the full diagnostic MRI dataset introduced in Section 2.3, 426 conventional breast DCE-MR exams (from 399 patients aged 23-89 years) were from patients that were diagnosed with a single lesion (Figure 1.5, Table 3.1). A subset of 76 exams (6 minimal, 18 mild, 26 moderate, 11 marked, and 15 unknown BPE) from 73 patients were previously used in developing the breast segmentation methods (refer to Section 2.7). The remaining exams from 350 exams from 326 patients with known radiologist BPE ratings (99 minimal, 159 mild, 78 moderate, and 14 marked

BPE), were used for independent testing of the proposed BPE AI algorithm. Radiologist BPE ratings were acquired from the radiologist reports available from prior clinical review. For each exam, the breast containing the diagnosed lesion is termed the "affected" breast, and the contralateral breast is termed the "unaffected" breast.

Table 3.1: Prevalence of radiologist BPE ratings contained within the dataset of 426 DCE-MR exams from 399 patients. All exams from a given patient were in either the training set (refer to Section 2.7) or the test set.

(No. of exams)	**Minimal**	**Mild**	**Moderate**	**Marked**	**Unknown**	**Total**
Training Set	6	18	26	11	15	76
Test Set	99	159	78	14	0	350
Total	105	177	104	25	15	426

3.3 Electronic Lesion Removal

A well-established, in-house, automated 3D fuzzy c-means (FCM) clustering approach was used to segment the lesions from the DCE-MR volumes. [91] The lesion sizes, approximated by the square root of the lesion area at the center lesion slice, ranged between two and 65 mm. To electronically remove the lesions, the lesion area defined by the FCM segmentation was replaced with a value equivalent to the average intensity of the pixels bordering the lesion segmentation on the second post-contrast subtraction image slice. This process was repeated on each slice that passed through the lesion before projecting the maximum pixel values from all available volume slices to produce a new MIP that excluded the influence of the lesion. The breast masks generated from the U-Net outputs were used to retain only the pixels belonging to both breasts, the affected breast, and the unaffected breast on the second post-contrast subtraction MIP with the lesion removed (Figure 3.3). For comparison across input image parameters, this method was also

conducted using first post-contrast subtraction images and also using average-intensity projections to produce images of the affected breast without the influence of the lesion.

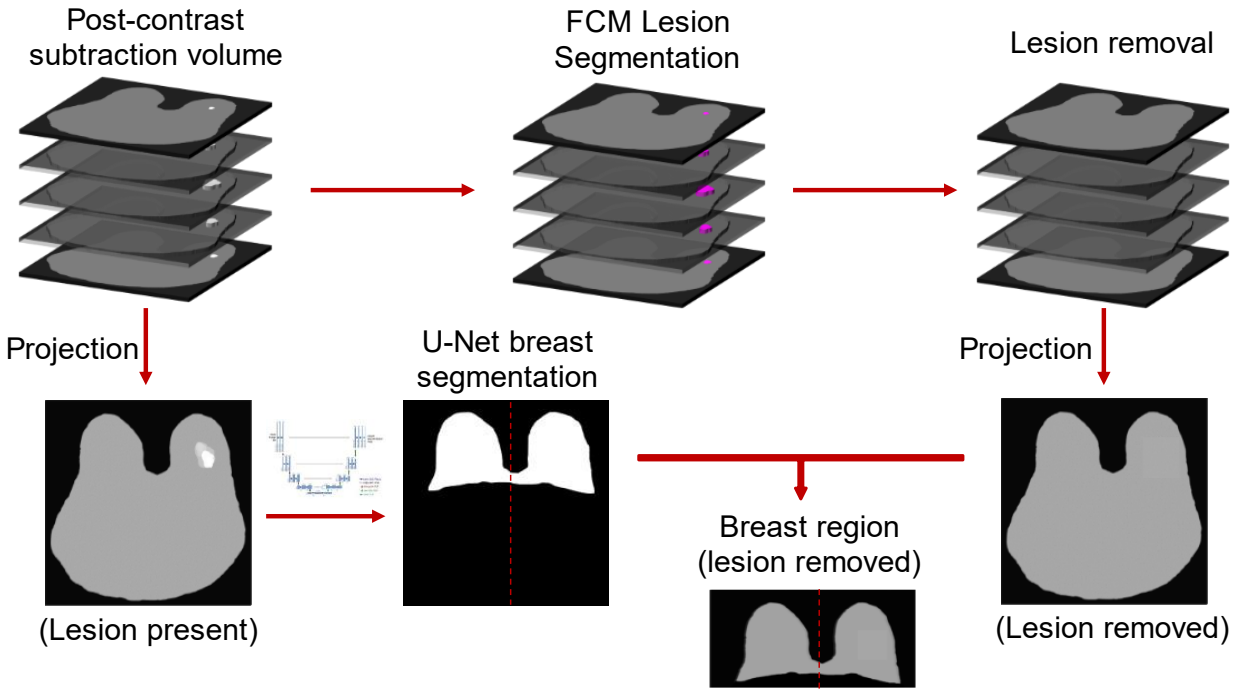

Figure 3.3: Flowchart of the method for electronic lesion removal, image projection, and breast segmentation from a post-contrast subtraction breast DCE-MRI. Lesion and breast segmentations were performed using fuzzy c-means (FCM) clustering and U-Net CNN, respectively. The breast segmentation was vertically split at center to select the affected breast region from the projection image excluding the lesion (Refer to Section 2.7). Computer BPE scores were calculated in a separate rescaled MIP after implementation of our digital electronic lesion removal algorithm.

3.4 Computer BPE Score

For each of the defined breast regions (affected, unaffected, and both), the computer BPE scores were automatically calculated from the second post-contrast subtraction MIPs. Within each MIP, the pixel values were linearly rescaled so that the original pixel values ranging from 0 to 255 were scaled to a range of 0 to 1. To reflect the qualitative definitions of BPE assigned by radiologists based on the amount and intensity of the enhancement in FGT, the average pixel intensity of the pixels contained within each breast served as the computer BPE score (Figure 3.4).

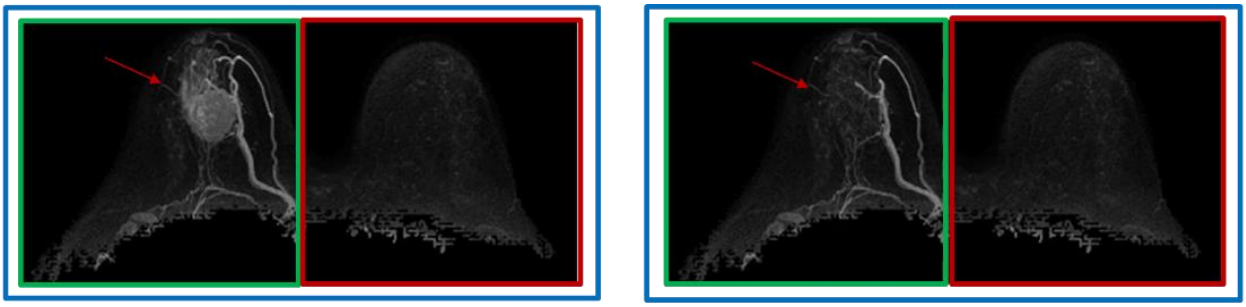

Figure 3.4: Computer BPE scores have been calculated from the affected breast (green box), unaffected breast (red box), and both breasts (blue box), before and after lesion (arrow) removal.

3.5 Evaluation of Computer BPE Score

To determine the strength and direction of the correlation of the computer BPE scores with radiologist BPE ratings, Kendall's tau-b was used in rank correlation with a t-test used to assess the statistical significance of the correlation. [107] To assess how lesion removal changes the computer BPE scores, the ratio of the computer BPE score calculated after lesion removal to the computer BPE score calculated before lesion removal was examined according to the lesion size for the second post-contrast subtraction MIP of each affected breast. To determine the predictive value of the computer-extracted BPE scores for BPE level classification tasks, receiver operating characteristic (ROC) analysis was performed using the proper binormal model. [108] Clinical radiologist BPE ratings were the only truth available for BPE assessment, so the performance of the computer BPE scores was compared to random guessing. ROC analysis was performed using computer BPE scores for binary classification of minimal versus marked BPE; it was also evaluated for binary classification of low (minimal, mild) versus high (moderate, marked) BPE (Figure 3.5). The statistical significance of the area under the ROC curve (AUC) relative to random guessing was determined using the z-test with Bonferroni corrections for multiple comparisons. [98]

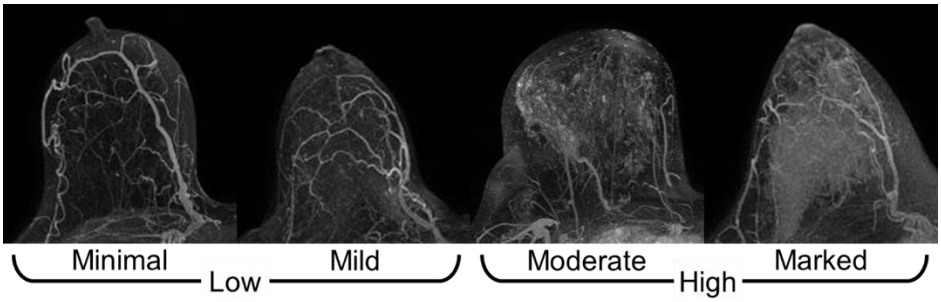

Figure 3.5: Clinical radiologist BPE ratings were used as the reference standard for receiver operating characteristic (ROC) analysis. ROC analysis was performed to determine the predictive value of computer BPE scores for binary classification of minimal vs. marked BPE and of low (minimal, mild) vs. high (moderate, marked) BPE.

Rank correlation and ROC analysis were also used to understand the effect of different image parameters on the calculated BPE. The minimal versus marked BPE and low versus high BPE tasks were thus evaluated for computer BPE scores calculated from the affected breast in each of the following image types (shown in Figure 3.6):

(i) First- or Second- post-contrast subtraction images
(ii) Maximum- or Average- Intensity Projections
(iii) Original or Rescaled pixel values

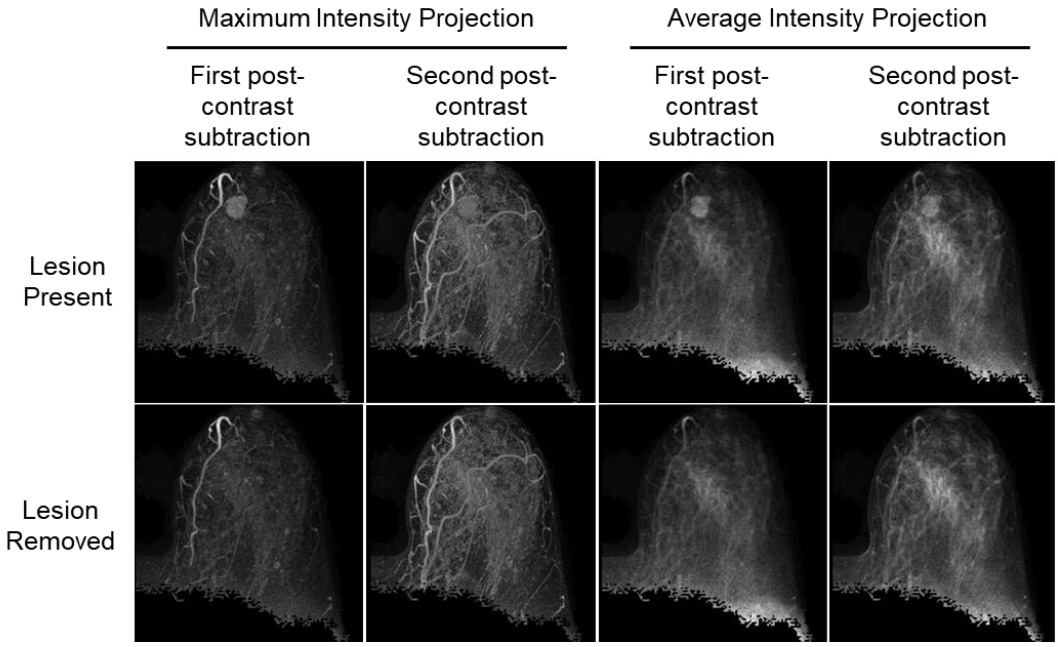

Figure 3.6: Example images of an affected breast from a case classified as Marked BPE by a radiologist. The computer BPE scores were calculated from the affected breast region in the post-contrast subtraction projection images after electronic lesion removal (bottom row).

3.6 Results

As mentioned in Section 3.2, the initial definition of our computer BPE score was calculated from the second post-contrast subtraction MIPs to reflect the qualitative assessment of BPE performed by radiologists. Prior to the results comparing the effect of imaging parameters on the computer BPE scores (Table 3.3), we evaluated the effect of breast region on the computer BPE scores calculated from the second post-contrast subtraction MIPs (Table 3.2). On the test set of 350 second post-contrast subtraction MIPs, a statistically significant positive correlation was found between the computer BPE scores and radiologist BPE ratings for all breast regions, before and after the lesion removal (Figure 3.7).

The ratio of the scores calculated after versus before lesion removal, sorted by size and BPE level, are shown with example cases of affected breasts (Figure 3.8). As would be expected, the computer BPE scores were reduced after the lesion removal; this was more pronounced for larger lesions and cases with low BPE levels.

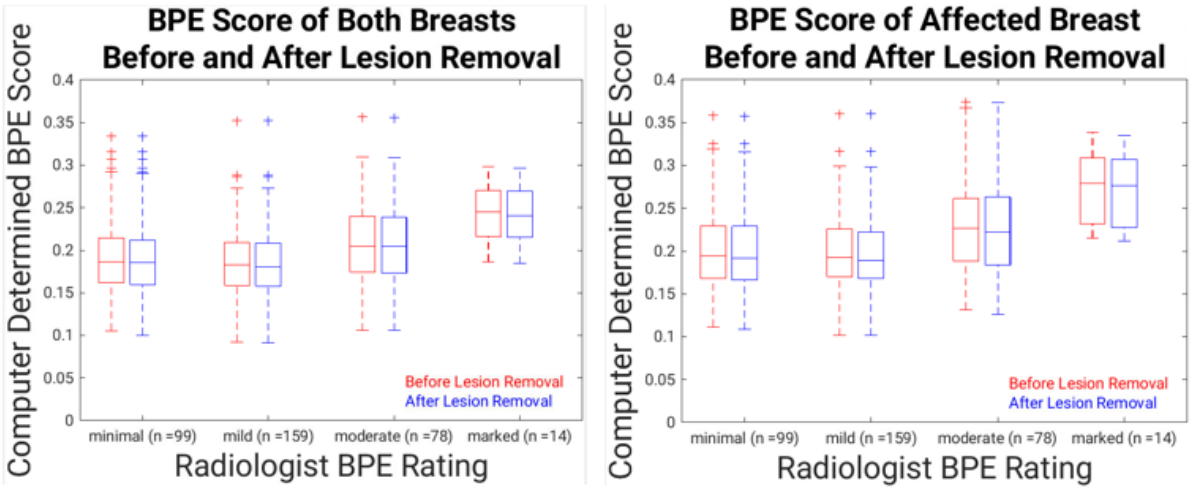

Figure 3.7: Positive correlation between all computer BPE scores (second post-contrast subtraction MIP) and the radiologist BPE ratings were statistically significant ($p < 0.001$). Computer BPE scores from unaffected breasts are not shown since there was no lesion to be removed.

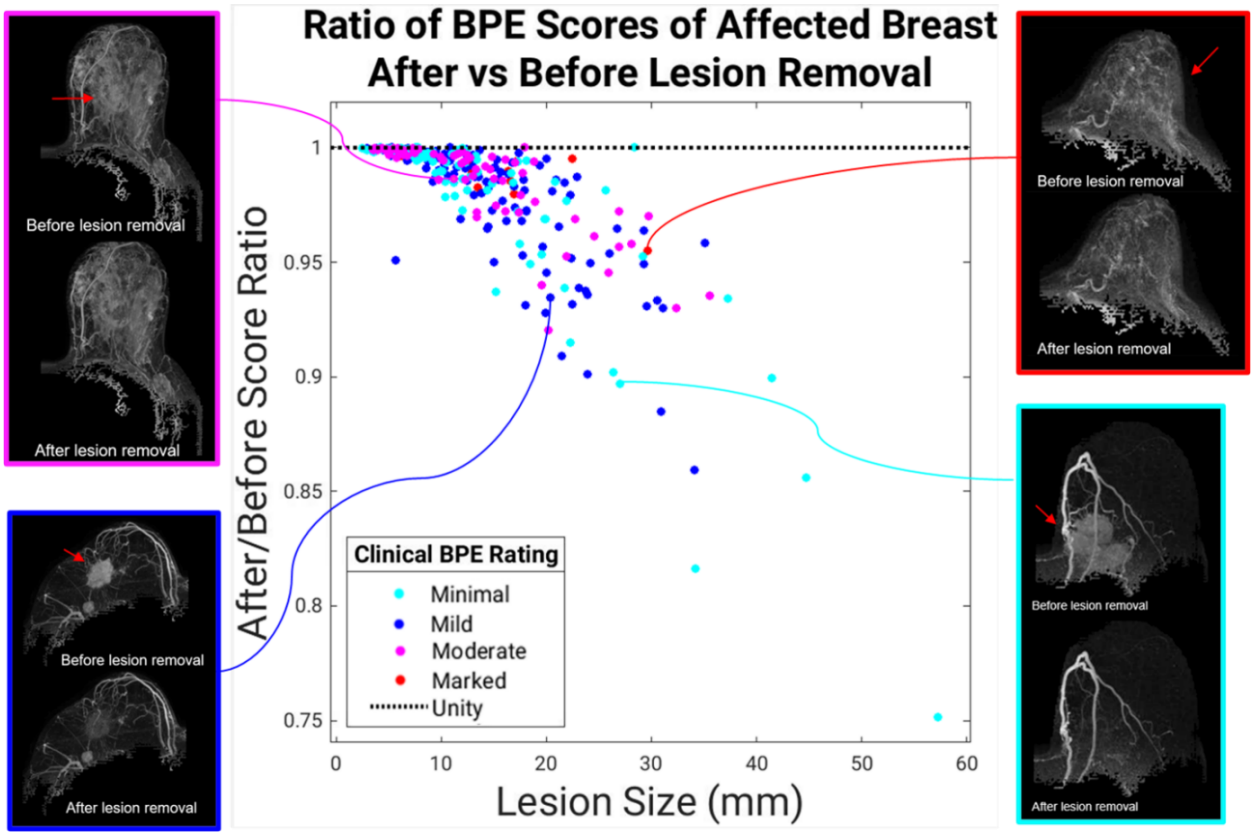

Figure 3.8: The ratio of the computer BPE scores (second post-contrast subtraction MIP) calculated after lesion removal to the score calculated before lesion removal for the affected breast, shown versus lesion size (n = 350). Results demonstrate the importance of lesion removal to avoid inflation of computer BPE estimations, especially in cases containing large lesions and low BPE levels.

The AUCs for the task of classifying minimal versus marked BPE and for the task of classifying low (minimal, mild) versus high (moderate, marked) BPE according to radiologist BPE ratings were calculated for each of the breast regions in the second post-contrast subtraction MIPs (Table 3.2). All classification tasks performed statistically significantly greater than random guessing (z-test). For all breast regions, the computer BPE scores yielded greater AUC results for minimal versus marked BPE than for low versus high BPE levels, which was expected because it is easier to distinguish between the two extreme BPE levels than the intermediate ones. The computer BPE scores from the affected breast, both before and after lesion removal, yielded greater

AUC results than the computer BPE scores from the unaffected breast for both classification tasks, thus the computer BPE scores from that region were used in subsequent evaluations.

Table 3.2: Effect of breast region used for computer BPE score. AUC results from ROC analysis for the task of BPE level classification using computer BPE scores calculated from the rescaled second post-contrast subtraction maximum-intensity projection (MIP). High BPE includes marked or moderate BPE, and low BPE includes mild or minimal BPE. Raw, uncorrected p-vales from the z-test are reported in the table. Asterisks indicate performance statistically significantly greater than random guessing. Statistical significance of the AUCs was assessed using the Bonferroni correction for thirteen comparisons. The bolded selection was used in subsequent analyses.

