THE UNIVERSITY OF CHICAGO

DEEP LEARNING AND RADIOMICS OF BREAST CANCER ON DCE-MRI IN ASSESSMENT OF MALIGNANCY AND RESPONSE TO THERAPY

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

NATALIA ANTROPOVA

CHICAGO, ILLINOIS

JUNE 2018

For My Parents and Grandparents

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My successful completion of my PhD would not be possible without the support, encouragement, and advice of many people. First of all, I want to thank my PhD advisor, Dr. Maryellen Giger, for her mentorship, support, and guidance, which were crucial to all of my accomplishments. I thank her for letting me explore promising research directions, for knowing when I needed assistance, and for providing me with necessary research freedom. She played a significant role in the growth of my scientific writing and presentation skills. Her door was always open for any advice I needed.

I am grateful to my thesis committee, including Sam Armato, Patrick LaRiviere, and Dr. Abe, without whose help the completion of my dissertation work would not be possible. I appreciate the great discussions of new ideas, their advice, and truthful and rigorous evaluation of my work. Special thank you to Dr. Abe for letting me to come and observe the patient-doctor interactions and medical procedures to better understand the clinical perspective.

I would like to say thank you to all of the members of Giger lab. Special thank you to Dr. Hui Li for the technical and emotional support from the beginning to the end of my PhD and to Dr. Karen Drukker for the technical guidance in statistics. I would like to acknowledge Sasha Edwards and John Papaioannou for tirelessly and efficiently helping me with the organization of image and clinical data. Sasha Edwards' help was essential for my education on many topics related to breast cancer imaging.

I appreciate the collaborative and friendly environment created by the entire Committee on Medical Physics at the University of Chicago, including all of the students and faculty. Special thank you to Eyjolfur Gudmundsson for the support, cheerful discussions, and the essential coffee

breaks. I would also like to express gratitude to the following people in the medical physics department, who helped me on the technical and administrative sides: Chun-Wai for the software support; Ms. Ruth Magana, Ms. Julie Hlavaty, and Ms. Loretta Powell for the administrative support; Ms. Gloria Frazier for the help with my fellowship application.

I acknowledge the support of the following grants and fellowships: The NIH Training Award (NIH T32 EB002103-25); The NIH QIN Grant (NIH QIN U01CA195564); The NIH Pre-Doctoral Research Fellowship (NIH F31 CA221193-01A1); The Paul C. Hodges Alumni Society Research Award (The University of Chicago); and Machine Learning in Healthcare Workshop Travel Award (NIPS 2017).

Finally, I dedicate this dissertation to my parents and grandparents, who always believed in me and taught me the value of education. To my family's unconditional love and support. My mom especially has always had the most significant role in what I have achieved. She made the decision to bring me to the USA and gave me the opportunity to learn and become whoever I wanted to be.

ABSTRACT

Breast cancer is found in one in eight women in the United States and is expected to be the most frequently diagnosed form of cancer among women in 2018. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) plays a significant role in breast cancer diagnosis and prognosis. The interpretation of the DCE-MR images remains labor-intensive and can lead to erroneous clinical results, associated with subsequent unnecessary biopsies and patient stress. Computer-aided detection (CADe) and diagnosis (CADx) systems, or radiomics, have been developed to help reduce these errors. The conventional radiomics methods involve automatic segmentation of a lesion from the neighboring background and extraction of intuitive features, manually designed by the scientists and domain experts. Such features describe a lesion's size, shape, texture, and enhancement patterns. The recent advances in machine learning techniques have provided an alternative method for image assessment, where images are analyzed directly by deep learning models, in fully automated mode.

Radiomics has strong potential to lead clinicians towards more accurate and rapid image interpretation. Furthermore, it can serve as a "virtual digital biopsy," allowing for discovery of relationships between radiomics and the pathology/genomics from actual biopsies. The objective of this research is to analyze the already existing methods and to design new radiomics methods for breast DCE-MRI, in order to improve image-based clinical decisions. Specifically, the first part of the dissertation analyzes the robustness of conventional radiomics methods across MRI scanners of different manufacturers. The rest of the research develops accurate and robust deep learning-based models for automated breast lesion characterization, tailored to complex 4-

dimensional (4D) DCE-MRI data. These models are applied to two clinical tasks, lesion malignancy assessment and prediction of cancer's response to therapy. The research is concluded by testing the hypothesis that incorporating the two types of radiomics, deep learning-CADx and conventional-CADx, will enhance lesion characterization within the tasks of diagnosis and treatment response.

The research presents the following results. First, the robustness analysis reveals radiomics features that are generalizable across datasets acquired with MRI scanners of two major manufacturers, GE and Philips. The features that characterize lesions in terms of size are robust in their average values. The results demonstrate that an entropy feature, which quantifies randomness of pixel values of the lesion image, is robust in its classification performance in multiple clinical tasks. Second, a novel deep learning-based method is developed to assess breast lesion malignancy and response to therapy based on the DCE-MRI sequence. The work demonstrates that the lesion representation input into the deep learning pipeline needs to be carefully designed prior to the application of the algorithms. In particular, the size of the region of interest selected around a lesion and the DCE time point on which it is selected significantly affects the deep learning-based classification performance. We further show that DCE-MRI maximum intensity projection (MIP) images incorporate clinically useful information about the entire lesion volume and partly about the contrast enhancement and can be utilized as a lesion representation. Their use improves deep learning-based lesion classification compared to the classification based on a single MRI slice. Given that MIP images do not fully utilize the sequential enhancement patterns present in the DCE-MRI sequences, we develop a method that incorporates the temporal and volumetric

components using recurrent neural networks. Finally, the results support the hypothesis that deep learning-based methods are complementary to the conventional radiomics.

The medical significance of this research is that it has potential to improve DCE-MRI-based breast cancer diagnosis and prognosis. The clinical value of DCE-MRI is continuously increasing in breast cancer management. The developed deep learning methods and their fusion with conventional radiomics can reduce human burden and allow for more rapid and accurate analysis of the breast DCE-MR images.

Keywords: breast cancer, computer-aided diagnosis, DCE-MRI, deep learning, convolutional neural networks, recurrent neural networks, robustness, ROC analysis.

CHAPTER 1

INTRODUCTION

According to the statistics accumulated by the American Cancer Society, breast cancer is found in one in eight women in the United States.[1] It is the number one cancer found in women worldwide.[1] Among North American women, it is the second most frequently diagnosed form of cancer and the second leading cause of death. In 2018, it is estimated that 266,120 new invasive breast cancer cases and 63,960 breast carcinomas *in situ* cases will be diagnosed in the United States. Furthermore, 40,920 deaths are expected to occur due to breast cancer in the same year.[5] Despite these numbers, based on the statistics collected in 2015, the mortality rates have dropped 39% below the peak rates.[2] This decrease can be attributed to early detection as well as advanced treatment techniques for breast cancer.[3,4]

X-ray mammography is the most frequently used imaging modality for breast cancer diagnostic and screening practice. It is the suggested screening modality for women with reasonably good health. Additional screening methods, such as ultrasound, digital breast tomosynthesis (DBT), and magnetic resonance imaging (MRI), can be used for high-risk women, even though there is still not enough data to conclusively support their use.[5,6] For any imaging modality, the image interpretation can be time intensive and prone to human errors. Those can arise from such factors as readers' inexperience, physical state, large amounts of images to evaluate, and outside distractions. To increase accuracy and decrease the human burden, computer-

aided diagnosis (CADx) and detection (CADe) systems, or radiomics, have been developing to aid radiologists in image assessment.[7,8] The work presented in this dissertation aims to develop and analyze CADx methods for breast dynamic contrast-enhanced magnetic resonance images (DCE-MRIs).