	Minimal (n = 99) vs. Marked (n = 14) BPE AUC	Low (n = 258) vs. High (n = 92) BPE AUC
Both breasts	0.84 ± 0.04 (p = 9.21e-15) *	0.66 ± 0.03 (p = 7.94e-07) *
Both breasts, removed lesion	0.83 ± 0.04 (p = 4.76e-14) *	0.66 ± 0.03 (p = 5.39e-07) *
Affected breast	0.86 ± 0.03 (p = 2.69e-26) *	0.68 ± 0.03 (p = 3.92e-08) *
Affected breast, removed lesion	**0.87 ± 0.04 (p = 1.31e-21) ***	**0.68 ± 0.03 (p = 1.43e-08) ***
Unaffected breast	0.79 ± 0.05 (p = 8.83e-08) *	0.66 ± 0.03 (p = 6.82e-07) *

The results of the comparisons between computer BPE scores calculated from varying image types are shown in Table 3.3 and Figure 3.9 (affected breast scores only). Statistically significant positive correlations were found between the radiologist BPE ratings and the computer-extracted BPE scores from the rescaled images, except for the first post-contrast subtraction AIPs. Computer BPE scores performed statistically significantly greater than random guessing in minimal versus marked BPE level classification, except for the first post-contrast subtraction AIPs. Computer BPE scores performed statistically significantly greater than random guessing in low versus high BPE level classification, except for the original MIPs and the original first post-contrast subtraction AIPs. For all image types, the computer BPE scores yielded greater AUC results for minimal versus marked BPE than for low versus high BPE levels. Computer BPE scores from rescaled images yielded greater AUC results than computer BPE scores from original images

in both BPE level classification tasks. ROC curves showed computer BPE scores from second post-contrast projections yielded greater AUC results than computer BPE scores from first post-contrast projections, and computer BPE scores from MIPs yielded greater AUC results than computer BPE scores from AIPs (Figure 3.9). The computer BPE scores of the rescaled second post-contrast MIPs statistically significantly outperformed other rescaled image types for minimal versus marked BPE classification ($p < 0.05$, z-test).

Table 3.3: Effect of breast imaging parameters used for the computer BPE score. Results from Kendall's rank correlation and ROC analysis for computer BPE scores calculated from the affected breast region. High BPE includes marked or moderate BPE, and low BPE includes mild or minimal BPE. Raw, uncorrected p-vales from the t-test or z-test are reported in the table. Asterisks on tau-b values indicate a statistically significant correlation between the computer BPE scores and radiologist BPE ratings. Asterisks on AUC results indicate performance statistically significantly greater than random guessing. Statistical significance was assessed using the Bonferroni correction for thirteen comparisons.

Quantitative Value	Projection Image Type	Post-Contrast Subtraction Timepoint	Kendall's rank correlation tau-b (n = 350)	Minimal (n = 99) vs. Marked (n = 14) BPE AUC	Low (n = 258) vs. High (n = 92) BPE AUC
Mean pixel intensity of original image	Maximum	First	0.043 (p = 0.299)	0.69 ± 0.06 (p = 1.42e-3)*	0.58 ± 0.03 (p = 0.016)
		Second	0.075 (p = 0.067)	0.78 ± 0.06 (p = 9.51e-7)*	0.58 ± 0.03 (p = 0.013)
	Average	First	0.083 (p = 0.043)	0.79 ± 0.05 (p = 5.27e-10)*	0.60 ± 0.03 (p = 5.32e-3)
		Second	0.090 (p = 0.030)	0.74 ± 0.05 (p = 3.65e-6)*	0.60 ± 0.03 (p = 2.66e-3)*
Mean pixel intensity of rescaled image	Maximum	First	0.132 (p = 1.32e-3)*	0.78 ± 0.06 (p = 3.88e-7)*	0.63 ± 0.03 (p = 1.29e-4)*
		Second	0.186 (p = 5.64e-6)*	0.87 ± 0.04 (p = 1.31e-21)*	0.68 ± 0.03 (p = 1.44e-8)*
	Average	First	0.119 (p = 3.86e-3)	0.69 ± 0.07 (p = 7.85e-3)	0.61 ± 0.03 (p = 1.18e-3)*
		Second	0.160 (p = 9.46e-5)*	0.77 ± 0.06 (p = 1.84e-6)*	0.63 ± 0.03 (p = 5.55e-5)*

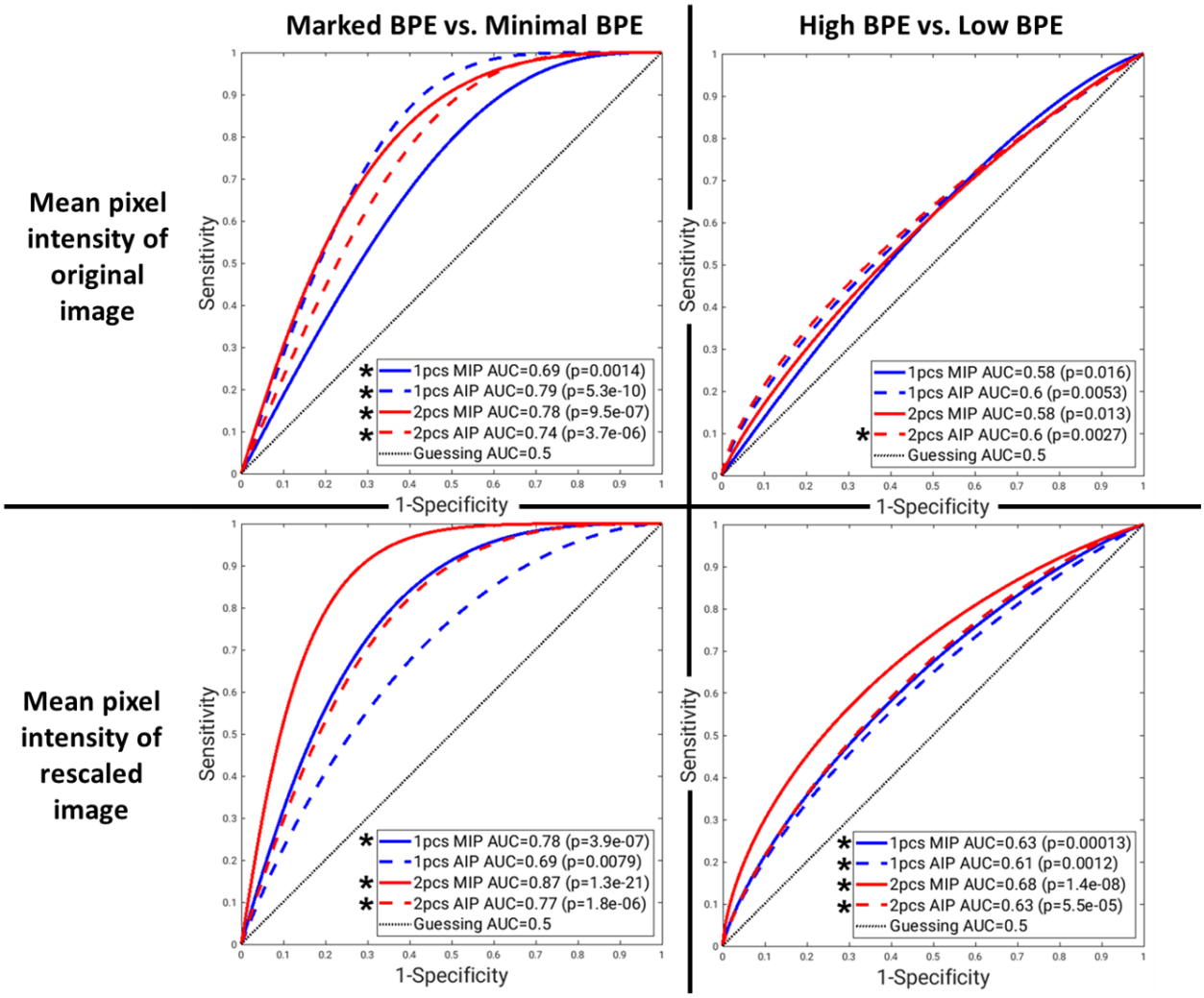

Figure 3.9: ROC curves for the binary classification tasks of marked BPE (n = 14) vs. minimal BPE (n = 99) (left) and high (marked or moderate) BPE (n = 92) vs. low (mild or minimal) BPE (n = 258) (right) using the mean pixel intensity of the original (top) and rescaled (bottom) image of the affected breast. Asterisks indicate classification performance statistically significantly greater than random guessing. Raw, uncorrected p-values are reported from the z-test; statistical significance for performance greater than random guessing was assessed after a Bonferroni correction for thirteen comparisons. 1pcs, 2pcs: first- and second-post-contrast subtraction; MIP, AIP: maximum- and average-intensity projection.

3.7 Additional Evaluation of the Influence of Magnet Strength on BPE Assessment

After completion of the previous analyses to determine that the computer BPE scores calculated from the rescaled second post-contrast subtraction MIP of the affected breast were the most suitable image type to use when predicting BPE levels (Table 3.3), we evaluated the potential

effects that magnet strength may have on the computer BPE scores (second post-contrast subtraction MIP only). In addition, we investigated the potential influence of magnet strength on the qualitative BPE assessment performed by radiologists. Using the computer BPE scores from our dataset of 350 exams (255 1.5T, 95 3.0T), the difference in computer BPE scores across magnet strengths was evaluated using the t-test. To determine the strength and direction of the correlation between computer BPE scores and radiologist BPE ratings, Kendall's tau-b was used in rank correlation with a t-test used to assess the statistical significance of the correlation. [107] ROC analysis was performed to determine the predictive value of computer BPE scores calculated from 1.5T only, 3.0T only, or all exams for the binary tasks of classifying high (marked or moderate) BPE versus low (minimal or mild) BPE and for classifying marked BPE versus minimal BPE relative to a radiologist rating. [108] The statistical significance of the area under the ROC curve (AUC) relative to random guessing was determined using the z-test with Bonferroni corrections for multiple comparisons. [98] Further, the use of each magnet strength over time and the prevalence of radiologist BPE ratings assigned for 1.5T or 3.0T images were explored.

The distribution of computer BPE scores for each magnet strength are shown in Figure 3.10. The general range and shapes of the computer BPE score distributions appeared similar between 1.5T and 3.0T, and the results of the t-test failed to show a statistically significant difference between the computer BPE scores from the 1.5T and 3.0T images ($p = 0.05$). For each magnet strength, the computer BPE scores had a statistically significant positive correlation with the radiologist BPE ratings across the four BPE levels ($p < 0.001$, t-test). Based on the boxplots in Figure 3.10 and Figure 3.11, it appears that the computer scoring may depend on whether BPE scores were calculated from 1.5T or 3.0T images. However, the results of the t-tests used to assess

differences between the computer BPE scores from the 1.5T and 3.0T images failed to show a statistically significant difference at each radiologist BPE rating ($p > 0.05$, t-tests).

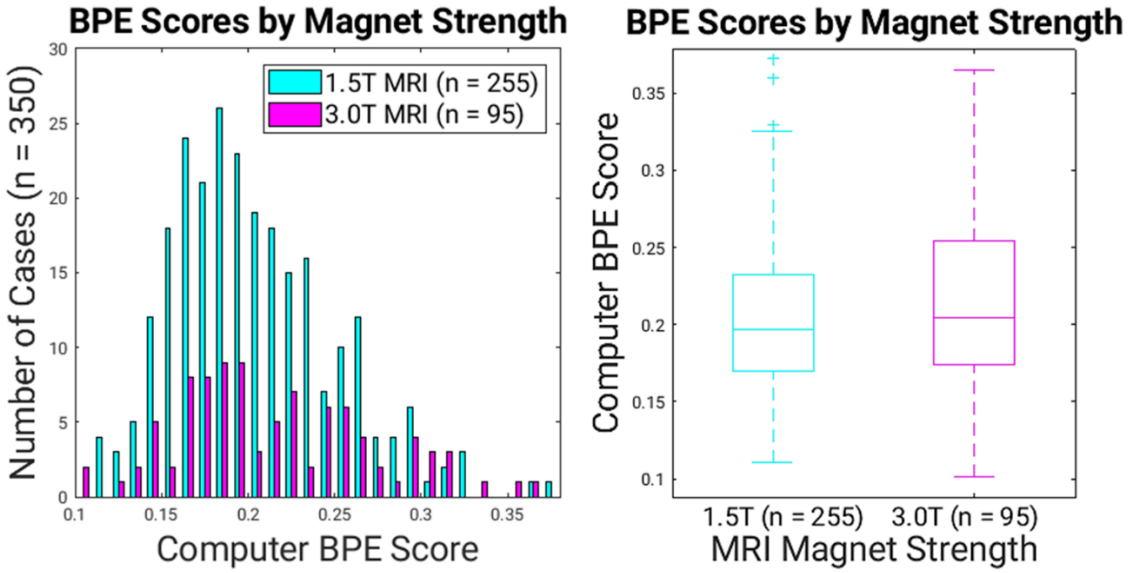

Figure 3.10: Histogram (left) and boxplot (right) of computer BPE scores (second post-contrast subtraction MIP) for each magnet strength. Results of the t-test failed to show a statistically significant difference between the computer BPE scores from the 1.5T and 3.0T images ($p = 0.05$).

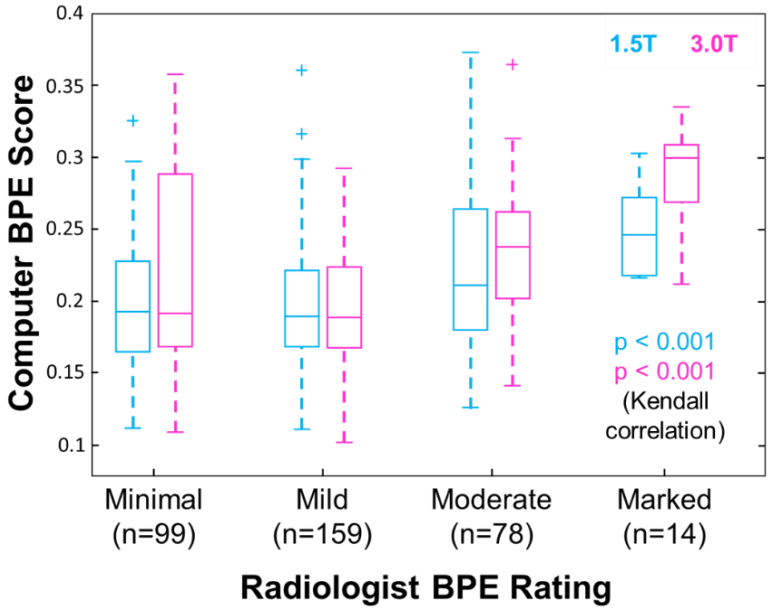

Figure 3.11: Kendall's tau-b results showed statistically significant positive correlations between the computer BPE scores (second post-contrast subtraction MIP) and the radiologist BPE ratings of 1.5T and 3.0T images ($p < 0.001$, t-test). Results of t-tests failed to show a statistically significant difference between computer BPE scores of 1.5T and 3.0T images at each radiologist BPE rating ($p > 0.05$).

Although we failed to show statistically significant differences between the computer BPE scores resulting from 1.5T or 3.0T images, we performed ROC analysis to see how the magnet strength might affect the BPE level classification performance of the scores. The computer BPE scores performed statistically significantly greater than random guessing in both classification tasks: high (marked or moderate) BPE versus low (minimal or mild) BPE and marked BPE versus minimal BPE (Figure 3.12). As would be expected due to the similarity between the intermediate BPE levels (mild and moderate), the computer BPE scores yielded greater AUC results for marked versus minimal BPE than for high versus low BPE classification tasks. In the high versus low BPE task, the computer BPE scores from 3.0T cases performed statistically significantly greater than the computer BPE scores from the combined 1.5T and 3.0T cases ($p < 0.05$, z-tests). The marked versus minimal BPE task failed to show a statistically significant difference in classification performance between magnet strengths ($p > 0.05$, z-tests).

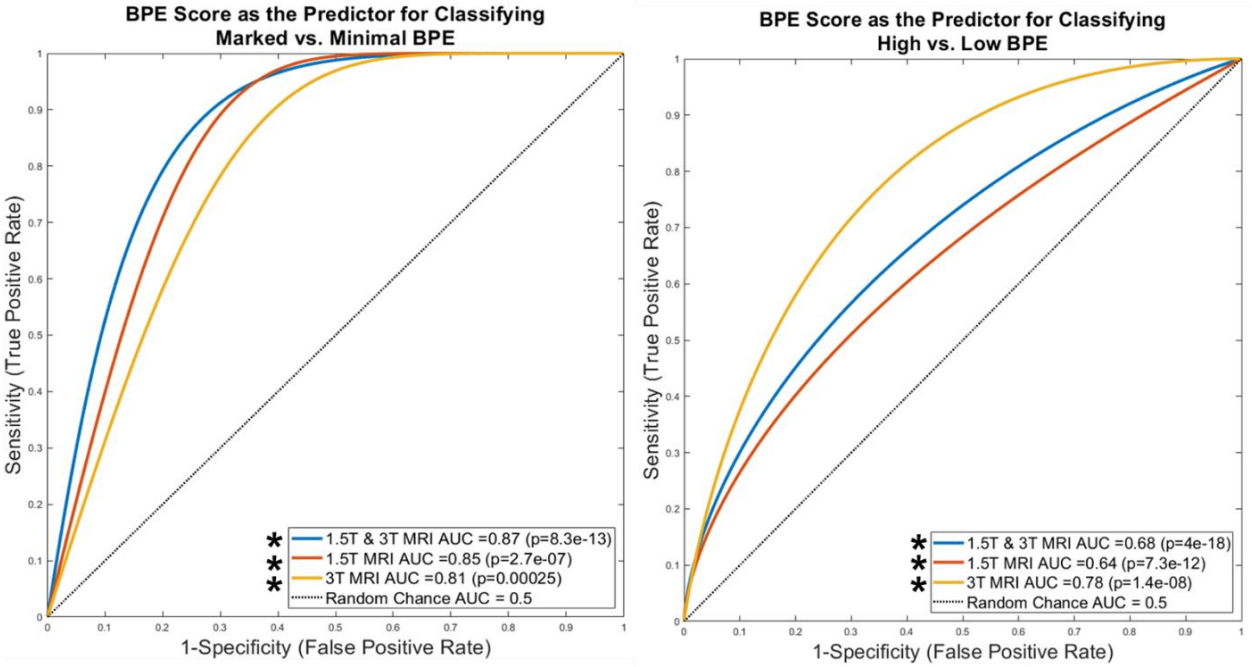

Figure 3.12: ROC curves showing the performance of the computer BPE scores (second post-contrast subtraction MIP) in the binary classification tasks of marked BPE vs. minimal BPE and

Figure 3.12 (continued): high (marked or moderate) BPE vs. low (mild or minimal) BPE. Raw, uncorrected p-values are reported from the z-test. Asterisks indicate classification performance statistically significantly greater than random guessing after multiple comparisons corrections. Computer BPE scores calculated on 3.0T cases performed statistically significantly greater compared to 1.5T cases in high vs. low BPE ($p < 0.05$, z-test).

The BPE level classification performance was determined relative to the radiologist BPE ratings assigned during the initial clinical review, so we also investigated the potential effects of magnet strength on our reference standard by looking at the prevalence of each radiologist BPE rating assigned to 1.5T or 3.0T images. Figure 3.13 shows the prevalence of each magnet strength used for patient scans over the period of our dataset. Figure 3.14 demonstrates a relatively increased occurrence of above minimal BPE ratings for 3.0T than for 1.5T cases, potentially indicating that radiologists perceived stronger BPE levels in 3.0T MRI compared with 1.5T MRI. We speculate that it is possible that 3.0T cases evaluated by radiologists familiar with 1.5T enhancement appearance may have overestimated BPE ratings as the use of 3.0T scanners increased. Likewise, that 1.5T cases acquired more recently could be underestimated by radiologists familiar with 3.0T enhancement appearance. Although further investigation is needed, we note that it is possible that radiologist BPE ratings may depend on magnet strength.