The introduction of this dissertation starts with an overview of the clinical breast imaging modalities, with the emphasis on the role of DCE-MR imaging in breast cancer screening and diagnostic workup. It further presents the history and state-of-the-art of CADx and CADe methods and discusses their applications for breast MR imaging and gives an overview of deep learning methods. The introduction concludes with the statement of the goals and the outline of the dissertation research.

1.1. Breast Cancer Imaging

Breast cancer affects a significant part of the world's population. Thanks to the advanced imaging techniques that lead to early disease detection and to the effective treatment methods, breast cancer mortality rates have been drastically decreasing.

Distinct methods of breast cancer screening are performed for women of different age groups and of different breast cancer risk associations. One of the methods utilized is clinical breast examination, which plays an important role in the checkup of younger women. It is done by a trained medical personnel by visual inspection and palpation of the entire breast and is recommended to be performed every 3 years for women of ages between 20 and 30.[3]

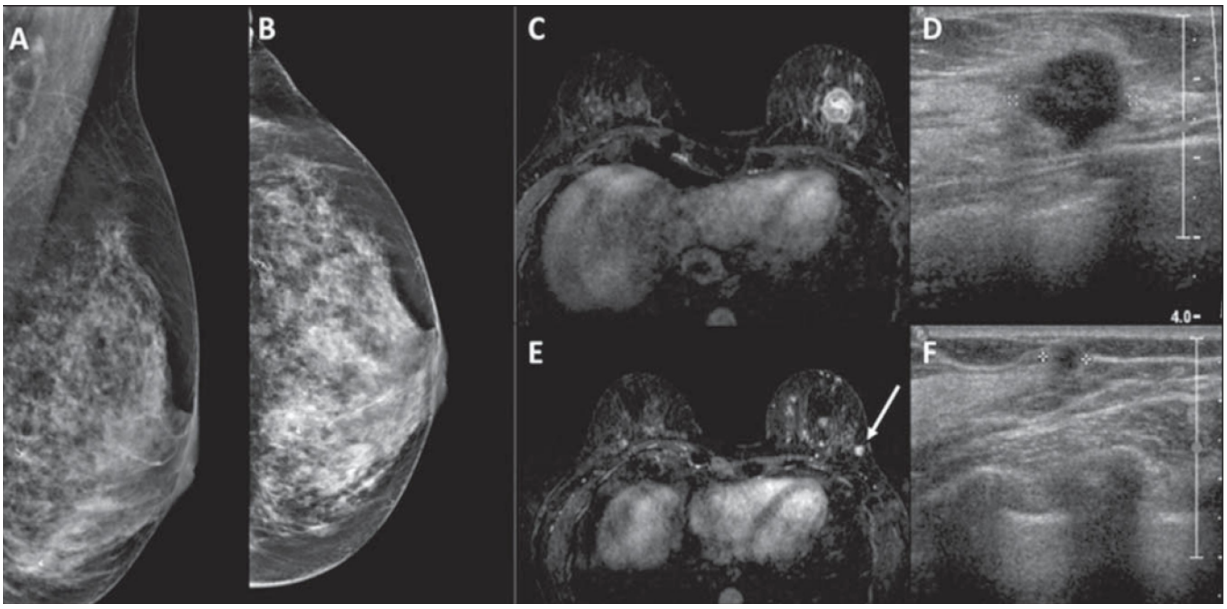


Figure 1.1. Examples of large and small breast lesions imaged with mammography (A, B), MRI (C, E), and ultrasound (D, F).

X-ray mammography is the recommended and most common imaging modality for breast cancer screening (Figure 1.1).[9] It has been shown to reduce mortality rates in women of ages between 40 and 74.[5] Various agencies have different recommendation on breast cancer screening with mammography. The U.S. Preventive Services Task Force (USPSTF) recommends starting mammography screening at the age of 40. Women of 50 years of age and older should perform biennial screening. Younger women are suggested to perform screening with frequency based on the trade-offs between the harm of X-ray radiation from mammography and their individual risk of having breast cancer.[5] The Canadian and European agencies, The Canadian Task Force on Preventive Health Care[10] and the European Cancer Observatory[11], have slightly more conservative screening recommendations with suggested mammography imaging frequency of every two or three years for women between ages 50 and 69. The recommendations vary mainly due to the trade-offs between the number of breast cancers averted and the number of false-positive findings.

For the women younger than age 50, the number of false positives tends to be high, which often leads to unnecessary emotional stress and invasive procedures.[10]

Mammography uses low-energy X-rays of around 30kVp and produces a 2-dimensional (2D) image of a breast. In earlier years, film mammography was used, but advances in computer processing speed, data storage, and imaging resolution replaced it with digital mammography. In 2012, the Food and Drug Aministration approved 3D mammography, or tomosynthesis, and it has been adapted in the clinical workflow since then.[3] Tomosynthesis is performed in combination with 2D mammography and has been shown to increase cancer detection rates and reduce false positive calls. Mammography also remains the main modality for breast cancer diagnosis. However, diagnosis with mammography can be difficult for dense breasts. Furthermore, mammography provides a projection of the anatomical structure of the breast, which does not include any diagnostically important information about the physiology of the breast lesions.

Women with a lifetime risk of breast cancer of 20-25% and greater are recommended to receive an annual MRI screening in addition to mammography (Figure 1.1). [3,12] The level of the risk of breast cancer can be evaluated through a patient's family history, genetic testing, and clinical history. Women with breast or ovarian cancer running in their family are identified as higher-risk patients. Genetic testing can reveal mutations in two genes, BRCA1 and BRCA2, which are associated with breast and ovarian cancers. Furthermore, clinical factors that are associated with breast cancer risk include diagnosis with Hodgkins disease and prior radiation treatments.[12]

Several studies have been conducted to evaluate the effectiveness of screening MRI for the high-risk patient group and compared it with mammography. The results showed that MRI has

71% -100% sensitivity, compared to 16-40% for mammography. On the other hand, MRI has significantly lower specificity than mammography. Table 1.1 summarizes the sensitivity and specificity values obtained by three clinical studies for the following breast imaging methods - mammography alone, MRI alone, and combination of mammography and MRI. Another benefit of MRI over mammography is that it does not use ionizing radiation. However, the imaging procedure is time consuming and expensive. Furthermore, its low specificity means more false-positive detections and thus unnecessary clinical procedures and patient stress. The role of MRI in breast cancer management is further detailed in Section 1.2 of this dissertation.

Table 1.1. Sensitivity and specificity values for breast screening mammography alone, MRI alone, and combination of mammography and MRI for three distinct studies.

Study	Sensitivity			Specificity		
	Mammography	MRI	Mammography + MRI	Mammography	MRI	Mammography + MRI
529 patients[13]	33%	71%	93%	95%	90%	96%
649 patients[14]	40%	77%	94%	93%	81%	77%
1,909 patients[15]	40%	71%	NA	95%	90%	NA

Patients with lifelong risk who cannot undergo MRI procedure are recommended to have sonography/ultrasound screening in addition to mammography (Figure 1.1). Sonography also plays an important role in screening of patients with intermediate risk and who have high breast tissue density.[3,16] Breast density has been demonstrated to be a significant risk factor for breast cancer, with women with denser breast tissue having two to six times higher risk than the women

with less-dense breast tissue.[16] Several studies have shown that the use of sonography increases the breast cancer detection rate. Frequently, mammographically occult cancers can be detected with sonography, regardless of the breast tissue density. Compared to mammography, sonography does not use any ionizing radiation. Instead it uses high-frequency sound waves to produce the image of the breast. However, like with MRI, sonography results in high false-positive rates and takes a long time to perform.

Other imaging modalities, such as PET, digital infrared thermal imaging (thermography), sono-elastography, electrical impedance scanning, and optical imaging have been infrequently used for breast cancer screening and diagnosis and have been evaluated only on small patient samples.[3]

1.2. The Role of DCE-MRI in Breast Management

1.2.1. MRI Physics

MR images are generated with the application of strong magnetic fields, typically 1.5 Tesla (T) or 3T, magnetic gradients, and radiofrequency pulses, which excite hydrogen nuclei to produce a measurable signal. Hydrogen is used to produce the MR images is because it is the most abundant nucleus in a human body and thus it is able to produce the strongest signal. There are about 3.3×10^{22} water molecules in one milliliter of water, or 6.6×10^{22} hydrogen nuclei. In the absence of external magnetic field, the nuclear spins are randomly oriented. When they are placed in a strong magnetic field, they align either parallel or antiparallel with it, with a slight majority aligning

parallel and producing a ‘net’ magnetization (of approximately 1 spin per million), that can be detected and measured.

In addition to the application of a strong magnetic field, an oscillating magnetic field, or radiofrequency (RF) pulse, is applied orthogonally to the direction of the main magnetic field at the proton resonance frequency during MR image acquisition. The RF pulse is applied only for a few milliseconds, making the aligned hydrogen nuclei precess around the direction of the main magnetic field at increasing angles. This makes the total magnetization tip from its original alignment into the transverse plane of the main magnetic field. Once the RF pulse is off, the magnetization returns back to its original state. The frequency at which the nuclei are precessing, and therefore the required frequency of the RF pulse, is directly proportional to the strength of the main magnetic field. In an MRI machine, gradient coils are used to create a spatially variant main magnetic field in order to produce varying precessional frequencies in different parts of the body. When the excited hydrogen nuclei relax, they emit RF signals at a range of frequencies, which are recorded by receiving coils. These different frequencies encode the specific locations in the body from which the signal is originating. The recorded MRI signal is strongly influenced by magnetic interactions of the nuclei with their local environment in the body. Such interactions affect how quickly the hydrogen nuclei can return to the original magnetization state, which is characterized by the longitudinal relaxation time, T_1, and the transverse relaxation time, T_2. MR image tissue contrast can be changed through varying the weighting of the T_1 and T_2 relaxation times by changing the pulse repetition time (TR) and echo time (TE), respectively, in the image acquisition sequence.

1.2.2. Breast DCE-MRI

At the University of Chicago Medical Center, routine clinical breast MRI studies include axial Turbo spin echo (TSE) T2-weighted images and a dynamic contrast-enhanced study using one pre- and six post-contrast fat-saturated axial T1-weighted images. T2-weighted MR imaging is valuable in detecting water and fat signal, including cysts, dilated ducts and fat-containing masses such as hamartoma and lymph node. Since it also does not require administration of a contrast agent, it is often recommended to be performed prior to the contrast enhanced imaging.[17] However, the T2 images do not show the physiology of the lesions and might not always be clinically sufficient. Diffusion-weighted MR imaging (DWI) also carries diagnostically and prognostically useful information for breast cancer.[18,19] DWI measures the diffusion of water molecules in breast tissues with the apparent diffusion coefficient, which is recognized to be a diagnostic parameter. Evaluation of T2-weighted and DWI breast images is outside of the scope of this dissertation. The dissertation performs analysis of the breast DCE-MRIs.

The DCE-MRI procedure involves intravenous injection of a gadolinium-based contrast agent, which enhances breast tissues over time. The enhancement pattern is then observed in the DCE-MRI sequence, which consists of one MR image acquired prior to and multiple MR images acquired after the contrast injection. Gadolinium administered in small doses affects the microenvironment by reducing the T_1 relaxation time, therefore, increasing its signal in a T_1-weighted MRI acquisition.[20]

One advantage of the DCE-MRI is that it shows not only the morphological structure of the breast, but also its physiology. The contrast enhancement of the lesions and the surrounding parenchyma are different, due to the difference in the vascular and capillary permeability of

these tissues. This allows for easier visual and computerized discrimination of the lesion and surrounding tissue. DCE-MRI allows for visualization of spatial and temporal variations of lesion angiogenesis. It is presented in the time-signal intensity curve, or kinetic curve, which carries some of the most diagnostically useful information. The kinetic curve characterizes the uptake and washout patterns of the contrast agent by the tissue (Figure 1.2). The enhancement patterns are classified into three categories: Type I is a progressive enhancement pattern, Type II is a plateau pattern, and Type III is a washout pattern. Type I enhancement typically retains increasing signal intensity over time, in both the initial and delayed enhancement phase. Tumors with Type I kinetic curves are considered to be mostly benign, with only about 9% of them being malignant. Type II pattern is characterized by plateauing enhancement signal after the initial rise.[21,22] Type II enhancements are often indicative of tumor malignancy. Finally, Type III enhancement has decrease contrast enhancement after the initial rise and is a strong indicator for lesion malignancy.