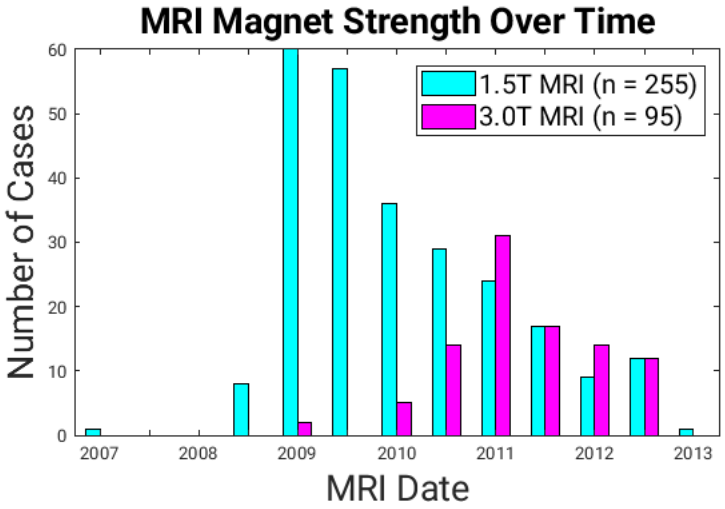

Figure 3.13: Histogram of the number of cases acquired using either 1.5T or 3.0T DCE-MRI over the time period of the diagnostic MRI dataset (n = 350 cases).

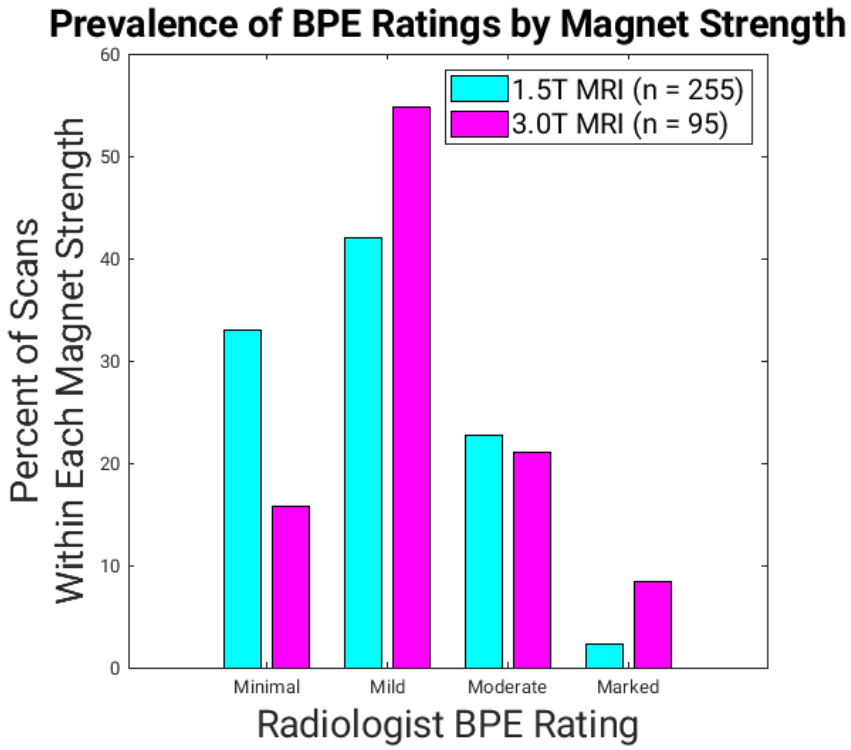

Figure 3.14: Prevalence of each radiologist BPE rating in the dataset of 350 diagnostic MRIs, split by magnet strength. Note the greater prevalence of above minimal ratings in the 3.0T cases than the 1.5T cases.

3.8 Discussion

In current clinical settings, radiologist BPE ratings are subjectively assigned based on the relative volume and intensity of enhancement in normal fibroglandular breast tissue after contrast injection for DCE-MRI. This chapter presented an automated computer algorithm for assessment of BPE and investigated the effect of using various breast DCE MR image types. The results of this work demonstrated the promising performance of an automatic BPE scoring method, which yields computer BPE scores in classifying marked versus minimal BPE across various image viewing projections and DCE timepoints. The method of calculating BPE scores from breast DCE-MR MIP images was not influenced by the contrast enhancement within lesions, which currently causes intra-observer variability in clinical BPE level assessment, because the algorithm includes an electronic removal of the lesion.

The automatically calculated computer BPE scores from all breast regions had a statistically significant correlation with the radiologist BPE ratings, with the exception of one image type; thus, the computer BPE scores had a positive correlation with increasing BPE. The ratio of the computer BPE scores calculated after to before lesion removal demonstrated the importance of electronic lesion removal to avoid inflation of BPE estimations, especially in cases containing large lesions and low BPE levels. Although the computer BPE scores from the second post-contrast subtraction MIPs of the affected and unaffected breasts appeared similar in boxplots, computer BPE scores of the affected breast yielded greater AUC results than those of unaffected breast in the prediction of radiologist BPE ratings. Based on the computer BPE scores from all breast regions, the classification of minimal versus marked BPE yielded greater AUC results than the classification of low versus high BPE, which was expected because it is easier to distinguish between the two extreme BPE levels than the intermediate ones.

The electronic removal of the lesion from the affected breast improved the predictions for the minimal versus marked task, but not for the low versus high task; this may be due to the complexity of the BPE levels considered in each task. Given that the removal of lesions had the greatest impact on reducing the computer BPE scores for minimal BPE cases, the lesion removal would improve the classification of minimal versus marked BPE. In the low versus high task, however, the large prevalence of mild and moderate BPE cases contribute to the difficulty of the task due to the similarity between intermediate BPE levels that exists even after lesion removal. Additionally, the AUC results for computer BPE scores calculated from various image projections and post-contrast subtraction timepoints demonstrated the flexibility of the algorithm in BPE level classification tasks. Comparisons between the original and rescaled versions of the maximum- and

average-intensity projections (MIP and AIP) created from the first or second post-contrast subtraction images of the affected breast demonstrated that the computer BPE scores calculated from the rescaled, second post-contrast subtraction MIP yielded the greatest overall AUC results. Therefore, of the scores evaluated in this study, the best computer-generated representation of the relative intensity and volume of enhancement qualitatively assessed by radiologists were the computer BPE scores of the rescaled, second post-contrast subtraction MIP.

Results of our additional investigations suggested radiologist assessment of BPE may not be consistent due to differences in perceived enhancement in 1.5T and 3.0T breast DCE-MR images. We failed to show statistically significant differences in computer BPE scores computed from 1.5T or 3.0T images for all radiologist BPE ratings, although we observed improved high versus low BPE classification performance from the 3.0T scores compared to 1.5T or combined scores. While the performance of our method for computing BPE scores is promising, further assessment is needed to ensure the robustness of a computer method for calculating BPE scores against variations in magnet strength.

3.9 Limitations and Future Work

A limitation in our method for automatic BPE assessment was based on the segmentation of breast regions; future investigations could assess the variability in computer BPE scores based on the precision of the masks that define breast regions. In this work, the computer BPE scores were calculated from MIPs that often contained major vasculature, which contain bright pixels that may inflate the computer estimation of BPE. Future investigations should aim to remove the influence of the vasculature's enhancement, as we have already done with lesions, in order to produce a more accurate representation of the fibroglandular tissue enhancement. The only truth we had

available to assess the performance of our computer BPE scores for BPE level classification tasks were the radiologist BPE ratings assigned during initial clinical review; thus, our ROC analyses were limited to comparisons against random guessing performance. Additionally, a single case-level radiologist BPE rating was used as truth for the computer BPE scores calculated from individual breasts, so the potential influence of BPE asymmetry on the performance of our BPE scoring method could not be assessed. Another limitation of this study was that, although breast MRIs are typically acquired during the second week of the menstrual cycle, we lack data regarding the timing of the acquisition relative to the patients' cycles. Thus, we could not evaluate the reported influence hormonal status may have on T1 relaxation times and signal intensity. [109] Further investigation of variability in the reference standards used for algorithm development may improve the overall performance of our method in BPE classification tasks. Future assessment of a screening dataset, including 1.5T and 3.0T data matched across patients, could also increase our understanding of BPE interpretation. Additionally, the techniques developed in this chapter could also be translated to abbreviated or ultrafast dynamic contrast enhancement sequences; the electronic lesion removal may allow for BPE assessment during the early phase of contrast uptake. The robustness of the computer BPE algorithm should also be assessed across datasets that include images acquired from multiple institutions and MRI manufacturers.

CHAPTER 4

BPE SCORING ON A HIGH-RISK SCREENING DATASET

4.1 Breast Cancer Risk Assessment

Estimation of a woman's lifetime risk for breast cancer beyond the general population risk is based on a series of known risk factors, such as hormonal and reproductive status, genetic mutation status (e.g., *BRCA1*, *BRCA2*, or *PALB2*), personal history of breast disease, radiation exposure, or family history of breast and ovarian cancers. [14–18] Although the thresholds may vary depending on which model for risk assessment is being used, an individual's lifetime risk for developing invasive breast cancer can be categorized into average- (<15%), intermediate- (15-19%), or high (>20%) risk. [110] Statistical models, such as Gail, Claus, BRCAPRO, and IBIS (Tyrer-Cuzick), have been established for clinical risk assessment; each model incorporates a selection of the known risk factors into their prediction of future breast cancer (Table 4.1). These types of models have been successful for use in the general population, however, while some models have been validated in a high-risk clinic, the accuracy of other models is limited for high-risk patients. [14, 111] Additionally, mammographic breast density has only recently been incorporated into some models to refine clinical risk assessment, and background parenchymal enhancement (BPE) has yet to be included, although they are both known risk factors for breast cancer. [18, 104, 112, 113]

Computer-extracted tumor features from breast magnetic resonance imaging (MRI) have played a role in classification tasks for diagnosing malignant versus benign lesions, and they have more recently been used as prognostic markers to distinguish between noninvasive and invasive lesions. [86, 88] Similar to the methods in which tumor features can be used as prognostic markers, other image-based biomarkers have the potential to be used to predict prognosis or be factored into

risk assessment models. BPE, for instance, may be able to indicate physiologically active breast tissue and therefore serve as a biomarker for tissue prone to malignant transformation. [29, 30, 102] Factoring in qualitative or quantitative values of BPE may further improve the predictive value of clinical risk models, which would be particularly useful for high-risk patients that undergo screening MRI. One experimental image-only model found that it was possible to identify image characteristics that are associated with long-term risk for a high-risk population and to improve risk predictions versus a traditional model. [104] Table 4.2 includes additional image-based studies in which radiologist BPE ratings and computer-generated BPE values have been used alone, or in combination with clinical patient characteristics, to predict breast cancer.

Table 4.1: Examples of breast cancer risk assessment models currently used clinically.

Model	**Clinical Features**
Gail [15]	*Personal Info*: Age *Hormonal/reproductive factors*: age at menarche, age at first live birth *Personal history*: breast biopsies, atypical ductal hyperplasia *Family history*: first-degree relatives with breast cancer
Claus [16]	*Personal Info*: Age *Hormonal/reproductive factors*: N/A *Personal history:* N/A *Family history*: first- or second-degree relatives with breast cancer, age of onset of breast cancer in relative
BRCAPRO [17]	*Personal Info*: Age *Hormonal/reproductive factors*: N/A *Personal history:* N/A *Family history*: first- or second-degree relatives with breast cancer, age of onset of breast cancer in relative, bilateral breast cancer in relative, ovarian cancer in a relative, male breast cancer

Table 4.1 (continued): Examples of breast cancer risk assessment models currently used clinically.

Model	Clinical Features
IBIS (Tyrer-Cuzick) [18]	*Personal Info*: Age, BMI *Hormonal/reproductive factors*: age at menarche, age at first live birth, age at menopause, hormone replacement therapy use *Personal history*: breast biopsy, atypical ductal hyperplasia, lobular carcinoma in situ *Family history*: first- or second-degree relatives with breast cancer, age of onset of breast cancer in relative, bilateral breast cancer in relative, ovarian cancer in a relative

Table 4.2: Examples of image-based breast cancer risk assessment studies that have incorporated image-based features and a selection of clinical patient characteristics.

Model Reference	Dataset	Features	Key Result
Grimm et. al. [114]	High risk screening MRI: 61 patients with cancer, 122 control patients	*Clinical*: N/A *Image-based*: BI-RADS BPE	Patients with greater than minimal BPE were two and a half times more likely to develop future cancer
Lee et. al. [115]	Surveillance MRI, prior primary cancer treated with surgery: 109 patients with second cancer, 2559 patients with no second cancer	*Clinical*: age, family history, BRCA status, initial cancer stage, hormone status, treatments, time between diagnosis/treatment/imaging, mammographic density *Image-based*: BI-RADS BPE	Greater than minimal BPE at surveillance MRI was associated with increased future second cancer risk
Lam et. al. [100]	High risk screening MRI: 23 patients developed cancer, 23 did not	*Clinical*: N/A *Image-based*: BI-RADS BPE, Quantitative intensity metrics for BPE (based on a single first post-contrast subtraction slice)	Greater BPE parameters (area, higher mean, SD, & quartiles of intensity) were found for cancer subjects than control
Saha et. al. [105]	High risk screening MRI: 133 patients, 46 developed cancers within 2 years	*Clinical*: N/A *Image-based*: BI-RADS BPE, 8 automatic features of BPE	Imaging features remained independently predictive of subsequent development of cancer

Table 4.2 (continued): Examples of image-based breast cancer risk assessment studies.

Model Reference	Dataset	Features	Key Result
Arasu et. al. [116]	4,247 women in Breast Cancer Surveillance Consortium (BCSC); 176 developed cancers	*Clinical*: breast density, first-degree family history, menopausal status, MRI indication, and BCSC risk score *Image-based*: BI-RADS BPE	Increased levels of BPE demonstrated significantly increased future breast cancer risk; high BPE and high BCSC 5-year risk score increased risk.
Niell et. al. [106]	High risk women without personal history of cancer: 19 developed cancer, 76 controls	Clinical: sociodemographic features Image-based: BI-RADS BPE, 91 quantitative BPE measures	Women subsequently diagnosed with breast cancer were more likely to have mild, moderate, or marked BPE
Portnoi et. al. [104]	1,656 consecutive MRIs from 1,183 high risk screening patients	Clinical: Traditional risk factors (Tyrer-Cuzick), cancer history Image-based: BI-RADS BPE, Deep learned features	Deep learning model improved individual risk discrimination compared to Tyrer-Cuzick model or traditional risk factors.

Growing datasets from the population of high-risk women undergoing screening MRI will allow for investigation of the relationships between the imaging biomarkers present on MRI and mammograms, such as BPE and breast density. Based on the studies referenced in Table 4.2, a risk assessment model that combines the predictive power of imaging features and clinical risk factors would be expected to provide improved individualized assessments compared to the established clinical models in Table 4.1. Before such a risk model can be established, a robust method for quantifying useful BPE values must be validated for a high-risk dataset. Therefore, in this chapter, we have implemented the BPE scoring algorithm developed in Chapter 3 on an independent test set of high-risk screening patients. In an independent validation of the BPE scoring algorithm, we investigated the correlation of the computer BPE scores with the radiologist BPE ratings and we assessed the BPE level classification performance of the computer BPE scores. Further, we

performed exploratory ROC analysis to evaluate the potential role that the computer BPE scores may have in predicting cancer versus negative or benign (i.e., non-cancer) diagnoses.

4.2 Dataset

In order to perform an independent validation of the methods developed in Chapter 3, we implemented the BPE scoring algorithm on a new dataset of high-risk screening patients. This dataset included 490 conventional breast DCE-MR exams retrospectively collected between 2009-2022 from 313 high-risk screening patients (aged 23-91) at the University of Chicago under a HIPAA-compliant IRB-approved (16-0400) protocol. Each patient had previously been classified as high-risk based on a known personal history of cancer, family history of cancer, or positive genetic mutation, and had therefore been recommended for screening MRI. The dataset of high-risk patients included 303 patients that were negative or benign and 10 patients that developed cancer over the course of screening (Table 4.3).

Routine bilateral breast MRIs were acquired using a Philips Achieva scanner with 1.5T or 3.0T magnet strength. The breast DCE-MRI protocol included a fat-saturated 3D T1 weighted spoiled gradient-echo sequence that was used to acquire pre- and post-contrast images with a temporal resolution of 60-75 seconds (TE = 2.3-2.7 ms, TR = 4.7-5.3 ms, flip angle = 10-12°, in-plane resolution = 0.56-0.80 mm, FOV = 29.9-44.0 cm, matrix = 376-528 x 376-472, slice thickness = 1.6-2.0 mm, interslice gap = 0.8-1.0 mm). The most-recent MRI scan available for each patient is referred to as the "diagnostic scan," and the second most-recent MRI scan is referred to as the "first prior scan." Of the patients with more than one MRI scan available, the time elapsed between the first prior and the most-recent diagnostic scans ranged from approximately six to 36 months. All scans had radiologist BPE ratings available from prior clinical review (Table 4.4).

Of the 303 patients that were negative or benign, seventeen patients had a benign lesion identified in at least one scan. In one patient, the same benign lesion was identified in both the prior and diagnostic scan. In three patients, the benign lesion was identified in the first prior scan, and the diagnostic scan was negative. In five patients, the first prior was negative, and the benign lesion was identified in the diagnostic scan. Of the ten patients that developed cancer, seven patients had a cancerous lesion identified on the diagnostic scan; all but one of those patients had first prior scans available. For two patients, there was no lesion specifically identified, but cancer was known at the time of the diagnostic scan. There was one patient diagnosed with cancer following the most-recent negative scan available, so their only scan was considered the first prior.

Table 4.3: Number of high-risk screening patients in the dataset. Patients belonging to the benign group had a benign diagnosis at either the first prior or diagnostic scan, and negative patients had negative diagnosis for all scans available.

	Developed Cancer	**Benign**	**Negative**
First prior scan only	1	-	-
Diagnostic scan only	2	8	125
Both scans available	7	9	161
Total patients	10	17	286

Table 4.4: Prevalence of radiologist BPE ratings present in the dataset of high-risk screening patients. The number of scans includes all diagnostic and first prior scans available.

	Developed Cancer	**Benign**	**Negative**	**Total**
Minimal BPE	6	4	151	161
Mild BPE	8	8	169	185
Moderate BPE	2	12	103	117
Marked BPE	1	2	24	27
Total scans	17	26	447	490

4.3 Independent Validation of BPE Scoring Technique

The BPE scoring method developed in Chapter 3 was implemented on the independent dataset of high-risk screening patients without any modification of the trained algorithm (Figure 4.1).

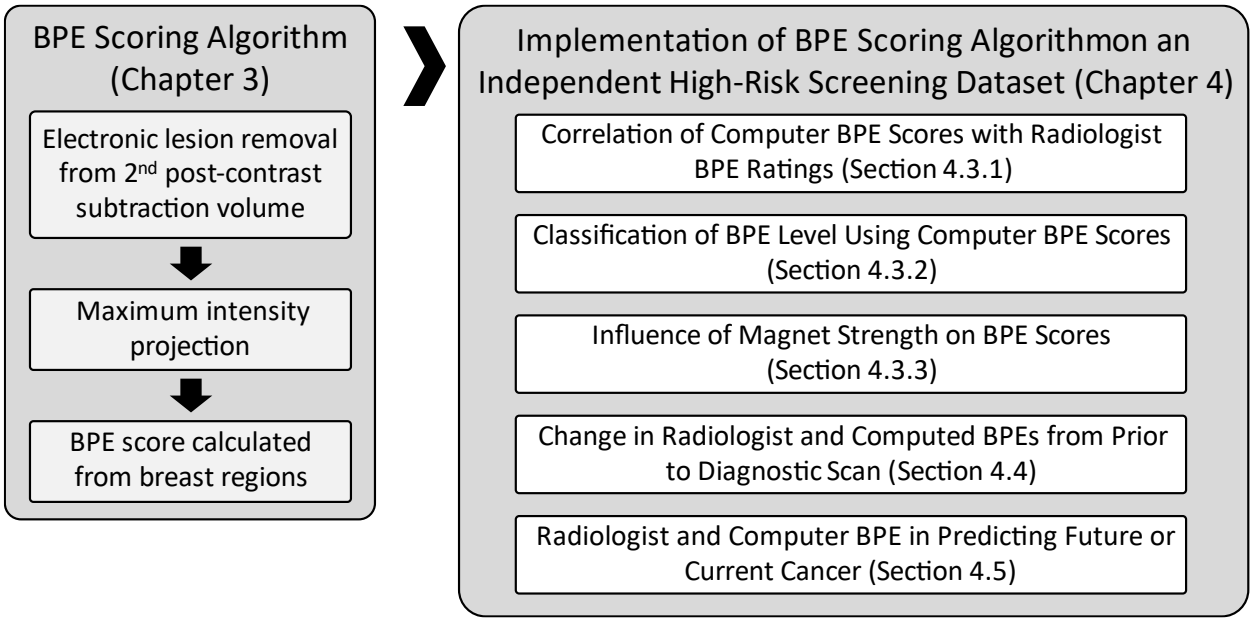

Figure 4.1: Organization of the evaluations performed using the BPE scoring algorithm developed in Chapter 3 (Figure 3.3) implemented on an independent dataset (Section 4.2) of high-risk screening MRIs.

Based on the dataset of diagnostic MRI patients that was used to develop the BPE scoring method in the prior chapter (Section 3.2), the computer BPE scores calculated from the affected breast on the second post-contrast subtraction MIP after lesion removal yielded the greatest AUC results in BPE level classification tasks (refer to Table 3.2). However, when assessing a screening patient, the presence of a lesion may not be known prior to the scan. Therefore, the computer BPE scores calculated from each of the breasts for each patient were evaluated in this chapter. For each scan that had a lesion identified, segmentations had previously been acquired using an in-house, automated 3D fuzzy c-means (FCM) approach; these segmentations were used for the electronic removal of the lesions prior to BPE scoring, as described in Section 3.3. [91]

For the remainder of Chapter 4, the only computer BPE scores evaluated were from breast regions on the second post-contrast subtraction MIPs following electronic lesion removal. To determine the strength and direction of the correlation of the computer BPE scores with radiologist BPE ratings on the independent dataset, Kendall's tau-b was used in rank correlation with a t-test used to assess the statistical significance of the correlation. [107] To determine the predictive value of the computer-extracted BPE scores for BPE level classification tasks on the independent dataset, receiver operating characteristic (ROC) analysis was performed using the proper binormal model. [108] Clinical radiologist BPE ratings were the only truth available for BPE assessment, so the performance of the computer BPE scores was compared to random guessing. ROC analysis was performed using computer BPE scores for binary classification of minimal versus marked BPE; it was also evaluated for binary classification of low (minimal, mild) versus high (moderate, marked) BPE (Figure 3.5). The statistical significance of the area under the ROC curve (AUC) relative to random guessing was determined using the z-test with Bonferroni corrections for multiple comparisons. [98]

4.3.1 Correlation of Computer BPE Scores with Radiologist BPE Ratings

The distribution of the computer BPE scores of the independent high-risk screening appeared similar to the distribution of the computer BPE scores of the previous dataset of diagnostic patients (Figure 4.2), and the results of a t-test failed to show a statistically significant difference between the computer BPE scores on independent datasets ($p > 0.05$, t-test). A single outlier in the high-risk screening dataset had a computer BPE score greater than 0.5 in the left breast; the abnormally high score was due to a poor breast segmentation that resulted in a computer BPE score calculated from a region containing only a few pixels from the chest wall.

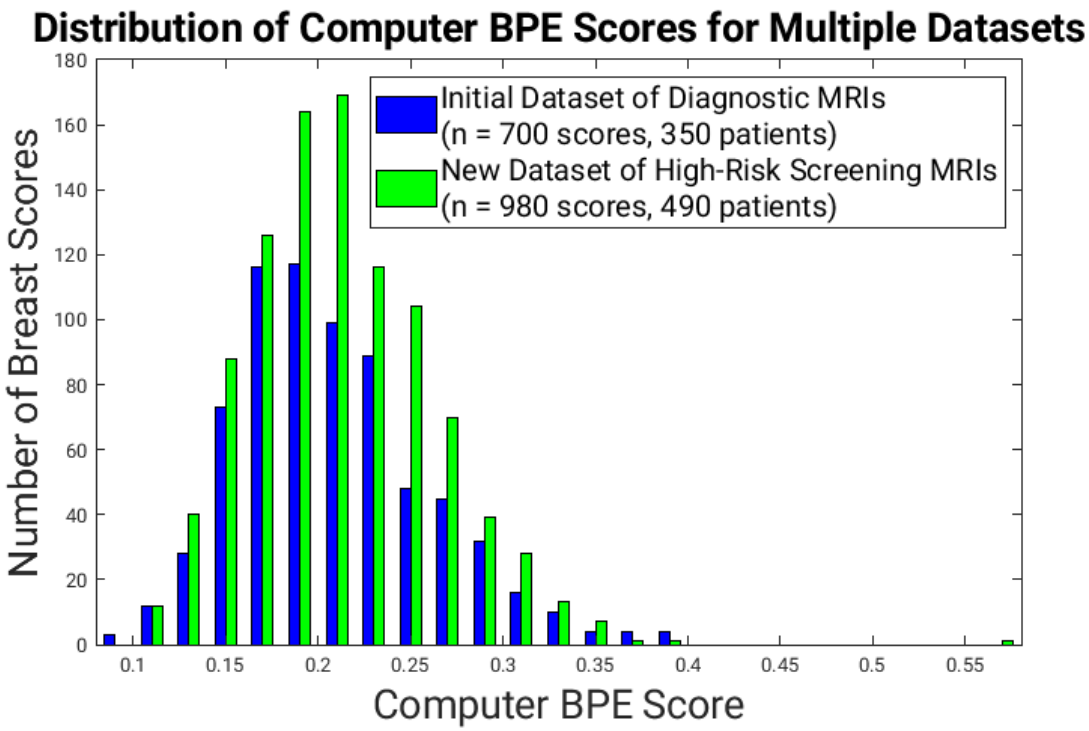

Figure 4.2: Histogram of computer BPE scores computed from each breast in all scans available from the datasets used in Chapters 3 and 4. Computer BPE scores were calculated from the second post-contrast subtraction MIP after electronic lesion removal. Histogram includes scores from each individual breast from all available scans. Results of a t-test failed to show a statistically significant difference between the computer BPE scores calculated on independent datasets ($p > 0.05$).

As seen in Figure 4.3, the scores calculated from the left and right breasts of each patient appeared similar, and the results of a t-test failed to show statistically significant difference between computer BPE scores calculated from the left and right breasts in each radiologist BPE rating group ($p > 0.05$, t-test). As with the previous dataset, the computer BPE scores of the high-risk screening dataset had a statistically significantly positive correlation with radiologist BPE ratings, although we failed to show a statistically significant correlation of computer BPE scores to radiologist BPE ratings for the subset of cases that developed cancer (Table 4.5 – Table 4.7).

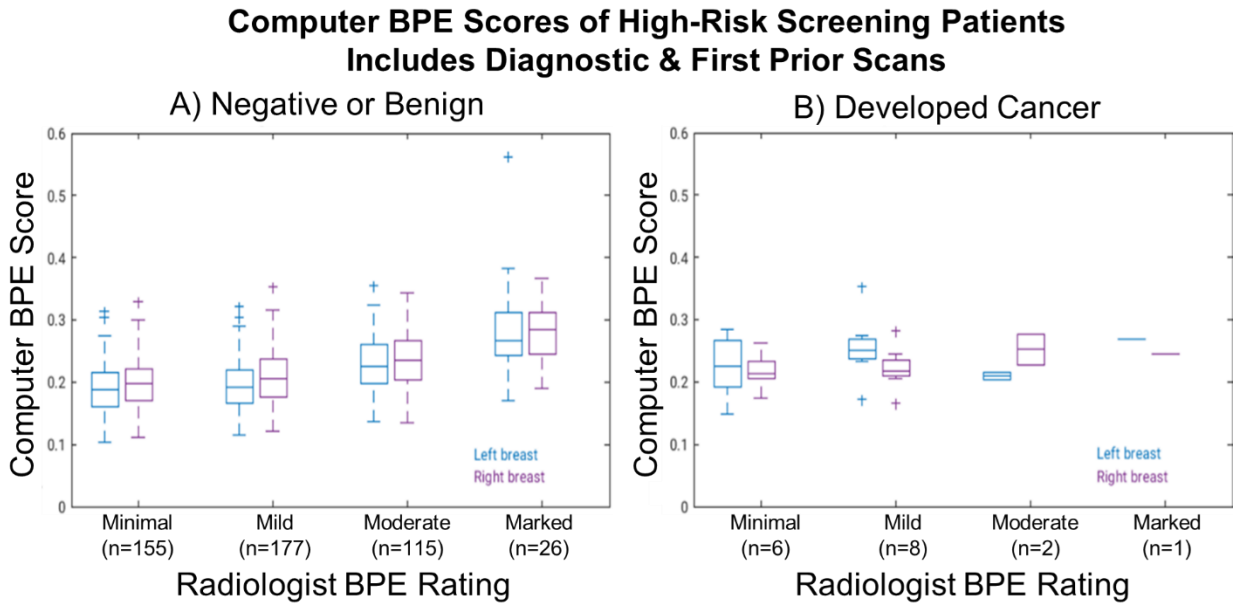

Figure 4.3: Positive correlations between the computer BPE scores and radiologist BPE ratings were statistically significant for all breast regions of negative or benign patients ($p < 0.001$, t-test) and results of the t-test failed to show a statistically significant correlation for all breast regions in patients that developed cancer ($p > 0.05$). Table 4.5 contains the details for the associated results from Kendall's rank correlation.

The computer BPE scores calculated from all breast regions of all the patients had statistically significant positive correlations with the radiologist BPE ratings, including both scans, only the diagnostic scan, and only the first prior scan (Table 4.5 – Table 4.7). Additionally, the combined group of benign and negative patients and the group of negative patients had statistically significant positive correlations with the radiologist BPE ratings, including both scans, only the diagnostic scan, and only the first prior scan (Table 4.5 -Table 4.7). For the small number of patients that had benign lesions and for the patients that developed cancer, the computer BPE scores failed to show a statistically significant correlation with the radiologist BPE ratings based on both scans, only the diagnostic scan, and only the first prior scan, with the exception of the affected breasts in the first prior scan of patients that developed cancer.

Table 4.5: Results from Kendall's rank correlation including the diagnostic and first prior scans. Raw, uncorrected p-vales from the t-test are reported in the table. Asterisks indicate a statistically significant correlation between the computer BPE scores and radiologist BPE ratings. Statistical significance was assessed using the Bonferroni correction for multiple comparisons.

	Breast Region Scores	Kendall's rank correlation tau-b
All patients ($n_{patients}$ = 313, n_{scans} = 490)	Both	0.280 (p = 1.40e-30) *
	Right	0.286 (p = 8.16e-17) *
	Left	0.276 (p = 1.22e-15) *
Developed Cancer ($n_{patients}$ = 10, n_{scans} = 17)	Both	0.150 (p = 0.279)
	Right	0.268 (p = 0.192)
	Left	0.089 (p = 0.686)
	Affected	0.375 (p = 0.065)
	Unaffected	-0.018 (p = 0.964)
Negative and Benign ($n_{patients}$ = 303, n_{scans} = 473)	Both	0.286 (p = 7.07e-31) *
	Right	0.290 (p = 1.44e-16) *
	Left	0.285 (p = 4.15e-16) *
Benign lesions ($n_{patients}$ = 17, n_{scans} = 26)	Affected	0.356 (p = 0.024)
	Unaffected	0.3113 (p = 0.049)
Negative ($n_{patients}$ = 286, n_{scans} = 447)	Right	0.281 (p = 6.16e-15) *
	Left	0.285 (p = 3.19e-15) *

Table 4.6: Results from Kendall's rank correlation including the diagnostic scan only. Raw, uncorrected p-vales from the t-test are reported in the table. Asterisks indicate a statistically significant correlation between the computer BPE scores and radiologist BPE ratings. Statistical significance was assessed using the Bonferroni correction for multiple comparisons.

	Breast Region Scores	Kendall's rank correlation tau-b
All patients ($n_{patients}$ = 313, n_{scans} = 313)	Both	0.255 (p = 4.99e-9) *
	Right	0.256 (p = 3.01e-9) *
	Left	0.257 (p = 2.60e-9) *
Developed Cancer ($n_{patients}$ = 10, n_{scans} = 10)	Both	-0.026 (p = 0.916)
	Right	0.105 (p = 0.771)
	Left	-0.053 (p = 0.923)
	Affected	0.158 (p = 0.628)
	Unaffected	-0.211 (p = 0.498)
Negative and Benign ($n_{patients}$ = 303, n_{scans} = 303)	Both	0.263 (p = 1.77e-17) *
	Right	0.259 (p = 3.38e-09) *
	Left	0.268 (p = 9.24e-10) *
Benign lesions ($n_{patients}$ = 17, n_{scans} = 17)	Affected	0.433 (p = 0.031)
	Unaffected	0.329 (p = 0.103)
Negative ($n_{patients}$ = 286, n_{scans} = 286)	Right	0.244 (p = 6.44e-8) *
	Left	0.266 (p = 4.43e-9) *

Table 4.7: Results from Kendall's rank correlation including the first prior scan only. Raw, uncorrected p-vales from the t-test are reported in the table. Asterisks indicate a statistically significant correlation between the computer BPE scores and radiologist BPE ratings. Statistical significance was assessed using the Bonferroni correction for multiple comparisons.

	Breast Region Scores	**Kendall's rank correlation tau-b**
All patients ($n_{patients}$ = 178, n_{scans} = 178)	Both	0.322 (p = 1.79e-15) *
	Right	0.460 (p = 1.05e-10) *
	Left	0.427 (p = 2.79e-9) *
Developed Cancer ($n_{patients}$ = 8, n_{scans} = 8)	Both	0.482 (p = 0.024)
	Right	0.564 (p = 0.100)
	Left	0.477 (p = 0.179)
	Affected	0.824 (p = 7.14e-3) *
	Unaffected	0.390 (p = 0.286)
Negative and Benign ($n_{patients}$ = 170, n_{scans} = 170)	Both	0.324 (p = 5.22e-15) *
	Right	0.336 (p = 1.05e-8) *
	Left	0.315 (p = 8.28e-8) *
Benign lesions ($n_{patients}$ = 9, n_{scans} = 9)	Affected	0.204 (p = 0.574)
	Unaffected	0.272 (p = 0.431)
Negative ($n_{patients}$ = 161, n_{scans} = 161)	Right	0.339 (p = 2.04e-8) *
	Left	0.319 (p = 1.33e-7) *

Figure 4.4 combines the left and right breast computer BPE scores using all available scans for each group of patients: the combined group of negative and benign patients and the patients that developed cancer. From Figure 4.4, it appeared that the computer BPE scores tend to be greater for the cancer cases than the negative and benign cases in the lower radiologist BPE rating levels. Using the computer BPE scores of the first prior and the diagnostic scans, the results of the t-test failed to show a statistically significant difference between the computer BPE scores of the cancer patients and those of the negative and benign patients for moderate or marked BPE ($p > 0.05$), however, there was a statistically significant difference between the computer BPE scores of the cancer patients and those of the negative or benign patients for the minimal and mild BPE groups ($p < 0.05$). Figure 4.5 displays the computer BPE scores for each radiologist BPE rating, including the first prior scan only (note there were no cancer cases rated moderate BPE). Using only the

computer BPE scores of the first prior scan, results of the t-test failed to show a statistically significant difference between the computer BPE scores of the cancer patients and those of the negative or benign patients for minimal and marked BPE ($p > 0.05$), and there was a statistically significant difference between the computer BPE scores of the cancer patients and those of the negative or benign patients for the mild BPE group ($p < 0.05$). We speculate that for patients who tend to have weaker background enhancement, computer BPE scores may be able to differentiate a future cancer diagnosis from a negative or benign diagnosis. For the patients who tend to have stronger enhancement, the radiologist BPE ratings may be a sufficient indication of increased risk for cancer. Therefore, the computer BPE scores have the potential to serve as a useful factor to predict cancer versus non-cancer diagnoses, particularly in patients with low BPE.

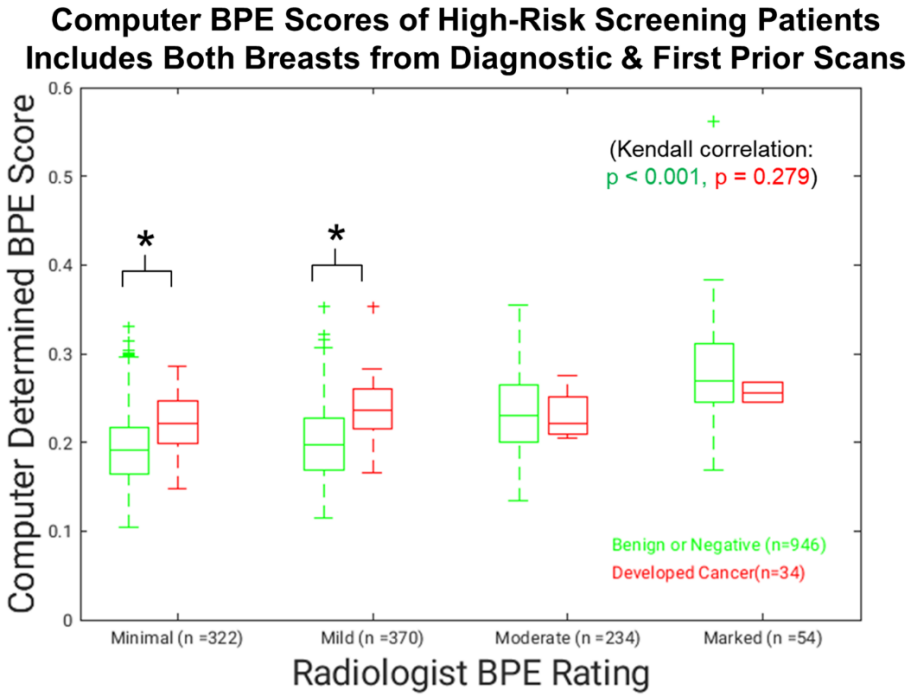

Figure 4.4: Based on all scans available, a statistically significant difference (asterisk) was found between the computer BPE scores of patients that developed cancer and those of benign or negative patients for the minimal and mild radiologist BPE rating groups ($p < 0.05$, t-test). Positive correlations between the computer BPE scores and radiologist BPE ratings, including all breasts of all patients, were statistically significant ($p < 0.001$, t-test). Additional results of the associated Kendall rank correlations are in Table 4.5. ($n_{patients} = 313$, $n_{scans} = 490$, $n_{breasts} = 980$)

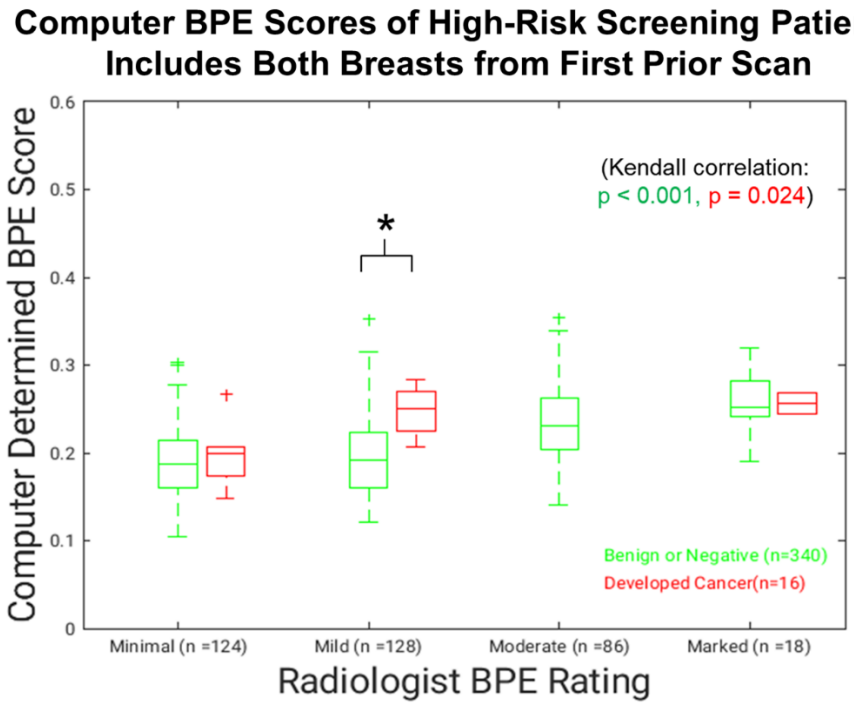

Figure 4.5: Based on the first prior scans only, a statistically significant difference (asterisk) was found between the computer BPE scores of patients that developed cancer and those of benign or negative patients for the mild radiologist BPE rating group ($p < 0.05$, t-test). Positive correlations between the computer BPE scores, including all breasts of all patients, and radiologist BPE ratings were significant ($p < 0.001$, t-test). Additional results of the associated Kendall rank correlations are in Table 4.7. ($n_{patients} = 178$, $n_{scans} = 178$, $n_{breasts} = 356$)

4.3.2 Classification of BPE Level Using Computer BPE Scores

As with the previous dataset of diagnostic scans (refer to Section 3.6), the computer BPE scores calculated from the high-risk screening dataset performed well in many of the binary BPE level classification tasks (Table 4.8). In the marked versus minimal BPE classification, computer BPE scores performed statistically significantly greater than random guessing for most of the breast regions scored (Table 4.8). We failed to show statistical significance greater than random guessing for the computer BPE scores of the unaffected breasts of the cancer patients and each of the breast regions for the patients with benign lesions (Table 4.8). The high versus low BPE task was expected to have lower performance than the marked versus minimal BPE task due to the difficulty in the task of distinguishing between the intermediate levels relative to distinguishing the two

extreme levels of BPE. Overall, the computer BPE scores yielded greater AUC results for the marked versus minimal BPE task than the high versus low BPE task (Table 4.8). The results showed that computer BPE scores from the patients in the combined benign and negative group, all patients, and the negative patients only each performed statistically significantly greater than random guessing, however the patients with known benign or cancer lesions failed to perform statistically significantly greater than random guessing in many of the classification tasks (Table 4.8). In some cases, the AUC could not compute due to a complete overlap of values in each group.