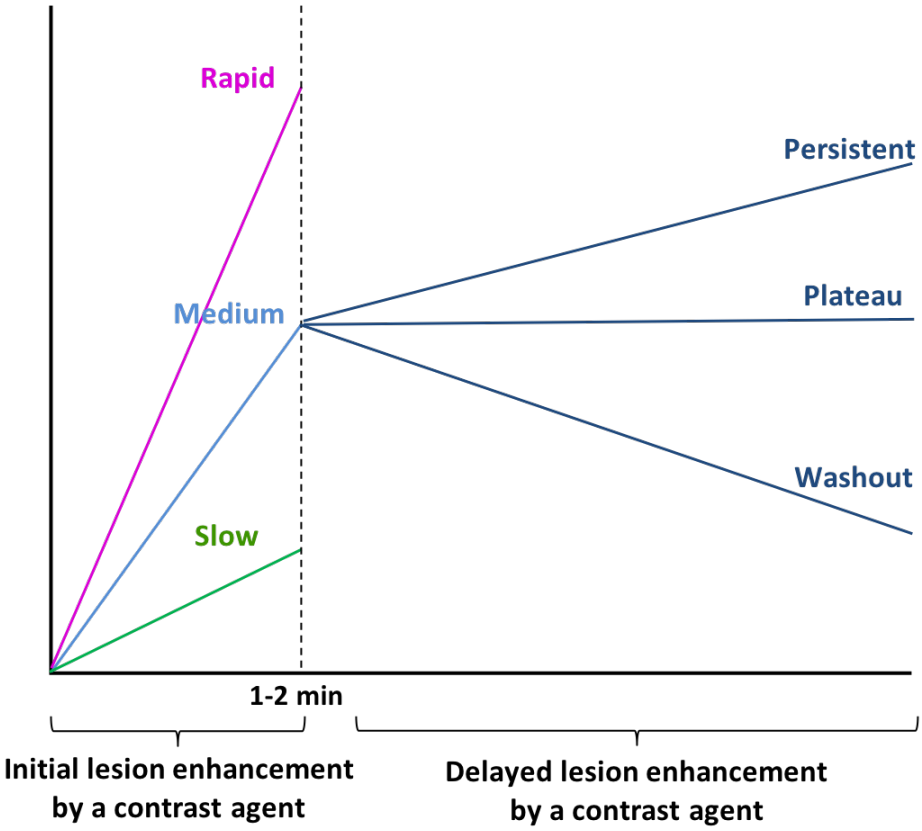


Figure 1.2. Contrast enhancement patters observed in breast DCE-MRI. Based on these enhancements, breast lesions are characterized as Type I, Type II, or Type III.

As described in Section 1.2.1, breast MRI is currently recommended for breast cancer screening of women in high-risk groups and as a follow-up exam for inconclusive findings with mammography. It has superior sensitivity to screening mammography, but lacks specificity and requires long scanning times. These conclusions were derived mainly based on the datasets acquired on 1.5T scanners.[13–15] Since then, it has been shown that it is possible to decrease false-positive rates of DCE-MRI by imaging patients with scanners that use higher magnetic field strengths (sensitivity increases from 92.8% to 94.5%).[23,24] Another study proposed using abbreviated protocols, or first post-contrast subtracted (FAST), to decrease image acquisition and evaluation times.[25] In the abbreviated protocol, only one pre-contrast and one post-contrast image

are acquired during the scan and are subsequently used to produce maximum intensity projection images (MIPs). Even though the study failed to demonstrate decrease in false positive rates, the proposed protocol requires only 3 minutes for image acquisition and 3 seconds for MIP image evaluation. More recent studies have shown supporting results for the utility of FAST in decreasing the scanning time.[26]

Furthermore, an ultrafast DCE-MRI sequence has been proposed to improve breast cancer diagnosis. [27,28] The sequence involves acquiring images at a high temporal resolution of 6.6 – 6.9 seconds in the first minute after the contrast injection, followed by the regular image sequence with high spatial and lower temporal resolutions. The proposed ultrafast technique has potential for aiding radiologists in lesion discrimination and is under further investigation.

Besides its diagnostic utility, DCE-MRI is suitable for breast cancer preoperative staging,[29,30] delineating the disease extent, and determining whether a cancer patient is responding or not responding to neoadjuvant chemotherapy.[31,32] For the patients undergoing neoadjuvant chemotherapy, it is important to evaluate as early as possible whether the patient is responding to the prescribed therapy regimen, to avoid potentially harmful effects and unnecessary costs. Therefore, MR imaging is usually performed to monitor early signs of treatment response. Following the neoadjuvant chemotherapy, MRI is utilized to identify possible residual disease.[30]

1.3. Computer-aided Diagnosis and Detection

Despite the advanced imaging techniques available in clinics, expert image interpretation remains time consuming, prone to human error, and, sometimes, not available. Computer-aided detection (CADe) and diagnosis (CADx) systems, also termed radiomics, have been developing since the mid-1980s to assist radiologists to make better clinical judjements.[7,8] CADe systems aim to localize abnormalities in medical images or suggest suspicious regions to the medical experts. On the other hand, CADx systems are designed to perform diagnostic and prognostic classification tasks. Those classify previously-identified lesions and suspicious regions and serve as a supplemental opinion to the clinicians.

The first CAD methods were developed for analysis of chest radiographs and breast mammograms. Since then, successful automated image analysis was performed on various imaging modalities, including mammography, computed tomography (CT), MRI, and PET, and for various diseases, such as breast, lung, colon, and prostate cancers, osteoporosis, cerebrovascular disease, diabetic retinopathy, interstitial disease, and many more.[33–36] The goal of these automated analysis systems is to reduce human errors, intra- and inter-reader variations, and evaluation times, and to make medical image interpretation more accessible.

Currently studied diagnostic radiomics systems can be separated into two types. We refer to one type as the conventional CADx and to another as the deep learning-based CADx. The conventional CADx has been around since the start of CADx development. It involves automatic segmentation of a lesion from the neighboring background and extraction of intuitive hand-crafted features, carefully developed by the scientists and domain experts over the years. These features are then used to train classification models for various clinical questions. Figure 1.3 presents a

schematic of the breast DCE-MRI CADx, developed at the Giger Lab at the University of Chicago. It extracts features that describe the lesion's size, shape, texture, and enhancement patterns.[37,38,40,41]