Table 4.8: Results of ROC analysis for computer BPE scores in BPE level classification on the independent dataset. Computer BPE scores were calculated from diagnostic and first prior MRIs, when available. High BPE includes marked or moderate BPE, and low BPE includes mild or minimal BPE. Raw, uncorrected p-vales from the z-test are reported in the table. Asterisks on AUC results indicate performance statistically significantly greater than random guessing. Statistical significance of the AUCs was assessed using the Bonferroni correction for multiple comparisons.

(No. MRI scans) [minimal, mild, moderate, marked]	**Breast Region Scores**	**Marked vs. Minimal BPE AUC ± SE**	**High vs. Low BPE AUC ± SE**
All patients ($n_{patients}$ = 313, n_{scans} = 490) [161, 185, 117, 27]	Both	0.865 ± 0.029 (p = 1.47e-35) *	0.731 ± 0.017 (p = 1.11e-40) *
	Right	0.907 ± 0.028 (p = 8.11e-47) *	0.741± 0.024 (p = 6.22e-24) *
	Left	0.850 ± 0.045 (p = 1.33e-14) *	0.726 ± 0.025 (p = 1.53e-19) *
Developed Cancer ($n_{patients}$ = 10, n_{scans} = 17) [6, 8, 2, 1]	Both	0.871 ± 0.097 (p = 1.37e-4) *	0.590 ± 0.105 (p = 0.395)
	Right	0.908 ± 0.114 (p = 3.32e-4) *	0.856 ± 0.089 (p = 6.54e-5) *
	Left	0. 908 ± 0.114 (p = 3.32e-4) *	0.506 ± 0.168 (p = 0.973)
	Affected	0.908 ± 0.114 (p = 3.32e-4) *	0.795 ± 0.133 (p = 0.026)
	Unaffected	0.784 ± 0.211 (p = 0.179)	Could not compute
Negative and Benign ($n_{patients}$ = 303, n_{scans} = 473) [155, 177, 115, 26]	Both	0.867 ± 0.030 (p = 3.61e-35) *	0.738 ± 0.017 (p = 3.07e-42) *
	Right	0.909 ± 0.029 (p = 2.29e-46) *	0.742 ± 0.024 (p = 1.13e-23) *
	Left	0.854 ± 0.046 (p = 1.70e-14) *	0.737 ± 0.025 (p = 3.22e-21) *
Benign lesions ($n_{patients}$ = 17, n_{scans} = 26) [4, 8, 12, 2]	Both	Could not compute	0.651 ± 0.061 (p = 0.013) *
	Affected	Could not compute	0.690 ± 0.085 (p = 0.025)
	Unaffected	Could not compute	0.650 ± 0.089 (p = 0.091)

Table 4.8 (continued): Results of ROC analysis for computer BPE scores in BPE level classification on the independent dataset.

(No. MRI scans) [minimal, mild, moderate, marked]	Breast Region Scores	Marked vs. Minimal BPE AUC ± SE	High vs. Low BPE AUC ± SE
Negative ($n_{patients}$ = 286, n_{scans} = 447) [151, 169, 103, 24]	Both	0.891 ± 0.035 (p = 4.02e-29) *	0.755 ± 0.025 (p = 1.15e-24) *
	Right	0.902 ± 0.031 (p = 2.75e-39) *	0.746 ± 0.025 (p = 5.90e-23) *
	Left	0.891 ± 0.035 (p = 4.02e-29) *	0.755 ± 0.025 (p = 1.15e-24) *

4.3.3 Influence of Magnet Strength on Independent Set of Computer BPE Scores

Additionally, we checked for differences in computer BPE scores across MRI magnet strengths (Figure 4.6). As was true for the original dataset of diagnostic patients (refer to Section 3.7), we failed to show a statistically significant difference between the computer BPE scores for 1.5T and 3.0T for all patients (p = 0.300, t-test), for only patients that developed cancer (p = 0.243, t-test), and for only patients that were negative or benign (p = 0.204, t-test).

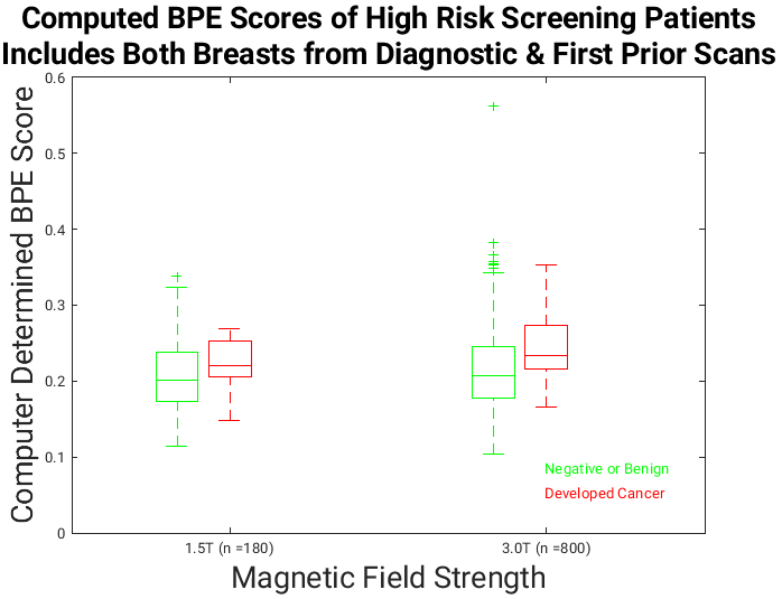

Figure 4.6: Computer BPE scores (second post-contrast subtraction MIP) for each magnet field strength, including all breasts from all scans available for all patients. ($n_{patients}$ = 313, n_{scans} = 490, $n_{breasts}$ = 980) The results of the t-tests failed to show a statistically significant difference between magnet strengths for each group of patients (p > 0.05).

We have previously speculated that radiologists more familiar with 1.5T enhancement appearance may have overestimated BPE ratings as the use of 3.0T scanners increased in the clinic between 2005-2017 (refer to section 3.7). Therefore, we assessed the prevalence of each magnet strength used for the high-risk screening dataset (2009-2022) and, further, the prevalence of each radiologist BPE rating by magnet strength. Figure 4.7 demonstrates the prominent use of 3.0T for this dataset, which we would expect to reduce potential effects of magnet strength on the radiologist perception of BPE levels. The results shown in Figure 4.8 shows the prevalence of radiologist BPE ratings assigned for each magnet strength. The distribution of each radiologist BPE rating assigned across magnet strengths suggests that after an adjustment period, radiologists may have been able to more consistently assign BPE ratings to images acquired with either magnet strength. However, this remains speculative until a reader study with patients matched across magnet strengths may be performed.

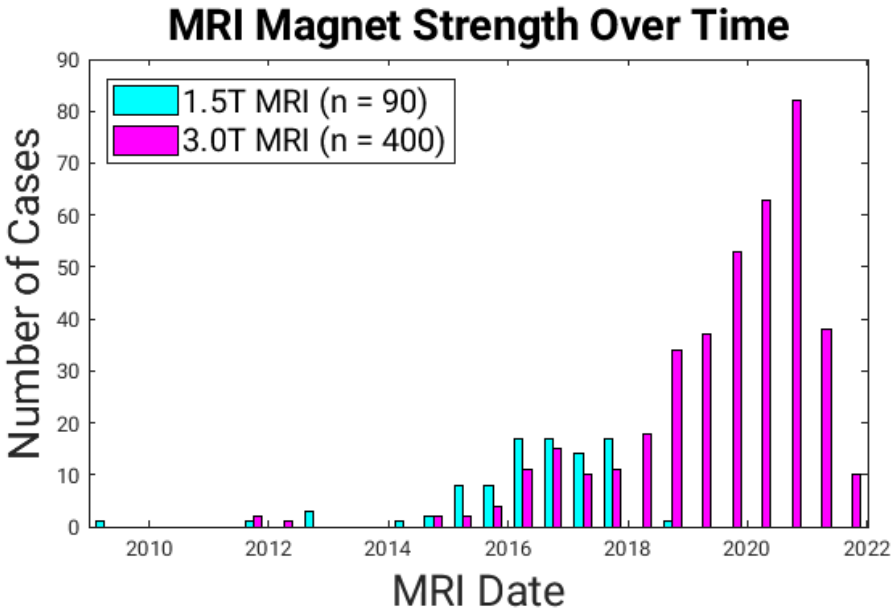

Figure 4.7: Histogram of the number of cases acquired using either 1.5T or 3.0T DCE-MRI over the time period of the high-risk screening MRI dataset ($n_{patients}$ = 313, n_{scans} = 490).

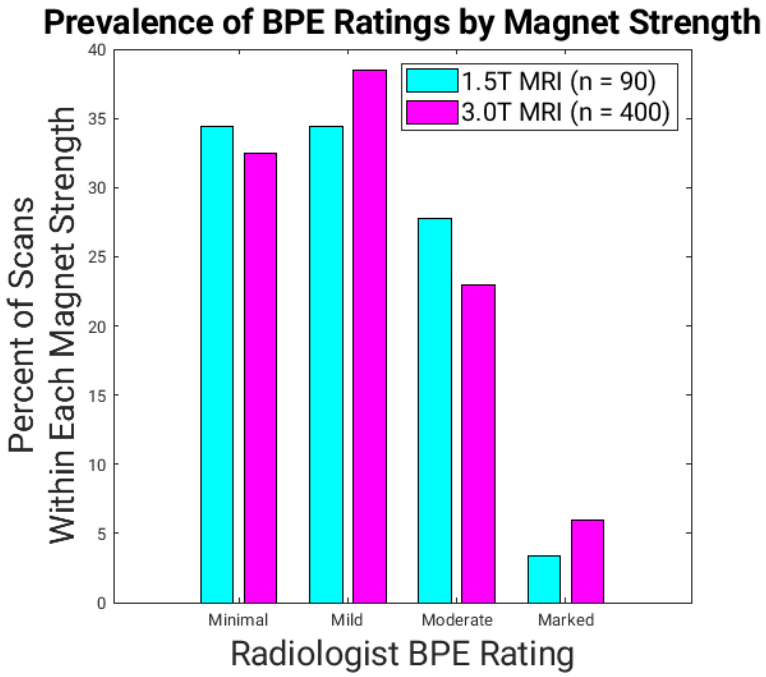

Figure 4.8: Prevalence of each radiologist BPE rating in the high-risk screening dataset ($n_{patients}$ = 313, n_{scans} = 490), split by magnet strength.

4.4 Change in Radiologist and Computer BPEs from Prior to Diagnostic Scan

Radiologist BPE ratings and computer BPE scores may change between the first prior and the most recent diagnostic scans. Figure 4.9 shows the radiologist BPE ratings assigned to each patient at their diagnostic scan and, if available, first prior scan. Based on the technique used in [117], we investigated the trend for changes in radiologist BPE ratings from the first prior to the scan at diagnosis. Given that we had at most one prior scan available for each patient, the line of best fit for each patient's change in BPE rating was exactly linear between the two datapoints available. The varying time between scans for each patient was accounted for in the horizontal component of the slope, and the vertical difference in each BPE rating was calculated in single unit increments (i.e., mild to moderate is +1, marked to mild is -2). By removing the y-intercept from the line of best fit, we could assess the relative trend for BPE levels from the first prior to the diagnostic scan for all patients and, further, whether there was a difference in the average trends for cancer versus

negative and benign patients (Figure 4.10). The same technique was used to investigate the computer BPE scores calculated from the first prior and diagnostic scans (Figure 4.11 and Figure 4.12), however the vertical slope component was based on the exact computer BPE scores that were calculated as unique values.

Based on the radiologist BPE ratings and the computer BPE scores, the results failed to show a statistically significant difference between the average trends for cancer patients or benign and negative patients (Figure 4.10 and Figure 4.12). For the computer BPE scores, there was a slight increase in the average trend for the cancer cases more than the negative or benign cases (Figure 4.12). Although we failed to demonstrate statistically significant results, we speculate that computer BPE scores may increase when radiologist BPE ratings do not in patients that developed cancer. This demonstrates the potential for the computer BPE scores to predict an increase in BPE, which is known to be associated with an increased risk of breast cancer.

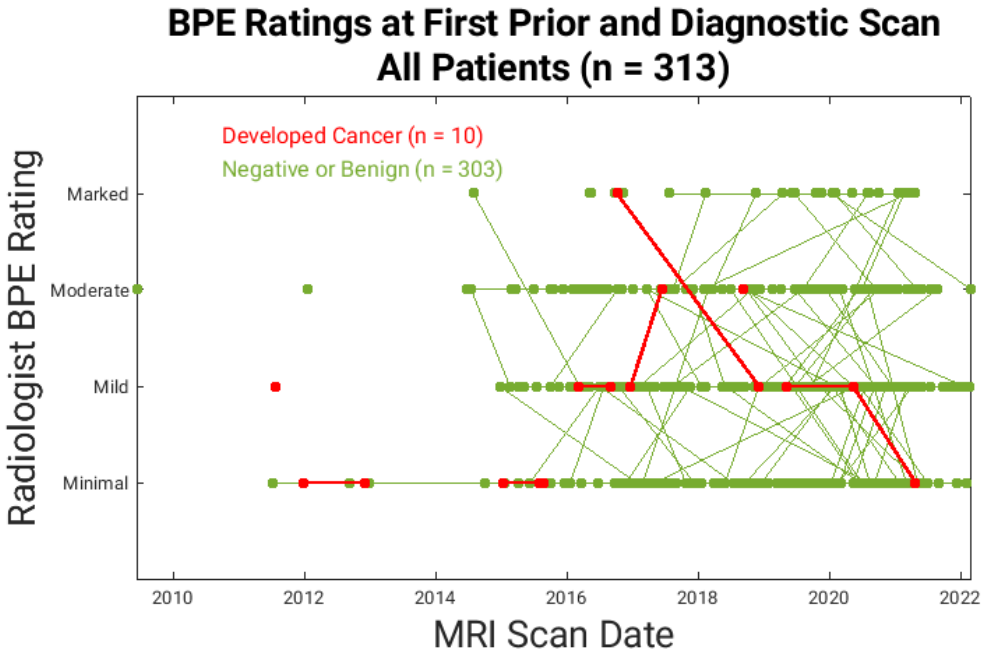

Figure 4.9: Radiologist BPE ratings assigned to each patient at the diagnostic scan and, if available, first prior scan. A single point indicates a patient that had only one scan; lines connect scans for each patient. ($n_{patients}$ = 313)

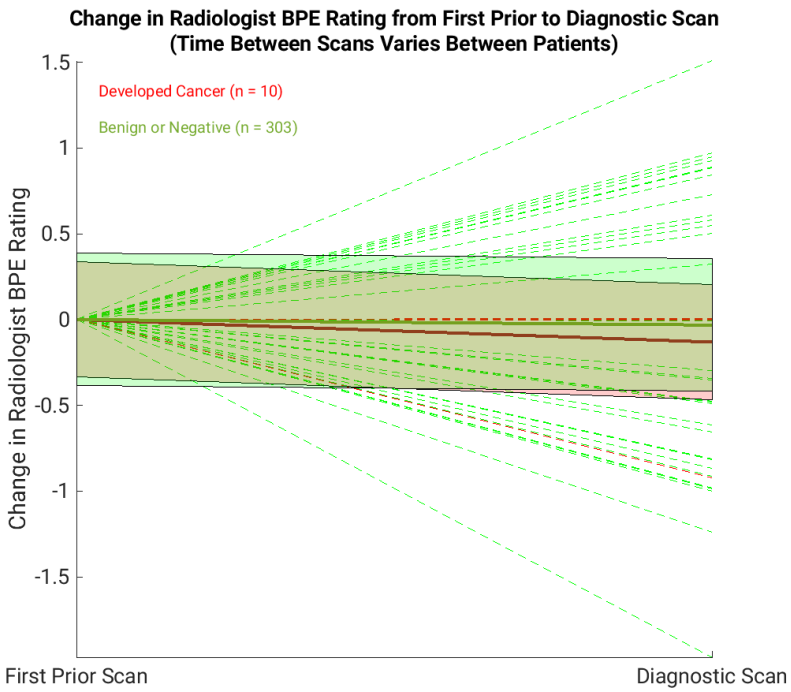

Figure 4.10: Change in radiologist BPE ratings assigned at first prior scan to the scan at diagnosis. Includes only patients with two scans available ($n_{patients}$ = 177). Dashed lines represent each patient trend, solid lines are the average trends for each group, and the standard deviation of the trends is shaded around the average.

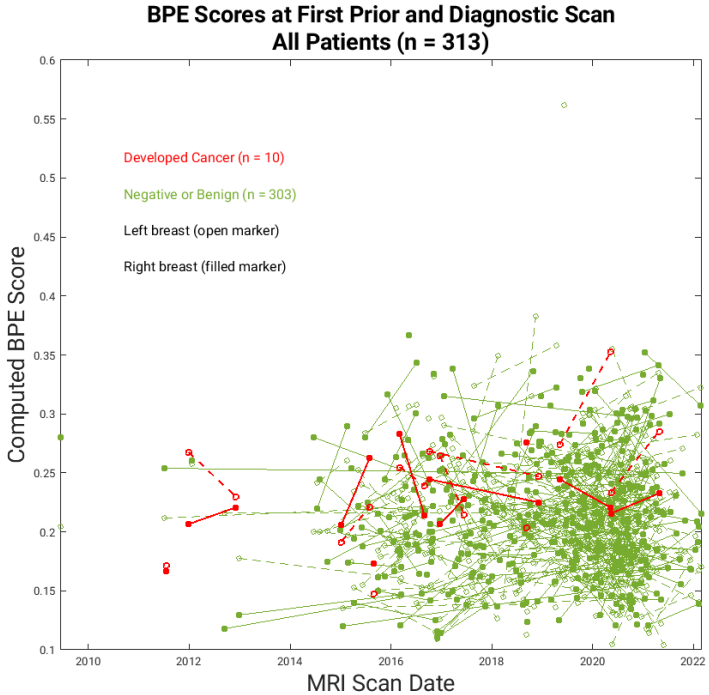

Figure 4.11: Computer BPE scores calculated for each breast at the diagnostic scan and, if available, first prior scan for each patient. A single point indicates a patient that had only one scan; lines connect scans for each patient. ($n_{patients}$ = 313)

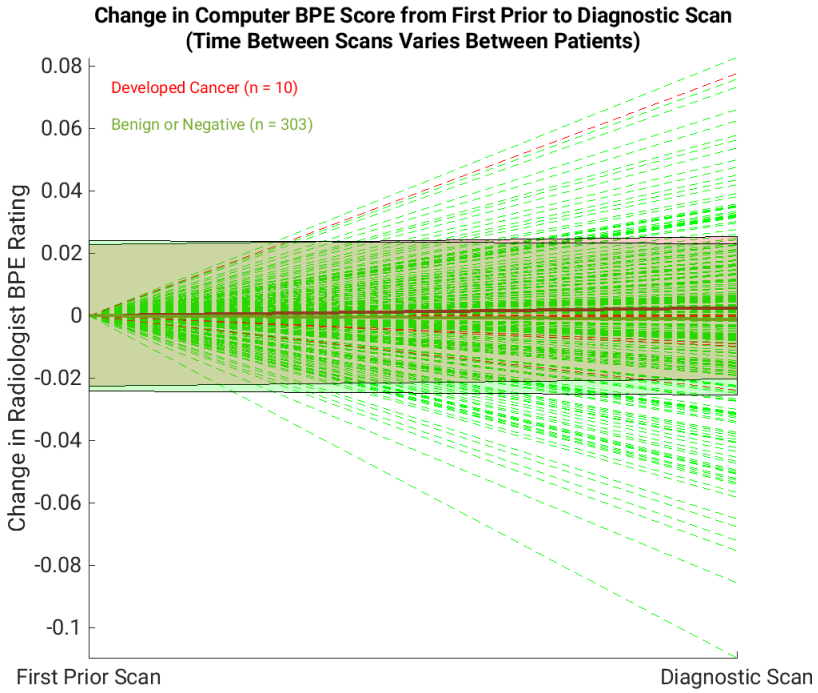

Figure 4.12: Change in computer BPE scores calculated at the first prior scan to the scan at diagnosis. Includes only those patients with two scans available ($n_{patients}$ = 177). Dashed lines represent each patient trend, solid lines are the average trends for each group, and the standard deviation of the trends is shaded around the average.

4.5 Radiologist and Computer BPE in Predicting Future or Current Cancer

Exploratory ROC analysis was performed using the proper binormal model for binary classification tasks of distinguishing cancer versus negative and benign patients using computer BPE scores and radiologist BPE ratings of diagnostic or first prior scans; clinical diagnosis at the time of the most recent diagnostic scan served as truth for these analyses (Figure 4.13). [108] The statistical significance of the area under the ROC curve (AUC) relative to random guessing was determined using the z-test with Bonferroni corrections for multiple comparisons. [98] Table 4.9 contains the statistical results for the classification tasks demonstrated in Figure 4.13.

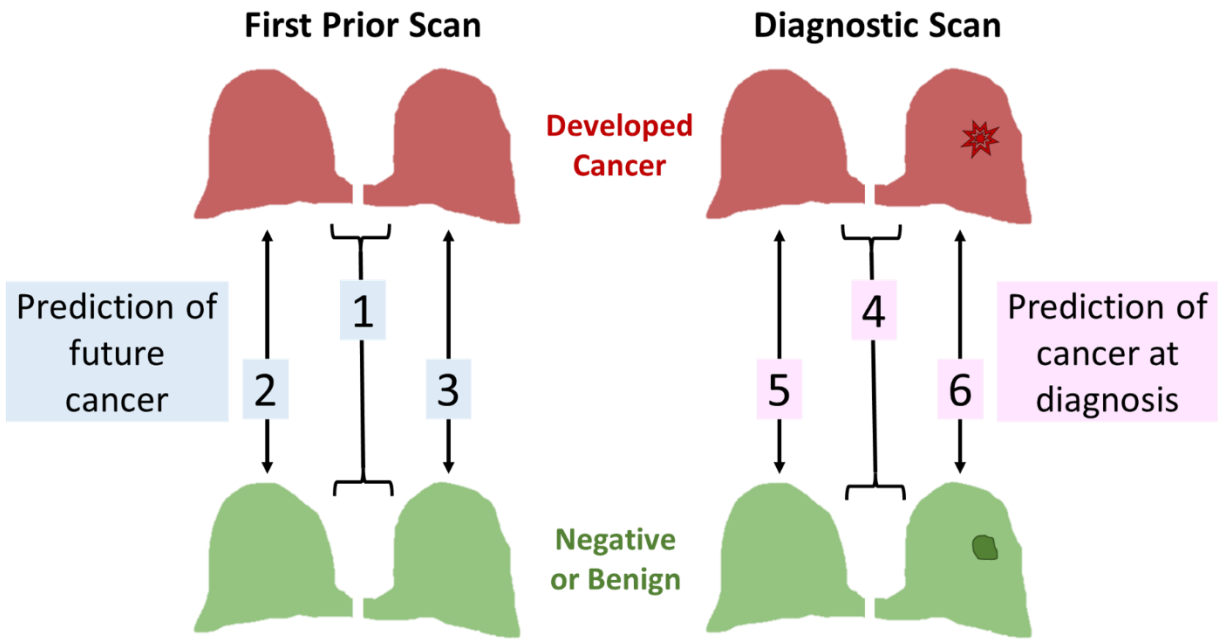

Figure 4.13: Visual representation of the exploratory ROC analysis performed using radiologist BPE ratings and computer BPE scores of the diagnostic or first prior scans in binary classification tasks of cancer versus non-cancer diagnoses. Figure 4.14 and Table 4.9 contain the associated AUC results.

Based on our observed trend of computer BPE being greater in cancer cases than in non-cancer cases (Figure 4.4 and Figure 4.5), in addition to the known association of stronger BPE presentation in patients that have developed cancer, we predicted the computer BPE scores would perform well in classification tasks between cancer and non-cancer cases. However, the statistical power of the results would be limited due to the limited number of cancer patients in the dataset. Given that we have aimed to assess the role that computer BPE scores may have in clinical risk assessment, the results we have based on prior scans only were of particular interest. For a breast cancer screening MRI, a potential "affected" breast is unknown, so the computer BPE scores of both breast regions were investigated.

The AUC results for the prediction of future cancer based on the first prior scan (tasks 1-3 in Figure 4.14) showed that the computer BPE scores calculated for both breasts or for the

unaffected breast performed statistically significantly greater than random guessing in cancer versus non-cancer classification, however the computer BPE scores failed to outperform random guessing for the affected breast (Table 4.9). For the same set of patient scans, the radiologist BPE ratings failed to perform statistically significantly greater than random guessing in the task of distinguishing between cancer and non-cancer diagnosis (tasks 1-3 in Table 4.9). Based on the diagnostic scan (tasks 4-6 in Figure 4.14), AUC results showed that the computer BPE scores calculated for both breasts or for the unaffected breast performed statistically significantly greater than random guessing in cancer versus non-cancer classification, however the computer BPE scores failed to outperform random guessing for the affected breast only (Table 4.9). For the same set of patient scans, the AUC for the radiologist BPE ratings could not compute due to a complete overlap of values in each group, thus the radiologist BPE ratings failed to perform statistically significantly greater than random guessing (tasks 4-6 in Table 4.9). These results further demonstrate the potential for computer BPE scores to be a useful addition to image-based breast cancer risk assessment models developed in the future.

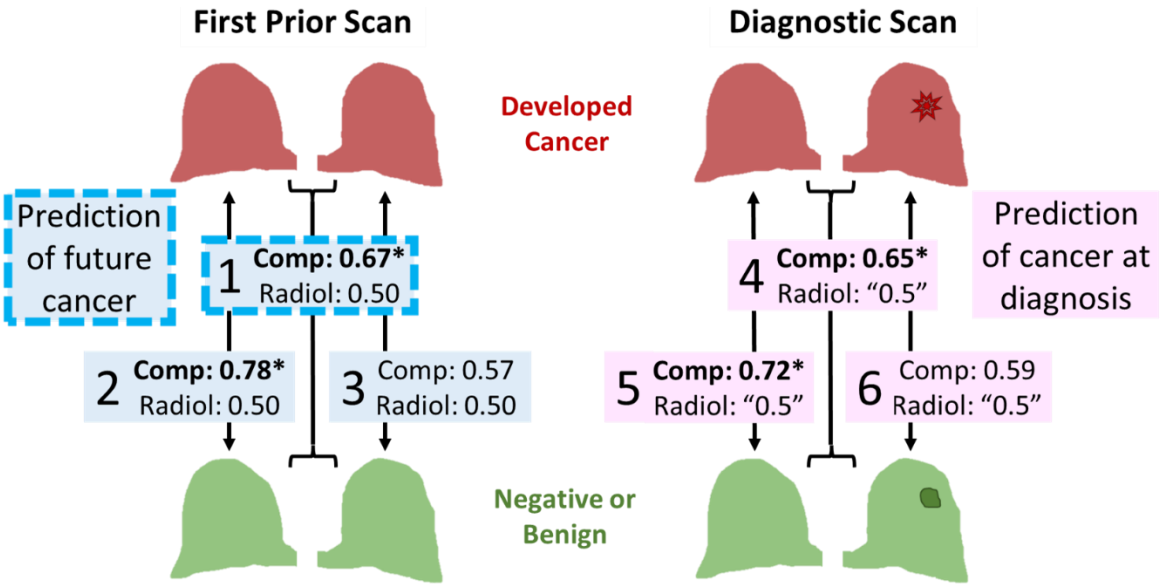

Figure 4.14: AUC results of the exploratory ROC analysis introduced in Figure 4.13. The most clinically relevant task is indicated by the dashed outline. ROC analysis was performed using computer BPE scores ("Comp") and radiologist BPE ratings ("Radiol") of the diagnostic or first prior scans in binary classification tasks of cancer versus non-cancer diagnosis. Asterisks indicate AUC values statistically significantly greater than random guessing after Bonferroni correction for three comparisons. AUC result displayed as "0.5" when ROC failed to compute to due a complete overlap of values in each group. Table 4.9 contains additional details associated with the exploratory ROC analysis.

Table 4.9: Exploratory results from ROC analysis using computer BPE scores or radiologist BPE ratings in various tasks. Data corresponds to Figure 4.13 and Figure 4.14. Raw, uncorrected p-vales from the z-test are reported in the table. Asterisks indicate AUC statistically significantly greater than random guessing. Statistical significance of the AUCs was assessed using the Bonferroni correction for three comparisons. For negative patients, the greatest of both breast scores for each patient were selected for the 'affected' breast comparisons. (n = number of scans)

	Task	Truth	"Positive class" for ROC analysis	"Negative class" for ROC analysis	Breast region(s)	Radiologist BPE rating AUC ± SE	Computer BPE score AUC ± SE
1	Distinguishing, based on BPE of first prior scan, future cancer diagnosis vs. future non-cancer diagnosis	Clinical diagnosis from diagnostic scan (either cancer or negative/benign)	First prior scan of patients that developed future cancer (n = 8)	First prior scan of negative and benign patients (n = 170)	Both	0.500 ± 0.089 (p = 0.997)	0.665 ± 0.062 (p = 8.25e-3) *
2					Unaffected	0.500 ± 0.089 (p = 0.997)	0.778 ± 0.082 (p = 7.08e-4) *
3					Affected	0.500 ± 0.089 (p = 0.997)	0.574 ± 0.085 (p = 0.381)
4	Distinguishing, based on BPE of diagnostic scan, cancer diagnosis vs. non-cancer diagnosis	Clinical diagnosis from diagnostic scan (either cancer or negative/benign)	Diagnostic scan of patients with known cancer (n = 9)	Diagnostic scan of negative and benign patients (n = 303)	Both	Could not compute	0.650 ± 0.056 (p = 7.62e-3) *
5					Unaffected	Could not compute	0.718 ± 0.071 (p = 2.10e-3) *
6					Affected	Could not compute	0.587 ± 0.092 (p = 0.349)

Although our dataset of cancer patients was limited, we performed supplementary exploratory ROC analysis between the first prior and diagnostic scans and the affected and unaffected breasts (tasks 7-14 in Figure 4.15). Table 4.10 contains the statistical results for the additional classification tasks demonstrated in Figure 4.15.

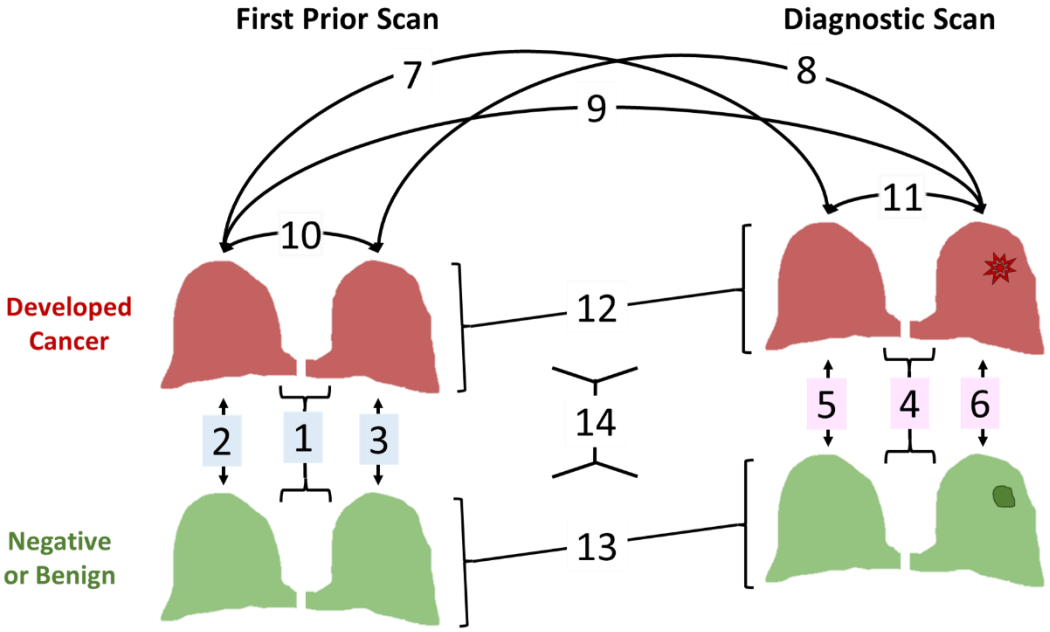

Figure 4.15: Visual representation of the supplementary exploratory ROC analyses performed using computer BPE scores and radiologist BPE ratings of the diagnostic or first prior scans in various binary classification tasks. Table 4.10 contains the associated AUC results for tasks 7-14.

Using the computer BPE scores or radiologist BPE ratings of the unaffected breast and affected breast failed to perform statistically significantly greater than random guessing in distinguishing diagnostic and prior scans of the patients that developed cancer (tasks 7-8, Table 4.10). Computer BPE scores and radiologist BPE ratings also failed to perform statistically significantly greater than random guessing in distinguishing the affected breast at diagnosis against the affect breast at the first prior scan for patients that developed cancer (task 9, Table 4.10). Using computer BPE scores and radiologist BPE ratings of the first prior scan of patients that developed cancer failed to perform statistically significantly greater than random guessing in distinguishing

the breast that would develop a future cancer from the unaffected breast (task 10, Table 4.10). Using computer BPE scores and radiologist BPE ratings of the diagnostic scan of patients that developed cancer failed to perform statistically significantly greater than random guessing in distinguishing the affected breast from the unaffected breast (task 11, Table 4.10).

An increase of BPE from first prior to diagnostic scan would be expected for patients that developed cancer, however, classification performance based on both breasts failed to outperform random guessing using the computer BPE scores and radiologist BPE ratings in the task of distinguishing prior and diagnostic scans of patients that developed cancer (task 12, Table 4.10). Although no change in BPE would be expected from first prior to diagnostic BPE for negative and benign cases, the AUC result from computer BPE scores was statistically significantly greater than random guessing in distinguishing prior and diagnostic scans of negative and benign patients; the radiologist BPE ratings failed to perform statistically significantly greater than random guessing for the same task (task 13, Table 4.10). Using the computer BPE scores of both breasts, including both diagnostic and prior scans, results showed performance statistically significantly greater than random guessing in distinguishing the patients that developed cancer from the negative and benign patients (task 14, Table 4.10).

Table 4.10: Supplementary exploratory results from ROC analysis using computer BPE scores or radiologist BPE ratings in various tasks. Data corresponds to Figure 4.15. Raw, uncorrected p-vales from the z-test are reported in the table. Asterisks on AUC values indicate statistically significantly greater than random guessing. Statistical significance of the AUCs was assessed using the Bonferroni correction for multiple comparisons. (n = number of scans)

	Task	**Truth**	**"Positive class" for ROC analysis**	**"Negative class" for ROC analysis**	**Breast region(s)**	**Radiologist BPE rating AUC ± SE**	**Computer BPE score AUC ± SE**
7	Distinguishing, using BPE, unaffected breast at diagnosis vs unaffected breast at first prior	Prior or diagnostic scan	Unaffected breast on diagnostic scan of patients with known cancer (n = 9)	Unaffected breast on first prior scan of patients that developed future cancer (n = 8)	Unaffected	0.540 ± 0.155 (p = 0.795)	0.507 ± 0.116 (p = 0.954)

Table 4.10 (continued): Supplementary exploratory results from ROC analysis using computer BPE scores or radiologist BPE ratings in various tasks. Data corresponds to Figure 4.15.

	Task	Truth	"Positive class" for ROC analysis	"Negative class" for ROC analysis	Breast region(s)	Radiologist BPE rating AUC ± SE	Computer BPE score AUC ± SE
8	Distinguishing, using BPE, cancer-affected breast at diagnosis vs cancer-affected breast at first prior	Prior or diagnostic scan	Affected breast on diagnostic scan of patients with known cancer (n = 9)	Affected breast on first prior scan of patients that developed future cancer (n = 8)	Affected	0.540 ± 0.155 (p = 0.795)	0.615 ± 0.115 (p = 0.316)
9	Distinguishing, using BPE, cancer-affected breast at diagnosis vs. unaffected breast at first prior	Prior or diagnostic scan	Affected breast on diagnostic scan of patients with known cancer (n = 9)	Unaffected breast on first prior scan of patients that developed future cancer (n = 8)	Single	0.540 ± 0.155 (p = 0.795)	0.576 ± 0.114 (p = 0.504)
10	Distinguishing, using BPE of first prior, breast side with future cancer vs. contralateral unaffected breast	Affected or unaffected sides	Affected breast on first prior scan of patients that developed future cancer (n = 8)	Unaffected breast on first prior scan of patients that developed future cancer (n = 8)	Single	0.500 ± 0.006 (p = 0.999)	0.501 ± 0.120 (p = 0.993)
11	Distinguishing, using BPE of diagnostic scan, breast with cancer vs. contralateral unaffected breast	Affected or unaffected sides	Affected breast on diagnostic scan of patients with known cancer (n = 9)	Unaffected breast on diagnostic scan of patients with known cancer (n = 9)	Single	0.500 ± 0.001 (p = 0.999)	0.546 ± 0.135 (p = 0.734)
12	Distinguishing, using BPE of scans of patients that developed cancer, first prior vs. diagnostic scan	Prior or diagnostic scan	Diagnostic scan of patients with known cancer (n = 9)	First prior scan of patients that developed cancer (n = 8)	Both	0.540 ± 0.155 (p = 0.795)	0.540 ± 0.080 (p = 0.612)
13	Distinguishing, using BPE of negative/benign patients, first prior vs. diagnostic scan	Prior or diagnostic scan	Diagnostic scan of negative/ benign patients (n = 303)	First prior scan of negative/ benign patients (n = 170)	Both	0.518 ± 0.029 (p = 0.548)	0.542 ± 0.015 (p = 6.14e-3) *
14	Distinguishing, using BPE of all scans available, patients that developed cancer vs. negative/ benign patients	Clinical diagnosis from diagnostic scan	Diagnostic and first prior scans of patients that developed cancer (n = 14)	Diagnostic and first prior scans of patients that were negative/ benign (n = 340)	Both	0.500 ± 0.068 (p = 0.999)	0.763 ± 0.031 (p = 6.54e-17) *

4.6 Discussion

In this chapter, we presented an independent validation of the BPE scoring algorithm developed in Chapter 3 using a new dataset of high-risk screening DCE-MRIs, and we reproduced the

findings of the BPE scoring algorithm applied to the original dataset. The computer BPE scores calculated on the new high-risk screening dataset failed to show a statistically significant difference from the computer BPE scores calculated on the prior dataset. The computer BPE scores of the new dataset showed statistically significantly positive correlation with the radiologist BPE ratings and performed statistically significantly greater than random guessing in many of the binary BPE level classification tasks, although the small number of cases that developed cancer limited the statistical significance of those results in some groups. Including the computer BPE scores calculated for all breasts from all scans available, and from the first prior scans only, there was a statistically significant difference between the computer BPE scores of patients that developed cancer and the computer BPE scores of negative or benign patients for the minimal or mild radiologist BPE ratings; this indicates the potential for computer BPE to differentiate a future cancer diagnosis from negative or benign diagnosis in patients with low BPE. We also assessed the effect of magnet strength on BPE for the high-risk screening dataset. Over the course of the dataset time period, there was a prominent increase in the use of 3.0T versus 1.5T MRI. A more consistent distribution of radiologist BPE ratings was assigned for each magnet strength than in the initial dataset, and we failed to show a statistically significant difference between the computer BPE scores calculated for each magnet strength.