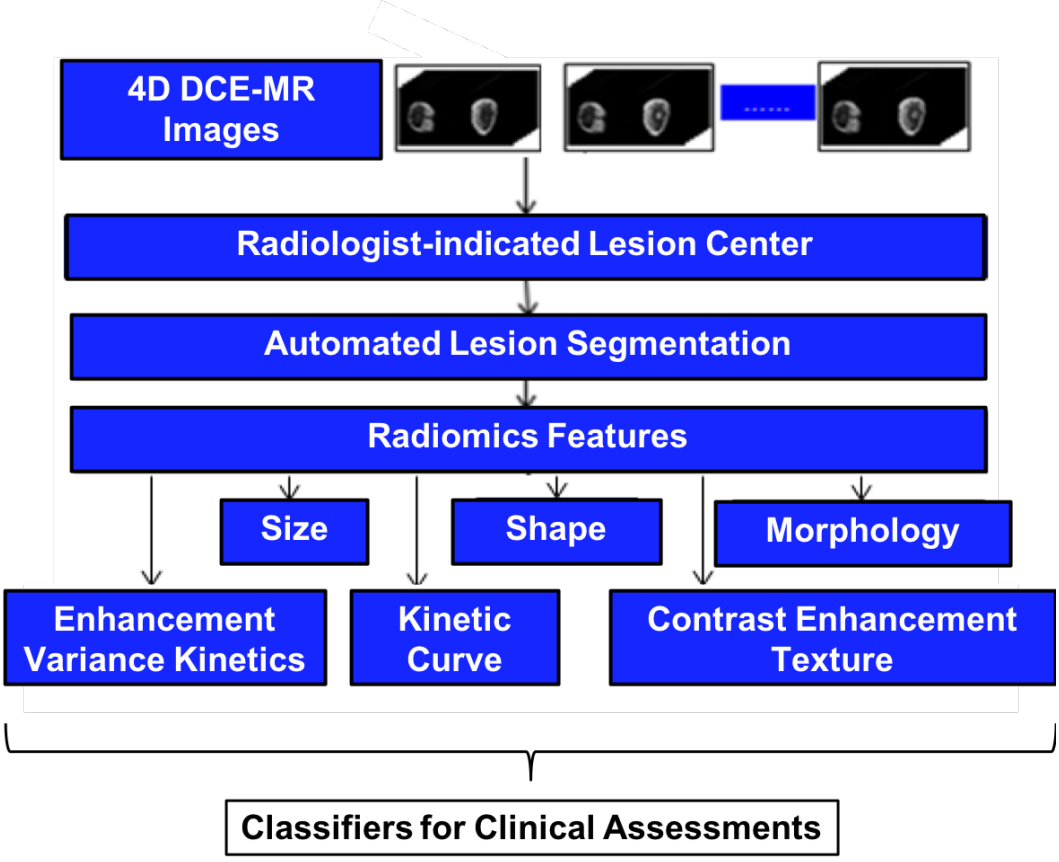


Figure 1.3. Breast DCE-MRI CADx. A DCE-MR image of a breast lesion gets input into a workstation, which first automatically segments a lesion based on a previously radiologist-indicated lesion center. From the segmented lesion, the system then extracted pre-defined lesion features.

Recent advances in machine learning methods, coupled with the increasing use of GPUs for faster, parallelized processing have provided an alternative method for medical image assessment, where images are analyzed by deep learning models using an end-to-end automated process. Some of the most powerful models in deep learning-based image analysis are convolutional neural networks (CNN). These have been applied to various visual tasks, including

medical, since the late 1970s. However, they gained substantial attention in all areas of research only recently, when efficient training of the networks became possible. Specifically, GPU-based ImageNet dataset classification with CNN AlexNet provided researchers with ability to efficiently and reliably train their models and compare the results using standardized representative benchmarks.[42,43] Since then, these methods have been rapidly adapted in automated medical image analysis research.[44]

Deep learning methods have several advantages over the conventional CADx. They eliminate the need for manual and labor-intensive design of conventional hand-crafted features and are able to analyze large volumes of image data very quickly. Furthermore, deep learning models are able to learn abstract image representation, sometimes not accessible to a human mind. The last point is both an advantage and a shortcoming of deep learning models. Even though these models see images in very complex ways, they lack human interpretability, which is often desirable for the clinicians and the patients. Deep learning methods for medical imaging have other shortcomings as well. Training of accurate and robust deep learning models requires large amounts of well-annotated data, which is often not available in the medical domain. Furthermore, medical imaging data is often high-dimensional. Specific models have to be designed to capture clinically useful information present in the entire image. Such models can be extremely difficult to train and their training and evaluation time might be lengthy.

Nevertheless, both conventional and deep learning-based CAD methods have been shown to significantly improve medical image assessment. However, they require careful design and robust evaluation prior to their application. Besides the development of radiomics methods themselves, a crucial step in the radiomics pipeline is careful preparation of clinical and image

data. Robust analytical models and well-prepared data are the essential components to practical radiomics.

1.4. Machine Learning

Machine learning is increasingly used to power various research areas. It has been applied to tasks ranging from smile detection in photographs to genomics data analysis. Machine learning provides us with the ability to augment the knowledge and efforts of the domain experts with a tool that can help us to analyze data, extract meaning from it, and, as a result, to make better predictions and decisions.

Machine learning is commonly divided into two general forms. Unsupervised learning is concerned with understanding the internal structures of the data in a fully automated manner. Unsupervised learning algorithms work with unlabeled data. Supervised learning, on the contrary, requires the training data to be manually labeled. The algorithms learn from the labeled training data to predict the correct label for the unlabeled test data.

In the case of conventional CADx, machine learning models are limited in their ability to process data in their original form and require their manual preprocessing. These models are applied to the pre-designed radiomic features computed from medical images. Therefore, the conventional CADx process is not end-to-end, requiring interventional stops along the classification pipeline.

1.4.1. Deep Learning Models

The machine learning field is currently going through a period of explosive development. The modern world is generating huge amounts of data and computational power is now more accessible and cheap, allowing researchers to develop more powerful solutions for automated data analysis much faster and with less effort. One of the most important factors in the ongoing machine learning renaissance is deep learning technology. Compared to the conventional machine learning methods, which require considerable research and design effort to capture higher level features of the data, deep learning models learn data representations at various levels of abstraction automatically. Examples of such models are convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Convolutional neural networks are computational models that are typically applied to data with repeated local structure, such as image and video data.[45] They have been state-of-the-art methods for image and video classification, image segmentation, and image generation tasks.[46–50] Recurrent neural networks are designed to work with sequential data. They learn the sequential dependencies and use that knowledge to make predictions or carry out classification. RNNs have been state-of-the-art models in speech recognition, image captioning, text classification, and text translation tasks.[51,52]

1.4.2. Convolutional Neural Networks

Convolutional neural networks are the essential models for image detection and classification tasks. Earlier neural network models included fully-connected multilayer neural networks, which can be used for both learning representation features and classifying data.

However, it is computationally unrealistic to apply a fully-connected network to images of any meaningful size. Instead, with convolutional neural networks, an input image is broken up into a grid of smaller pieces of identical size. Each of those is then convolved with the same input weights matrix, called filter or kernel (Figure 1.4). It is important to note that the meaning of the term "convolution" for neural networks is different than in mathematics and physics. The filters are then used to scan the full image, piece by piece, in the vertical and horizontal dimensions, making the convolutional layer. Following each convolution operation, CNNs apply an activation function to each of the output units in the feature maps. Those activations are typically rectified linear unit (ReLU) activations, but can also be *tanh* or sigmoid operations.[45,53] A typical convolution layer involves application of a few such filters, which together produce a tensor of feature maps, or an activation volume. The dimension of such a tensor is *width* x *height* x *depth*, with the depth dimension corresponding to the number of filters producing the current layer. An input to a convolutional layer can be not only an image, but also the feature maps produce by earlier layers of the CNN. In the CNNs, the filters' weight matrices are learned from the provided training data by applying an algorithm called backpropagation. [54]

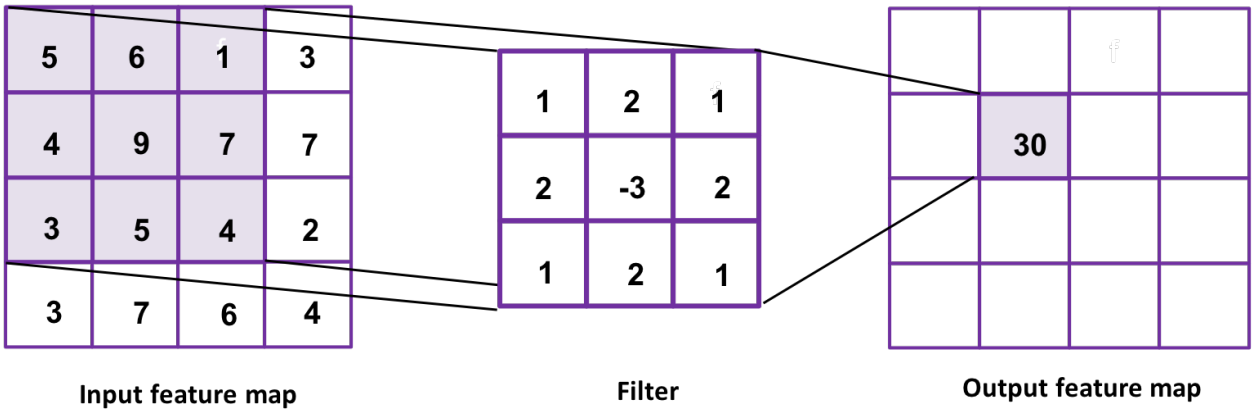


Figure 1.4. Convolution operation at one of the locations of the input feature map. It performs a dot product operation of the region in the feature map with the convolutional filter.