The dataset in this chapter included first prior and subsequent diagnostic scans for 177 patients, so we investigated the change in radiologist BPE ratings and computer BPE scores from the prior to diagnostic scans. The results failed to show a statistically significant difference between the average trends for cancer and negative or benign patients, although there was a slight increase in the average trend for the cancer cases more than the negative or benign cases. The role

of computer BPE in predicting cancer versus non-cancer was assessed in a number of exploratory binary classification tasks. The BPE levels of both breast regions in the prior scan would be the most clinically useful for predicting a screening patient's future risk of cancer. We found that the computer BPE scores of the prior scan performed statistically significantly greater than random guessing in distinguishing cancer versus non-cancer scans, and the radiologist BPE ratings failed to perform statistically significantly greater than random guessing in distinguishing cancer versus non-cancer scans. Although further validation of the BPE scoring technique is required, these exploratory results demonstrate the potential of computer BPE scores to predict a future cancer from prior scans, and further, to be a useful addition to radiologist BPE ratings for breast cancer risk assessment.

4.7 Limitations and Future Work

The analyses performed in this chapter were most notably limited by the small number of patients that were known to develop cancer over the course of their screening MRIs performed at our institution; this restricted our exploratory ROC analysis to be evaluated against random guessing. Also, we were limited to two scans per patient at most, which we considered the diagnostic and first prior scan. In the future, similar investigations should be performed on a larger dataset, including more prior scans per patient and more patients that were known to develop cancer. In addition to strengthening the statistical power of the results, the investigations would be more informative regarding the potential differences in how BPE levels change over time for patients that develop cancer compared to those who remain negative or have benign diagnoses. Additionally, the use of computer BPE scores calculated from the region containing both breasts, in contrast to including computer BPE scores from individual breasts, should be assessed in cancer

versus non-cancer classification tasks. Results of those investigations could provide information on whether computer and radiologist BPE assessments would be more useful based on both breasts or based on individual breasts.

The results of this chapter were limited by a few technical aspects of the BPE scoring algorithm. For the electronic removal of the lesions prior to the BPE score computation, the algorithm required the identification of the lesion on the radiologist report and, further, the existence of the associated FCM segmentation. In some patients, there may have been a previously identified benign lesion that remained stable and was unreported on the scan being evaluated in our study, thus some lesions may still be present in the regions used for calculating computer BPE scores. Additionally, the breast segmentation was limited. In two cases, the mask for the left breast was empty or restricted to a very small region; neither was due to prior mastectomy. In the future, the segmentation performance could be improved to produce more accurate breast masks. The masks could potentially exclude major vasculature, and ideally, the segmentation algorithm would include lesion segmentation from the breast regions without the previously identified FCM locations. Future investigations that include larger datasets would reveal more information about the potential role for computer BPE scores in predicting cancer and could therefore motivate computer BPE scores to be combined with radiologist BPE ratings and other known risk factors in a risk assessment model.

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

In summary, the novel contribution of this work has been the development of a machine learning algorithm for computer background parenchymal enhancement (BPE) assessment that includes the electronic removal of lesions. This work is particularly important in regard to breast cancer screening, especially for high-risk patients, and is motivated by the need for a deeper understanding of how breast magnetic resonance imaging (MRI) can be used for diagnosis and risk assessment in order to develop useful artificial intelligence (AI) systems. A crucial component of AI systems is proper segmentation of lesions and other breast regions before extraction of new quantitative values for clinically significant quantities.

In Chapter 1, we reviewed the development of breast imaging technologies and the advancement of computational power that have contributed to an evolution of AI in breast cancer screening practices. AI methods continue to be developed with the aim to improve the efficacy and efficiency of image interpretation. Many breast cancer AI systems are based on human-engineered or deep learning methods, and such AI systems serve as concurrent or secondary readers to the radiologist. Future advances in AI will allow the radiologists' role to be enhanced as they integrate multi-modality computer outputs with medical findings, however various challenges need to be addressed, including explainability, repeatability, and generalizability. Overall, we reviewed the growth of AI methods applied to breast cancer screening with a focus on providing a significant clinical benefit. In addition to aiding radiologist decision-making, various techniques for detection, diagnosis, and risk assessment can contribute to improved breast health care management.

In Chapter 2, we investigated the performance of various methods for segmentation of breast lesions and breasts from DCE-MRI. Under the conditions where a radiologist truth was unavailable to use for lesion segmentation performance evaluations, we employed a well-established FCM clustering approach to evaluate the U-Net CNN results. The results of the investigations demonstrated that using a 2D U-Net to yield quasi-3D segmentation of breast lesions from post-contrast subtraction DCE-MRIs statistically significantly outperformed a 3D U-Net, relative to the FCM reference. Additionally, compared to the radiologist, the 2D U-Net statistically significantly outperformed the 3D U-Net and FCM, thus, 2D U-Net could be an effective alternative to more complex segmentation techniques. Future investigations of U-Nets applied to breast cancer screening may be focused on identifying lesions from the whole breast, rather than from a pre-defined region of interest. Additionally, investigations of breast segmentation may be expanded upon by using a more complex U-Net, potentially for volumetric breast segmentation, or for identifying fibroglandular tissue and vasculature. Each of these segmentation techniques could then be integrated within an algorithm similar to the one developed in the following chapter.

In Chapter 3, we evaluated radiologist and computer background parenchymal enhancement assessment on DCE-MRI. We developed an automatic computer BPE scoring method that mimics radiologist visual assessment of BPE and includes electronic removal of lesion enhancement. We investigated computer BPE scores calculated from different breast regions and evaluated the effect of varying image types and magnet strengths on the computer BPE scores. The results of the investigations demonstrated performances statistically significantly greater than random guessing using the computer BPE scores in BPE level classification tasks across various viewing projections and DCE timepoints. We found a statistically significant positive correlation

between the computer BPE scores and the qualitative, radiologist-reported BI-RADS ratings of BPE; future investigations could include the correlation of the computer BPE scores with qualitative BI-RADS ratings of fibroglandular tissue (FGT), quantitative values calculated for FGT, mammographic breast density, and mammographic radiomic features. Although our method for BPE scoring accounted for some of the potential variations between manufacturers by including image rescaling, our images were acquired with a consistent protocol, so the robustness against other manufacturers or alternative protocols (e.g., abbreviated MRI) needs to be evaluated. Additionally, the MRI signal for different contrast agents may vary due to magnet strength or the hormonal activity of the patient; the influence of those factors on the BPE scoring technique should be evaluated.

In Chapter 4, the role of BPE in predicting breast cancer was explored. An independent validation of the BPE scoring algorithm developed in Chapter 3 was performed on an independent dataset of high-risk patients that underwent screening MRI. We reproduced the findings of our BPE scoring algorithm produced in the original dataset of diagnostic patients on an independent dataset of high-risk screening patients. On the independent dataset, the computer BPE scores demonstrated statistically significant positive correlations with radiologist BPE ratings and the computer BPE scores performed statistically significantly greater than random guessing in BPE level classification tasks. Although our high-risk screening dataset included a limited number of patients that were known to develop cancer over the course of screening, in patients with low radiologist BPE ratings there was a statistically significant difference between the computer BPE scores of patients that developed cancer and those of non-cancer patients. The results of our exploratory ROC analyses showed the potential for computer BPE scores to predict future cancer

from first prior scans. Further evaluation of the change in BPE over the course of more than two scans, and for a larger number of patients that developed cancer, would produce more convincing results. Future investigations involving enriched datasets with more cases with cancer outcomes would expand our understanding of the role that the computer BPE scores can have in predicting cancer. Further, the computer BPE scores could be combined with other quantitative and qualitative image-based values, along with known risk factors (e.g., genetic mutation status, history of breast disease, etc.) to potentially improve the predictive value of clinical risk models in the future.

In conclusion, this dissertation demonstrated the potential of AI in breast cancer screening, specifically for DCE-MRI, including lesion segmentation, background parenchymal enhancement evaluation, and prediction of future breast cancer. Ultimately, this work has the potential to encourage future incorporation of quantitative breast image analysis into the clinical workflow for radiologists and to improve patient care.

REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA Cancer J Clin*, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/CAAC.21708.

[3] M. B. Mainiero *et al.*, "ACR Appropriateness Criteria: Breast Cancer Screening," 2017. [Online]. Available: https://acsearch.acr.org/docs/70910/narrative/

[4] World Health Organization (WHO), "WHO position paper on mammography screening," 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK269545/pdf/Bookshelf_NBK269545.pdf

[5] D. Saslow *et al.*, "American Cancer Society Guidelines for Breast Screening with MRI as an Adjunct to Mammography," *CA Cancer J Clin*, vol. 57, no. 2, pp. 75–89, Mar. 2007, doi: 10.3322/canjclin.57.2.75.

[6] H. Shahid, J. Wiedenhoefer, C. Dornbluth, P. Otto, and K. Kist, "An overview of breast MRI," *Appl Radiol*, vol. 45, no. 10, pp. 7–13, Oct. 2016.

[7] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Comparison of Breast MRI Tumor Classification Using Human-Engineered Radiomics, Transfer Learning from Deep Convolutional Neural Networks, and Fusion Methods," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 163–177, 2020, doi: 10.1109/JPROC.2019.2950187.

[8] W. L. Bi *et al.*, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA Cancer J Clin*, pp. 1–31, Feb. 2019, doi: 10.3322/caac.21552.

[9] D. Sheth and M. L. Giger, "Artificial intelligence in the interpretation of breast cancer on MRI," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 5, pp. 1310–1324, 2020, doi: 10.1002/jmri.26878.

[10] C. H. Lee *et al.*, "Breast Cancer Screening with Imaging: Recommendations From the Society of Breast Imaging and the ACR on the Use of Mammography, Breast MRI, Breast Ultrasound, and Other Technologies for the Detection of Clinically Occult Breast Cancer," *Journal of the American College of Radiology*, vol. 7, no. 1, pp. 18–27, Jan. 2010, doi: 10.1016/j.jacr.2009.09.022.

[11] Y. Gao and S. L. Heller, "Abbreviated and Ultrafast Breast MRI in Clinical Practice," *RadioGraphics*, vol. 40, no. 6, pp. 1507–1527, Oct. 2020, doi: 10.1148/rg.2020200006.

[12] Y. Jiang, A. V. Edwards, and G. M. Newstead, “Artificial Intelligence Applied to Breast MRI for Improved Diagnosis,” *Radiology*, vol. 298, no. 1, pp. 38–46, Jan. 2021, doi: 10.1148/radiol.2020200292.

[13] Y. Gao, B. Reig, L. Heacock, D. L. Bennett, S. L. Heller, and L. Moy, “Magnetic Resonance Imaging in Screening of Breast Cancer,” *Radiol Clin North Am*, vol. 59, no. 1, pp. 85–98, Jan. 2021, doi: 10.1016/j.rcl.2020.09.004.

[14] E. Amir, O. C. Freedman, B. Seruga, and D. G. Evans, “Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models,” *JNCI Journal of the National Cancer Institute*, vol. 102, no. 10, pp. 680–691, May 2010, doi: 10.1093/jnci/djq088.

[15] M. H. Gail *et al.*, “Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually,” *JNCI Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–1886, Dec. 1989, doi: 10.1093/jnci/81.24.1879.

[16] E. Claus, N. Risch, and W. Thompson, “Genetic analysis of breast cancer in the cancer and steroid hormone study,” *Am J Hum Genet.*, vol. 48, no. 2, pp. 232–242, 1991.

[17] G. Parmigiani, D. A. Berry, and O. Aguilar, “Determining Carrier Probabilities for Breast Cancer–Susceptibility Genes BRCA1 and BRCA2,” *The American Journal of Human Genetics*, vol. 62, no. 1, pp. 145–158, Jan. 1998, doi: 10.1086/301670.

[18] J. Tyrer, S. W. Duffy, and J. Cuzick, “A breast cancer prediction model incorporating familial and personal risk factors,” *Stat Med*, vol. 23, no. 7, pp. 1111–1130, Apr. 2004, doi: 10.1002/sim.1668.

[19] *Code of Federal Regulations Title 21 - Mammography Quality Standards Act*. [Online]. Available: https://www.ecfr.gov/current/title-21/chapter-I/subchapter-I/part-900

[20] *Mammography Quality Standards Act Final Rule*. Food and Drug Administration, 2023.

[21] C. D’Orsi, E. Sickles, E. Mendelson, E. Morris, and E. Al., “ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System.” American College of Radiology, Reston, VA, 2013.

[22] D. Saslow *et al.*, “American Cancer Society Guidelines for Breast Screening with MRI as an Adjunct to Mammography,” *CA Cancer J Clin*, vol. 57, no. 2, pp. 75–89, Mar. 2007, doi: 10.3322/canjclin.57.2.75.

[23] R. M. Mann, N. Cho, and L. Moy, “Breast MRI: State of the Art,” *Radiology*, vol. 292, no. 3, pp. 520–536, Sep. 2019, doi: 10.1148/radiol.2019182947.

[24] J. H. Youk, E. J. Son, and J.-A. Kim, “Factors Influencing the Background Parenchymal Enhancement in Follow-Up Breast MRI after Adjuvant Endocrine Therapy,” *Investig Magn Reson Imaging*, vol. 19, no. 2, p. 99, 2015, doi: 10.13104/imri.2015.19.2.99.

[25] R. Rakow-Penner, B. Daniel, H. Yu, A. Sawyer-Glover, and G. H. Glover, “Relaxation times of breast tissue at 1.5T and 3T measured using IDEAL,” *Journal of Magnetic Resonance Imaging*, vol. 23, no. 1, pp. 87–91, Jan. 2006, doi: 10.1002/jmri.20469.

[26] H. Rahbar, S. C. Partridge, W. B. DeMartini, B. Thursten, and C. D. Lehman, “Clinical and technical considerations for high quality breast MRI at 3 tesla,” *Journal of Magnetic Resonance Imaging*, vol. 37, no. 4, pp. 778–790, Apr. 2013, doi: 10.1002/jmri.23834.

[27] B. Erguvan-Dogan, G. J. Whitman, A. C. Kushwaha, M. J. Phelps, and P. J. Dempsey, “BI-RADS-MRI: a primer.,” *AJR Am J Roentgenol*, vol. 187, no. 2, pp. W152-60, 2006, doi: 10.2214/AJR.05.0572.

[28] C. S. Giess, E. D. Yeh, S. Raza, and R. L. Birdwell, “Background parenchymal enhancement at breast MR imaging: Normal patterns, diagnostic challenges, and potential for false-positive and false-negative interpretation,” *Radiographics*, vol. 34, no. 1, pp. 234–247, 2014, doi: 10.1148/rg.341135034.

[29] G. J. Liao *et al.*, “Background parenchymal enhancement on breast MRI: A comprehensive review,” *Journal of Magnetic Resonance Imaging*, vol. 51, no. 1, pp. 43–61, Jan. 2020, doi: 10.1002/jmri.26762.

[30] V. King, J. D. Brooks, J. L. Bernstein, A. S. Reiner, M. C. Pike, and E. A. Morris, “Background Parenchymal Enhancement at Breast MR Imaging and Breast Cancer Risk,” *Radiology*, vol. 260, no. 1, pp. 50–60, Jul. 2011, doi: 10.1148/radiol.11102156.

[31] D. Sheth and H. Abe, “Abbreviated MRI and Accelerated MRI for Screening and Diagnosis of Breast Cancer,” *Topics in Magnetic Resonance Imaging*, vol. 26, no. 5, pp. 183–189, Oct. 2017, doi: 10.1097/RMR.0000000000000140.

[32] D. Leithner, L. Moy, E. A. Morris, M. A. Marino, T. H. Helbich, and K. Pinker, “Abbreviated MRI of the Breast: Does It Provide Value?,” *Journal of Magnetic Resonance Imaging*, vol. 49, no. 7, pp. e85–e100, Jun. 2019, doi: 10.1002/jmri.26291.

[33] A. R. Mootz, A. J. Madhuranthakam, and B. Dogan, “Changing Paradigms in Breast Cancer Screening: Abbreviated Breast MRI,” *Eur J Breast Health*, vol. 15, no. 1, pp. 1–6, Jan. 2019, doi: 10.5152/ejbh.2018.4402.

[34] M. L. Giger, “Machine Learning in Medical Imaging,” *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 512–520, Mar. 2018, doi: 10.1016/j.jacr.2017.12.028.

[35] M. L. Giger, “Future Perspectives: CAD to Quantitative Image Biomarkers, Phenotypes, and Imaging Genomics,” in *Computer-Aided Detection and Diagnosis in Medical Imaging*, Q. Li and R. M. Nishikawa, Eds., 1st Edition.2015, pp. 409–415.

[36] H. Li and M. L. Giger, “Breast cancer,” in *Radiomics and Radiogenomics Technical Basis and Clinical Applications*, R. Li, L. Xing, S. Napel, and D. L. Rubin, Eds., CRC Press, 2019, pp. 229–249.

[37] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, “Artificial intelligence in radiology,” *Nat Rev Cancer*, vol. 18, no. 8, pp. 500–510, Aug. 2018, doi: 10.1038/s41568-018-0016-5.

[38] I. El Naqa, M. A. Haider, M. L. Giger, and R. K. Ten Haken, “Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century,” *Br J Radiol*, vol. 93, no. 1106, Feb. 2020, doi: 10.1259/bjr.20190855.

[39] H.-P. Chan *et al.*, “Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience,” *Radiology*, vol. 237, no. 3, pp. 1075–1080, Dec. 2005, doi: 10.1148/radiol.2373041657.

[40] I. Reiser *et al.*, “Computerized mass detection for digital breast tomosynthesis directly from the projection images,” *Med Phys*, vol. 33, no. 2, pp. 482–491, Jan. 2006, doi: 10.1118/1.2163390.

[41] I. Reiser *et al.*, “Automated detection of microcalcification clusters for digital breast tomosynthesis using projection data only: A preliminary study,” *Med Phys*, vol. 35, no. 4, pp. 1486–1493, Mar. 2008, doi: 10.1118/1.2885366.

[42] K. Drukker, M. L. Giger, and C. E. Metz, “Robustness of computerized lesion detection and classification scheme across different breast US platforms,” *Radiology*, vol. 237, no. 3, pp. 834–840, Dec. 2005, doi: 10.1148/radiol.2373041418.

[43] B. Sahiner *et al.*, “Malignant and benign breast masses on 3D US volumetric images: Effect of computer-aided diagnosis on radiologist accuracy,” *Radiology*, vol. 242, no. 3, pp. 716–724, Mar. 2007, doi: 10.1148/radiol.2423051464.

[44] K. G. A. Gilhuijs, M. L. Giger, and U. Bick, “Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging,” *Med Phys*, vol. 25, no. 9, pp. 1647–1654, Sep. 1998, doi: 10.1118/1.598345.

[45] W. Chen, M. L. Giger, and U. Bick, “A Fuzzy C-Means (FCM)-Based Approach for Computerized Segmentation of Breast Lesions in Dynamic Contrast-Enhanced MR Images1,” *Acad Radiol*, vol. 13, no. 1, pp. 63–72, Jan. 2006, doi: 10.1016/j.acra.2005.08.035.

[46] P. Gibbs and L. W. Turnbull, “Textural analysis of contrast-enhanced MR images of the breast,” *Magn Reson Med*, vol. 50, no. 1, pp. 92–98, Jul. 2003, doi: 10.1002/mrm.10496.

[47] W. Chen, M. L. Giger, H. Li, U. Bick, and G. M. Newstead, "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magn Reson Med*, vol. 58, no. 3, pp. 562–571, Sep. 2007, doi: 10.1002/mrm.21347.

[48] G. Ertaş, H. Ö. Gülçür, and M. Tunacı, "An interactive dynamic analysis and decision support software for MR mammography," *Computerized Medical Imaging and Graphics*, vol. 32, no. 4, pp. 284–293, Jun. 2008, doi: 10.1016/j.compmedimag.2008.01.004.

[49] N. Bhooshan, M. L. Giger, S. A. Jansen, H. Li, L. Lan, and G. M. Newstead, "Cancerous Breast Lesions on Dynamic Contrast-enhanced MR Images: Computerized Characterization for Image-based Prognostic Markers," *Radiology*, vol. 254, no. 3, pp. 680–690, Mar. 2010, doi: 10.1148/radiol.09090838.

[50] W. Chen, M. L. Giger, U. Bick, and G. M. Newstead, "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," *Med Phys*, vol. 33, no. 8, pp. 2878–2887, Jul. 2006, doi: 10.1118/1.2210568.