Pooling layers are another essential part of CNNs. The pooling operation downsamples the input feature maps (Figure 1.5). The common pooling operations are max-pooling and average-pooling. Max-pooling takes a maximum value from a specified region, while average-pooling outputs the average of all of the values in that region. Since the pooling operation is performed in the height and width dimensions, it does not change the depth dimension.

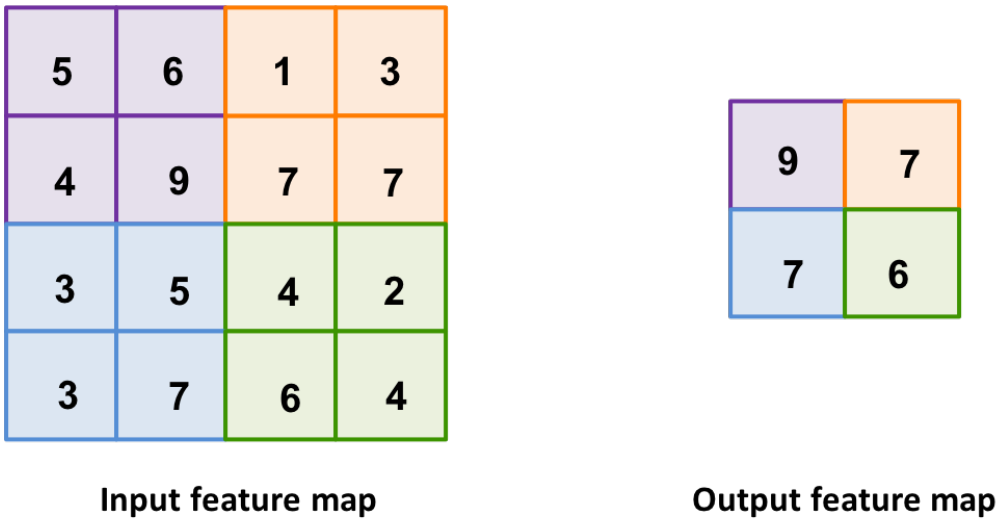


Figure 1.5. Pooling operation with filter size 2x2 and stride 2. It down samples the input volume spatially, preserving the depth dimension.

After the convolutional and pooling layers have projected input data into the abstract representations, fully-connected layers are used to complete the classification task. Those layers have every unit in the previous layer connected to every unit in the next layer (Figure 1.6). The weights of the fully-connected layer are also found with the backpropagation algorithm.

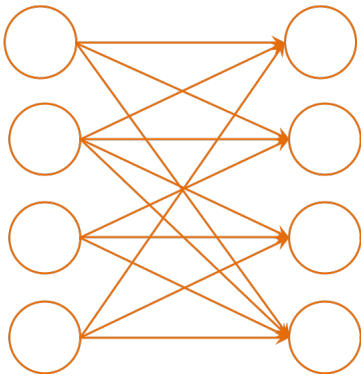


Figure 1.6. Fully-connected layer of a neural network. Each unit in the input layer is connected to each unit in the output layer.

There are many other techniques, such as dropout, skip connections, batch normalization, and many more, that are applied in various CNN architectures.[45,55]

1.4.3. Application of CNNs to Medical Images

In medical image analysis, CNNs are applied for segmentation, classification, image generation, image reconstruction, and image registration tasks. Often, CNN architectures are designed to segment various organs or diseases in 2D and 3D medical images, for detection tasks, or as the first step in quantitative image analysis. The first and biggest contribution of deep learning methods in medical image analysis was to classify regions of interest in medical images. There exist two common ways of applying CNNs to medical image classification. First, a CNN architecture can be designed and trained 'from scratch' for the medical task. This method has a large limitation. The existing successful models for natural image analysis have a number of parameters on the order of a million. The available natural image datasets are also extremely large, making it possible to train CNN models that are highly accurate and robust. On the other hand, the medical domain has a lack of image data necessary for CNN training. Researchers have overcome

this problem by applying various data augmentation techniques to increase their dataset sizes or by utilizing transfer learning techniques.

In transfer learning, CNNs previously trained on large natural or medical image datasets are applied to a medical dataset of a limited size. There are commonly used architectures that have been proven to achieve very high performances and those have been frequently used in transfer learning. These CNNs are pre-trained on ImageNet, a large natural image dataset, which contains over a million of images of a thousand categories, such as cars, cats, horses, chairs, etc. The pre-trained CNNs can be either fine-tuned or use feature extractors from the small-sized medical dataset. In both cases, the weights obtained by training the network on the large dataset are used to initialize the weights of the network used for the medical image analysis. Earlier layers of the networks are responsible for features that are common to many types of images, such as shapes, gradients, and edges. This is used in fine-tuning, where the early layers of the CNN are kept frozen, while later layers get updated during training with the new data. The fine-tuned CNN then makes classifications of the new data. When a pre-trained CNN is used as a feature extractor, an input image is passed through the network with the initialized weights. This process produces an image representation by extracting features from the input image. Those features can be further used to build classifiers for the clinical task of interest.

1.5. Research objectives and scope of dissertation

As described in the introduction, the role of DCE-MRI for breast cancer diagnosis and management continues to grow. The imaging modality presents breast tissue in high resolution and carries information about its physiology. DCE-MRI continues advancing towards becoming a

faster and more accessible technique for breast cancer evaluation. However, the challenges with DCE-MRI assessment remain. The conventional DCE-MRI radiomics methods have been developed for breast lesion segmentation, feature extraction, and lesion characterization and have been successfully applied for various clinical tasks, such as predicting cancer recurrence and assessing breast cancer subtype.[37–39,56–58] The rise of deep learning methods gives a new way to automatically evaluate medical images. Those methods have not yet been developed for breast DCE-MRI. Thus, the work presented in this dissertation evaluates the conventional CADx for unstudied clinical tasks and develops new deep learning-based methods for breast DCE-MRI.

Specifically, Chapter 2 studies the robustness of the conventional CADx features[37–39,58] across the scanners of two major MRI manufacturers, GE and Philips. The robustness is studied across two data samples in terms of feature values and in terms of feature classification performance for lymph node and hormone receptor statuses.

Chapter 3 develops novel deep learning-based methods, tailored to the breast DCE-MR image complexities. It compares the utility of the conventional and deep learning-based CADx to that of their combination. The results are presented for the tasks of breast lesion malignancy assessment and prediction of breast cancer response to neoadjuvant chemotherapy.