[51] K. D. Kurz *et al.*, "Assessment of three different software systems in the evaluation of dynamic MRI of the breast," *Eur J Radiol*, vol. 69, no. 2, pp. 300–307, Feb. 2009, doi: 10.1016/j.ejrad.2007.10.003.

[52] M. Samulski, R. Hupse, C. Boetes, R. D. M. M. Mus, G. J. den Heeten, and N. Karssemeijer, "Using computer-aided detection in mammography as a decision support," *Eur Radiol*, vol. 20, no. 10, pp. 2323–2330, Oct. 2010.

[53] R. Hupse *et al.*, "Computer-aided detection of masses at mammography: Interactive decision support versus prompts," *Radiology*, vol. 266, no. 1, pp. 123–129, Jan. 2013, doi: 10.1148/radiol.12120218.

[54] M. L. Giger, N. Karssemeijer, and J. A. Schnabel, "Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer," *Annu Rev Biomed Eng*, vol. 15, no. 1, pp. 327–357, Jul. 2013, doi: 10.1146/annurev-bioeng-071812-152416.

[55] N. Baughan, L. Douglas, and M. L. Giger, "Past, Present, and Future of Machine Learning and Artificial Intelligence for Breast Cancer Screening," *J Breast Imaging*, vol. 4, no. 5, pp. 451–459, 2022, doi: 10.1093/jbi/wbac052.

[56] M. L. Giger, H. P. Chan, and J. Boone, "Anniversary paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM," *Med Phys*, vol. 35, no. 12, pp. 5799–5820, 2008, doi: 10.1118/1.3013555.

[57] I. Sechopoulos and R. M. Mann, "Stand-alone artificial intelligence - The future of breast cancer screening?," *The Breast*, vol. 49, pp. 254–260, Feb. 2020, doi: 10.1016/J.BREAST.2019.12.014.

[58] K. J. Geras, R. M. Mann, and L. Moy, “Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives,” *Radiology*, vol. 293, no. 2, pp. 246–259, Nov. 2019, doi: 10.1148/radiol.2019182627.

[59] Y. Jiang, M. F. Inciardi, A. V. Edwards, and J. Papaioannou, “Interpretation Time Using a Concurrent-Read Computer-Aided Detection System for Automated Breast Ultrasound in Breast Cancer Screening of Women With Dense Breast Tissue,” *American Journal of Roentgenology*, vol. 211, no. 2, pp. 452–461, Aug. 2018, doi: 10.2214/AJR.18.19516.

[60] J. C. M. van Zelst *et al.*, “Dedicated computer-aided detection software for automated 3D breast ultrasound; an efficient tool for the radiologist in supplemental screening of women with dense breasts,” *Eur Radiol*, vol. 28, no. 7, pp. 2996–3006, Jul. 2018, doi: 10.1007/s00330-017-5280-3.

[61] M. U. Dalmış, S. Vreemann, T. Kooi, R. M. Mann, N. Karssemeijer, and A. Gubern-Mérida, “Fully automated detection of breast cancer in screening MRI using convolutional neural networks,” *Journal of Medical Imaging*, vol. 5, no. 01, p. 1, Jan. 2018, doi: 10.1117/1.JMI.5.1.014502.

[62] F. Ayatollahi, S. B. Shokouhi, R. M. Mann, and J. Teuwen, “Automatic breast lesion detection in ultrafast DCE-MRI using deep learning,” *Med Phys*, vol. 48, no. 10, pp. 5897–5907, Oct. 2021, doi: 10.1002/mp.15156.

[63] M. U. Dalmiş *et al.*, “Artificial Intelligence–Based Classification of Breast Lesions Imaged With a Multiparametric Breast MRI Protocol With Ultrafast DCE-MRI, T2, and DWI,” *Invest Radiol*, vol. 54, no. 6, pp. 325–332, Jun. 2019, doi: 10.1097/RLI.0000000000000544.

[64] N. Antropova, B. Q. Huynh, and M. L. Giger, “A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets,” *Med Phys*, vol. 44, no. 10, pp. 5162–5171, Oct. 2017, doi: 10.1002/MP.12453.

[65] B. Q. Huynh, H. Li, and M. L. Giger, “Digital mammographic tumor classification using transfer learning from deep convolutional neural networks,” *Journal of Medical Imaging*, vol. 3, no. 3, p. 034501, Aug. 2016, doi: 10.1117/1.JMI.3.3.034501.

[66] N. Antropova, H. Abe, and M. L. Giger, “Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks,” *Journal of Medical Imaging*, vol. 5, no. 01, p. 1, Feb. 2018, doi: 10.1117/1.JMI.5.1.014503.

[67] Q. Hu, H. M. Whitney, and M. L. Giger, “A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI,” *Sci Rep*, vol. 10, no. 1, p. 10536, Dec. 2020, doi: 10.1038/s41598-020-67441-4.

[68] Q. Hu, H. M. Whitney, and M. L. Giger, "Radiomics methodology for breast cancer diagnosis using multiparametric magnetic resonance imaging," *Journal of Medical Imaging*, vol. 7, no. 04, Aug. 2020, doi: 10.1117/1.JMI.7.4.044502.

[69] "FDA Cleared AI Algorithms," *AI Central*. https://aicentral.acrdsi.org/ (accessed Mar. 24, 2022).

[70] S. M. Mckinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 2, p. 89, 2020, doi: 10.1038/s41586-019-1799-6.

[71] Y. Shen *et al.*, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med Image Anal*, vol. 68, p. 101908, Feb. 2021, doi: 10.1016/J.MEDIA.2020.101908.

[72] R. Masud, M. Al-Rei, and C. Lokker, "Computer-aided detection for breast cancer screening in clinical settings: Scoping review," *JMIR Med Inform*, vol. 7, no. 3, pp. 1–10, 2019, doi: 10.2196/12660.

[73] *Code of Federal Regulations Title 21 - Diagnostic Devices*. [Online]. Available: https://www.ecfr.gov/current/title-21/chapter-I/subchapter-H/part-892/subpart-B/section-892.2080

[74] A. Yala, T. Schuster, R. Miles, R. Barzilay, and C. Lehman, "A Deep Learning Model to Triage Screening Mammograms: A Simulation Study," *Radiology*, vol. 293, no. 1, pp. 38–46, Oct. 2019, doi: 10.1148/radiol.2019182908.

[75] S. E. Hickman, G. C. Baxter, and F. J. Gilbert, "Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations," *Br J Cancer*, vol. 125, no. 1, pp. 15–22, 2021, doi: 10.1038/s41416-021-01333-w.

[76] A. D. Lauritzen *et al.*, "An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload," *Radiology*, Apr. 2022, doi: 10.1148/radiol.210948.

[77] H. M. Whitney and M. L. Giger, "Artificial Intelligence in Medical Imaging," in *Quantitative Imaging in Medicine*, R. J. Nordstrom, Ed., AIP Publishing, 2021, pp. 7.1-7.22. doi: 10.1063/9780735423473_007.

[78] J. D. Fuhrman, N. Gorre, Q. Hu, H. Li, I. El Naqa, and M. L. Giger, "A review of explainable and interpretable AI with applications in COVID-19 imaging," *Med Phys*, vol. 49, no. 1, pp. 1–14, Jan. 2022, doi: 10.1002/mp.15359.

[79] M. L. Giger, "AI/Machine Learning in Medical Imaging," in *Molecular Imaging Principles and Practice (Second Edition)*, B. D. Ross and S. Sam Gambhir, Eds., Elsevier, 2021, pp. 1691–1702. doi: 10.1016/B978-0-12-816386-3.00052-1.

[80] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Harmonization of radiomic features of breast lesions across international DCE-MRI datasets," *Journal of Medical Imaging*, vol. 7, no. 01, p. 1, 2020, doi: 10.1117/1.jmi.7.1.012707.

[81] M. McNitt-Gray *et al.*, "Standardization in quantitative imaging: A multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets," *Tomography*, vol. 6, no. 2, pp. 118–128, Jun. 2020, doi: 10.18383/j.tom.2019.00031.

[82] P. Amstutz, K. Drukker, H. Li, H. Abe, M. L. Giger, and H. M. Whitney, "Case-based diagnostic classification repeatability using radiomic features extracted from full-field digital mammography images of breast lesions," in *Medical Imaging 2021: Computer-Aided Diagnosis*, K. Drukker and M. A. Mazurowski, Eds., SPIE, Feb. 2021. doi: 10.1117/12.2580743.

[83] "The Medical Imaging and Data Resouce Center (MIDRC)," 2020. https://www.midrc.org/

[84] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging 2013 26:6*, vol. 26, no. 6, pp. 1045–1057, Jul. 2013, doi: 10.1007/S10278-013-9622-7.

[85] I. El Naqa *et al.*, "AI in medical physics: guidelines for publication," *Med Phys*, vol. 48, no. 9, pp. 4711–4714, 2021, doi: 10.1002/mp.15170.

[86] N. Bhooshan, M. L. Giger, S. A. Jansen, H. Li, L. Lan, and G. M. Newstead, "Cancerous Breast Lesions on Dynamic Contrast-enhanced MR Images: Computerized Characterization for Image-based Prognostic Markers," *Radiology*, vol. 254, no. 3, pp. 680–690, Mar. 2010, doi: 10.1148/radiol.09090838.

[87] M. U. Dalmış *et al.*, "Using deep learning to segment breast and fibroglandular tissue in MRI volumes," *Med Phys*, vol. 44, no. 2, pp. 533–546, Feb. 2017, doi: 10.1002/mp.12079.

[88] M. L. Giger, N. Karssemeijer, and J. A. Schnabel, "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer," *Annual Review of Biomedical Engineering*, vol. 15. pp. 327–357, Jul. 2013. doi: 10.1146/annurev-bioeng-071812-152416.

[89] D. Wei *et al.*, "Fully automatic quantification of fibroglandular tissue and background parenchymal enhancement with accurate implementation for axial and sagittal breast MRI protocols," *Med Phys*, vol. 48, no. 1, pp. 238–252, Jan. 2021, doi: 10.1002/mp.14581.

[90] Y. Zhang *et al.*, "Automatic Breast and Fibroglandular Tissue Segmentation in Breast MRI Using Deep Learning by a Fully-Convolutional Residual Neural Network U-Net," *Acad Radiol*, vol. 26, no. 11, pp. 1526–1535, Nov. 2019, doi: 10.1016/j.acra.2019.01.012.

[91] W. Chen, M. L. Giger, and U. Bick, “A Fuzzy C-Means (FCM)-Based Approach for Computerized Segmentation of Breast Lesions in Dynamic Contrast-Enhanced MR Images,” *Acad Radiol*, vol. 13, no. 1, pp. 63–72, Jan. 2006, doi: 10.1016/j.acra.2005.08.035.

[92] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[93] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-net: Learning dense volumetric segmentation from sparse annotation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9901 LNCS, pp. 424–432, Jun. 2016, doi: 10.1007/978-3-319-46723-8_49.

[94] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species,” *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945, doi: 10.2307/1932409.

[95] F. Hausdorff, *Grundzüge der Mengenlehre (“Basics of Set Theory”)*. Leipzig Viet, 1914.

[96] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, vol. 1, no. 6, p. 80, Dec. 1945, doi: 10.2307/3001968.

[97] H. B. Mann and D. R. Whitney, “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other,” *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, Mar. 1947, doi: 10.1214/aoms/1177730491.

[98] C. E. Bonferroni, “Teoria statistica delle classi e calcolo delle probabilità (Statistical class theory and probability calculus),” *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[99] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” 2018, doi: 10.48550/arXiv.1804.03999.

[100] D. L. Lam, D. S. Hippe, A. E. Kitsch, S. C. Partridge, and H. Rahbar, “Assessment of Quantitative Magnetic Resonance Imaging Background Parenchymal Enhancement Parameters to Improve Determination of Individual Breast Cancer Risk,” *J Comput Assist Tomogr*, vol. 43, no. 1, pp. 85–92, 2019, doi: 10.1097/RCT.0000000000000774.

[101] G. P. Watt *et al.*, “Association of breast cancer with MRI background parenchymal enhancement: the IMAGINE case-control study,” *Breast Cancer Research*, vol. 22, no. 1, p. 138, Dec. 2020, doi: 10.1186/s13058-020-01375-7.

[102] B. N. Dontchos *et al.*, “Are Qualitative Assessments of Background Parenchymal Enhancement, Amount of Fibroglandular Tissue on MR Images, and Mammographic Density Associated with Breast Cancer Risk?,” *Radiology*, vol. 276, no. 2, pp. 371–380, Aug. 2015, doi: 10.1148/radiol.2015142304.

[103] D. Sheth and M. L. Giger, “Artificial intelligence in the interpretation of breast cancer on MRI,” *Journal of Magnetic Resonance Imaging*, vol. 51, no. 5, pp. 1310–1324, 2020, doi: 10.1002/jmri.26878.

[104] T. Portnoi *et al.*, “Deep learning model to assess cancer risk on the basis of a breast MR image alone,” *American Journal of Roentgenology*, vol. 213, no. 1, pp. 227–233, 2019, doi: 10.2214/AJR.18.20813.

[105] A. Saha *et al.*, “Machine learning-based prediction of future breast cancer using algorithmically measured background parenchymal enhancement on high-risk screening MRI,” *Journal of Magnetic Resonance Imaging*, vol. 50, no. 2, pp. 456–464, Aug. 2019, doi: 10.1002/jmri.26636.

[106] B. L. Niell *et al.*, “Quantitative Measures of Background Parenchymal Enhancement Predict Breast Cancer Risk,” *American Journal of Roentgenology*, vol. 217, no. 1, pp. 64–75, Jul. 2021, doi: 10.2214/AJR.20.23804.

[107] M. G. Kendall, *Rank Correlation Methods*. Griffin, 1970.

[108] C. E. Metz and X. Pan, “‘Proper’ Binormal ROC Curves: Theory and Maximum-Likelihood Estimation,” *J Math Psychol*, vol. 43, no. 1, pp. 1–33, Mar. 1999, doi: 10.1006/jmps.1998.1218.

[109] S. S. Kang, E. Y. Ko, B.-K. Han, J. H. Shin, S. Y. Hahn, and E. S. Ko, “Background parenchymal enhancement on breast MRI: Influence of menstrual cycle and breast composition,” *Journal of Magnetic Resonance Imaging*, vol. 39, no. 3, pp. 526–534, Mar. 2014, doi: 10.1002/jmri.24185.

[110] J. Johnson and F. Selchick, “What to know about the Tyrer-Cuzick score,” *Medical News Today*, 2021. https://www.medicalnewstoday.com/articles/tyrer-cuzick-score

[111] S. Chen and G. Parmigiani, “Meta-Analysis of BRCA1 and BRCA2 Penetrance,” *Journal of Clinical Oncology*, vol. 25, no. 11, pp. 1329–1333, Apr. 2007, doi: 10.1200/JCO.2006.09.1066.

[112] J. Cuzick and A. Brentnall, “Models for Assessment of Breast Cancer Risk,” *DI Europe: Breast Cancer*, 2016. https://ems-trials.org/riskevaluator/documents/dieurope.pdf

[113] J. Warwick *et al.*, “Mammographic breast density refines Tyrer-Cuzick estimates of breast cancer risk in high-risk women: findings from the placebo arm of the International Breast Cancer Intervention Study I,” *Breast Cancer Research*, vol. 16, no. 5, p. 451, Oct. 2014, doi: 10.1186/s13058-014-0451-5.

[114] L. J. Grimm *et al.*, “Relationship between Background Parenchymal Enhancement on High-risk Screening MRI and Future Breast Cancer Risk,” *Acad Radiol*, vol. 26, no. 1, pp. 69–75, 2019, doi: 10.1016/j.acra.2018.03.013.

[115] S. H. Lee *et al.*, “Background Parenchymal Enhancement at Postoperative Surveillance Breast MRI: Association with Future Second Breast Cancer Risk,” *Radiology*, vol. 306, no. 1, pp. 90–99, Jan. 2023, doi: 10.1148/radiol.220440.

[116] V. A. Arasu *et al.*, “Population-based assessment of the association between magnetic resonance imaging background parenchymal enhancement and future primary breast cancer risk,” *Journal of Clinical Oncology*, vol. 37, no. 12, pp. 954–963, 2019, doi: 10.1200/JCO.18.00378.

[117] J. R. Wilkie, M. L. Giger, M. R. Chinander, C. A. Engh, R. H. Hopper, and J. M. Martell, “Temporal radiographic texture analysis in the detection of periprosthetic osteolysis,” *Med Phys*, vol. 35, no. 1, pp. 377–387, 2008, doi: 10.1118/1.2820900.

LIST OF PUBLICATIONS AND PRESENTATIONS

Peer-Reviewed Journal Papers (* indicates shared first authorship)

Baughan N*, **Douglas L***, Giger ML: Past, present, and future of machine learning and artificial intelligence for breast cancer screening. Journal of Breast Imaging 4(5): 451-459. Sept/Oct 2022

Douglas L, Bhattacharjee R, Fuhrman J, Drukker K, Hu Q, Edwards A, Sheth D, Giger ML: U-Net Breast Lesion Segmentations for Breast DCE-MRI. (under review)

Douglas L, Fuhrman J, Hu Q, Edwards A, Sheth D, Abe H, Giger ML: Computerized assessment of breast parenchymal enhancement on DCE-MRI including electronic lesion removal. (under review)

Douglas L, Edwards A, Abe H, Giger ML: Independent validation of machine learning technique for computer background parenchymal enhancement assessment for high-risk dataset of breast DCE-MRIs. (in draft)

Conference Papers

Douglas L, Mondal T, Edwards A, Giger ML: Influence of magnet strength on background parenchymal enhancement evaluation. Proceedings SPIE 12467: 2023

Baughan N, **Douglas L**, Ballard M, Lee ES, Edwards A, Lan L, Li H, Giger ML: Association between DCE MRI background parenchymal enhancement and mammographic texture features. Proceedings SPIE 12033: 2022

Douglas L, Sheth D, Giger ML: Electronic removal of lesions for more robust BPE scoring on breast DCE-MRI. Proceedings SPIE 11597: 2021

Bhattacharjee R, **Douglas L**, Drukker K, Hu Q, Fuhrman JD, Sheth D, Giger ML: Comparison of 2D and 3D U-Net breast lesion segmentations on DCE-MRI. Proceedings SPIE 11597: 2021

Marshall E, Joswiak C, **Douglas L**, Pearson E, Lu ZF, Al-Hallaq H, Reiser: Development of a compact inkjet-printed patient-specific phantom for optimization of fluoroscopic image quality in neonates. Proceedings SPIE 10948: 2019

Scientific Presentations

Douglas L, Mondal T, Edwards A, Giger ML: "Influence of magnet strength on background parenchymal enhancement evaluation." Oral presentation at SPIE: Medical Imaging, Feb 2023.

Douglas L: "Machine Learning for Background Parenchymal Enhancement from Breast DCE-MRI." Oral presentation at the inaugural Emerging Leaders in Academic Medical Physics Symposium and Workshop, Madison, WI, Aug 2022.

Douglas L: "Machine Learning for Background Parenchymal Enhancement from Breast DCE-MRI." Poster presentation at the inaugural Emerging Leaders in Academic Medical Physics Symposium and Workshop, Madison, WI, Aug 2022.

Douglas L, Edwards A, Abe H, Giger ML: "Effect of various DCE-MRI image parameters on AI assessment of breast parenchymal enhancement." Oral presentation at AAPM Annual Meeting, Washington, DC, July 2022.

Douglas L, Sheth D, Giger ML: "Electronic removal of lesions for more robust BPE scoring on breast DCE-MRI." Virtual oral presentation at SPIE: Medical Imaging, Feb 2021.

Marshall E, Joswiak C, **Douglas L**, Pearson E, Lu ZF, Al-Hallaq H, Reiser: "Development of a compact inkjet-printed patient-specific phantom for optimization of fluoroscopic image quality in neonates." Poster presentation at SPIE: Medical Imaging, San Diego, California, Feb 2020.

Douglas L, Marshall E, Pearson E, Al-Hallaq H, Reiser I: "Temporal Stability and Energy Dependence of a Compact Iodine Inkjet-printed Phantom." Oral presentation at the AAPM Annual Meeting, San Antonio, Texas, July 2019.

Invited Talks

Pogue B, Armato S, **Douglas L**: Invited Panelist for the Society of Directors of Academic Medical Physics Programs (SDAMPP) discussion on "Nurturing Medical Physics Research Leaders," January 2023.