Chapter 4 develops a more complex deep learning-based pipeline for DCE-MR image classification. The work of Chapter 3 does not work with the entire DCE-MRI sequence. Instead, it chooses the best lesion representation from the 4D MR data. This chapter presents a method that allows incorporation of the temporal component of DCE-MRI into lesion classification.

Finally, chapter 5 summarizes the results of this dissertation research and proposes further research directions.

CHAPTER 2

ROBUSTNESS OF CONVENTIONAL HAND-CRAFTED RADIOMIC FEATURES DERIVED FROM BREAST DCE-MRIS

This section of the dissertation presents robustness analysis of the conventional hand-crafted radiomic features, derived from breast DCE-MRIs, across two major MRI scanner manufacturers. The robustness of the features is studied in terms of their average values and in terms of their predictive performance of breast cancer's hormone receptor and lymph node statuses. The evaluations are performed across two data samples collected with the two different MRI scanners.

2.1. Introduction

As discussed in Chapter 1, DCE-MRI plays a significant role in the assessment of breast cancer. DCE-MRI's radiomics methods have the potential to significantly improve image interpretation by reducing human errors and reading time. However, many challenges have to be solved before prognostic radiomics can be applied in a clinical setting. One of those challenges is variability of image data, which may arise from variations in scanning protocols, MR system manufactures, and magnet strengths. With sufficient training and experience, human readers may adjust their interpretation, while computational radiomics may be dependent on the differences in image acquisitions. For computerized image analysis to be clinically useful, radiomic systems need to generate consistent results when they analyze images acquired at different conditions. Such consistency may be achieved by standardization of image data prior to radiomic feature extraction

or through the harmonization of features themselves. To ensure robustness of automated analysis, variation in current radiomic features needs to be investigated as a basis for further exploration of data harmonization. The work of Chapter 2 performs robustness analysis of the conventional hand-crafted radiomic features derived from DCE-MRIs acquired on scanners of two manufacturers. Out of the three major MR manufacturers, General Electric (GE), Philips, and Siemens, only GE and Philips DCE-MRI radiomics are evaluated, due to lack of sufficient amount of available data from the Siemens scanner.

2.2. Clinical Tasks

The robustness of the breast DCE-MRI radiomics was evaluated in terms of their utility relative to four prognostic questions, which included lymph node status and hormone receptor status assessment. The tasks included characterization of each breast lesion into lymph node positive or negative (LN+ vs. LN-), estrogen-receptor positive or negative (ER+ vs. ER-), progesterone-receptor positive or negative (PR+ vs. PR-), and human epidermal growth factor receptor 2 positive or negative (HER2+ vs. HER2-). The pathology data were obtained from biopsy and subsequent clinical pathology reports.

2.2.1. Lymph Node Status

Lymph nodes are a part of the lymphatic system and consist of packed immune cells that protect organs from viruses, bacteria, and cancer cells. Lymph nodes are located throughout the human body and, if enlarged or swollen, can be good indicators of clinical problems of nearby organs. If breast cancer spreads outside of the primary lesion in the breast, it often targets the axillary nodes

underarm. Positive lymph node status (LN+) indicates the spread of cancer cells out of the lesion to the lymph nodes. Negative lymph node status (LN-) indicates absence of cancer cells in the lymph nodes. The lymph node status is an important clinical factor that determines breast cancer prognosis and guides its therapy.[59]

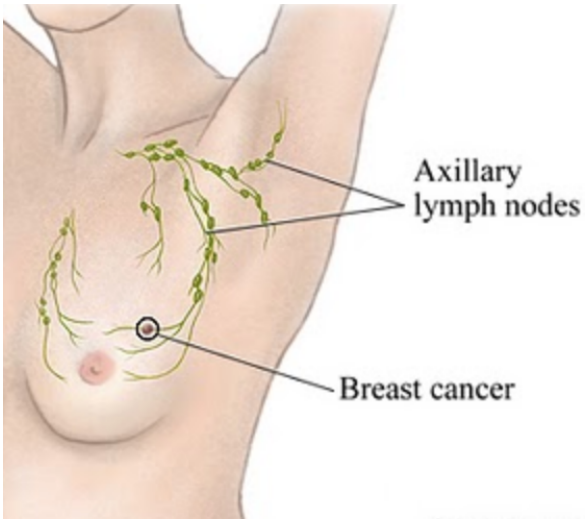


Figure 2.1. Schematic showing a lesion in the breast as well as the axillary lymph nodes near it.

2.2.2. Estrogen and Progesterone Hormone Receptor Status

Estrogen and progesterone receptors are proteins found inside cells and are responsible for binding with estrogen and progesterone hormones. Both of the hormones are central to cell growth. Breast cancers with ERs (PRs) found in their cells are called ER+ (PR+). ER- (PR-) breast cancers do not have ERs (PRs) in their cells. Knowledge of the hormone receptor status helps doctors decide what kind of treatment to prescribe to the cancer patient. Hormone therapies are given to patients with positive hormone status to slow down or stop cancer growth, by either lowering the amount of estrogen/progesterone or blocking it from the cells.[60]

2.2.3. Human Epidermal Growth Factor Receptor 2

HER2 is a protein receptor that promotes cancer cell growth. The protein is overexpressed in 20-30% of breast cells, making the cells grow and divide uncontrollably.[61] HER2+ breast cancers have excess of HER2 and are associated with higher risk of recurrence, aggressiveness, and immortality rates and are more likely to spread in the body outside of the primary lesion location. Special drugs designed specifically for HER+ breast cancer patients have been shown to be very effective in its treatment.[61]

2.3. DCE-MRI Datasets

Two DCE-MRI datasets were used in the robustness study. Patient data were collected under HIPPA-compliant protocols approved by Institutional Review Boards (IRB). Both datasets included only primary invasive breast cancer tumors (Table 2.1).

2.3.1. Dataset A – GE MRI Scanner

Dataset A was collected at four institutions - Memorial Sloan Kettering Cancer Center, Mayo Clinic, University of Pittsburgh Medical Center, and Roswell Park Cancer Institute. Images were acquired on 1.5 Tesla GE (GE Medical Systems, Milwaukee, Wisconsin, USA) scanners over a four-year period, 1999 - 2002. The clinical information for Dataset A had been retrospectively downloaded from the Cancer Genome Atlas (TCGA),[62] a dataset of de-identified cancer cases collected from the four US cancer centers using TCGA-Assembler.[63] The corresponding available MR images, matched to the TCGA cases, were obtained from The Cancer Imaging Archive (TCIA).[64] Out of the available data, we focused on 91 breast MRI cases, which had similar

manufacturers and imaging protocols. These breast cancer cases were previously used in image radio-genomic research conducted in the Giger lab.[56,57,65-67] All cases included one pre- and three to five post-contrast MRIs acquired with a T1-weighted 3D spoiled gradient-echo sequence. Gadolinium-based contrast agent was used for all of the acquisitions. The in-plane resolution of the images was between 0.53 mm and 0.86 mm.

2.3.2. Dataset B – Philips MRI Scanner

Dataset B was retrospectively collected at the University of Chicago Medical Center and included 332 invasive breast cancer cases, acquired on a 1.5 Tesla Philips MRI scanner (Philips Healthcare, Amsterdam, The Netherlands).[58] The clinical data were obtained from the pathology reports retrieved from the Center for Research Informatics Clinical Research Data Warehouse (CRI CRDW). From the dataset of 332 cases, pathology reports yielded 255, 266, 266, and 239 cases with known lymph node, PR, ER, and HER2 statuses. For all the cases, Gadodiamide (Omniscan, GE Healthcare) was used as a contrast agent. The images were acquired with T1-weighted 3D gradient echo sequence. The in-plane resolution of the images was 0.74 mm.

2.3.3. Dataset Distributions based on Lesion Size

Cases in Dataset A and Dataset B were separated into six categories based on lesion size and pathology: small positive, small negative, medium positive, medium negative, large positive and large negative (Table 2.1). This was separately performed for each of the four clinical tasks. TNM Classification of Malignant Tumors (TNM)[68] criteria were followed to categorize breast lesion size. Lesions were labeled as small if they had diameter <2 cm, medium if they had diameter of 2-5 cm, and large if they had diameter >5 cm.

Table 2.1. Total number of cases available in Dataset A (GE) and Dataset B (Philips). Cases are separated based on four clinical questions and three size categories.

Clinical Question/ Size Categories		Dataset A	Dataset B
LN+	Small positive	14	27
	Medium positive	11	24
	Large positive	19	39
	Total	*44 (48.9%)*	*90 (35.3%)*
LN-	Small negative	19	75
	Medium negative	19	38
	Large negative	8	52
	Total	*46 (51.1%)*	*165 (64.7%)*
Total for LN		90	255
ER+	Small positive	33	89
	Medium positive	26	46
	Large positive	18	49
	Total	*77 (84.6%)*	*184 (69.2%)*
ER-	Small negative	1	26
	Medium negative	4	19
	Large negative	9	37
	Total	*14 (15.4%)*	*82 (30.8%)*
Total for ER		91	266
PR+	Small positive	29	72
	Medium positive	25	37
	Large positive	18	38
	Total	*72 (79.1%)*	*147 (55.3%)*
PR-	Small negative	5	41
	Medium negative	5	27
	Large negative	9	51
	Total	*19 (20.9%)*	*119 (44.7%)*
Total for PR		91	266
HER2+	Small positive	6	18
	Medium positive	7	9
	Large positive	6	23
	Total	*19 (20.9%)*	*50 (20.9%)*
HER2-	Small negative	28	86
	Medium negative	23	48
	Large negative	21	55
	Total	*72 (79.1%)*	*189 (79.1%)*
Total for HER2		91	239

2.4. DCE-MRI Radiomic Features

DCE-MR images from Dataset A and Dataset B underwent quantitative radiomic analysis on the University of Chicago breast MRI radiomics workstation. Automatic segmentation of lesions from the DCE-MRI images was performed with a Fuzzy C-means (FCM) clustering-based algorithm.[40]

Based on the lesion segmentations, the workstation calculated 38 radiomic features for each lesion. These image features can be divided into six categories: size, shape, morphology, enhancement texture, kinetic curve assessments, and enhancement-variance kinetics. *Size* and *shape* features describe exclusively the geometry of the segmented lesion.[69] *Size* features include volume, effective diameter, surface area, and maximum linear size of the lesion. *Shape* features describe how irregular the shape of the lesion is and how similar it is to a sphere. *Morphological* descriptors account for the lesion's marginal enhancement, as well as the irregularity of its shape.[37,69,70] *Enhancement texture* descriptors portray the spatial and enhancement properties of the contrast-enhanced lesion on the first post-contrast MR image.[37] The last two groups of features, *kinetic curve assessments* and *enhancement-variance kinetics*, are extracted from the characteristic kinetic curve that shows intensity (lesion enhancement) variation over time.[41] The characteristic curve of the lesion is constructed from the output of Fuzzy C-means clustering of the kinetic curves. *Kinetic curve assessment* phenotypes describe the shape of the most-enhancing characteristic kinetic curve and represent the properties of uptake and washout of the contrast agent over the time period of dynamic imaging. *Enhancement-variance kinetics* features describe how the spatial enhancement of the lesion varies over imaging time.[38,41] The detailed description of individual features is provided in Table 2.2.

Table 2.2. Description of 38 radiomic features extracted from breast DCE-MR images of lesions.[38,41,69]

Feature Category	Feature Name	Feature Description
Size	1. Volume (mm^3)	Lesion volume
	2. Effective diameter (mm)	Diameter of a sphere that has the same volume as the lesion
	3. Surface area (mm^2)	Surface area of the lesion
	4. Maximum linear size (mm)	Maximum distance between any two voxels in the lesion
Shape	5. Sphericity	Describes how similar lesion's shape is to a sphere
	6. Irregularity	Describes how different lesion's surface is from the surface of a sphere
	7. Surface/volume ratio (1/mm)	Ratio of surface area to volume
Morphology	8. Margin sharpness	Mean of the image gradient at the lesion margin
	9. Variance of margin sharpness	Variance of the image gradient at the lesion margin
	10. Variance of radial gradient histogram	Degree to which the enhancement structure extends in a radial pattern originating from the center of the lesion
Enhancement texture	11. Energy	Measures image homogeneity
	12. Contrast	Measures local image variations
	13. Correlation	Measures image linearity
	14. Entropy	Measures the randomness of the gray-levels
	15. Sum of squares (Variance)	Measures the spread in the gray-level distribution
	16. Difference entropy	Measures the randomness of the difference of neighboring voxels' gray-levels
	17. Difference variance	Measures variations of difference of gray-levels between voxel pairs
	18. Inverse difference moment	Measures the image homogeneity
	19. Sum average	Measures the overall image brightness
	20. Sum entropy	Measures the randomness of the sum of gray-levels of neighboring voxels
	21. Sum variance	Measures the spread in the sum of the gray-levels of voxel pairs distribution

Table 2.2. Description of 38 radiomic features extracted from breast DCE-MR images of lesions.[38,41,69]

Feature Category	Feature Name	Feature Description
Enhancement texture	22. Information measure of correlation 1	Measures nonlinear gray-level dependence
	23. Information measure of correlation 2	Measures nonlinear gray-level dependence
	24. Maximum correlation coefficient	Measures nonlinear gray-level dependence
Kinetic curve assessments	25. Maximum enhancement	Maximum contrast enhancement
	26. Time to peak (s)	Time at which the maximum enhancement occurs
	27. Uptake rate (1/s)	Uptake speed of the contrast enhancement
	28. Washout rate (1/s)	Washout speed of the contrast enhancement
	29. Curve shape index	Difference between late and early enhancement
	30. Enhancement at first post-contrast time point	Enhancement at first post-contrast time point
	31. Signal enhancement ratio	Ratio of initial enhancement to overall enhancement
	32. Volume of most enhancing voxels (mm^3)	Volume of the most enhancing voxels
	33. Total rate variation ($1/s^2$)	Measures how rapidly the contrast will enter and exit from the lesion
	34. Normalized total rate variation ($1/s^2$)	Measures how rapidly the contrast will enter and exit from the lesion
Enhancement-variance kinetics	35. Maximum variance of enhancement	Maximum spatial variance of contrast enhancement over time
	36. Time to peak at maximum variance (s)	Time at which the maximum variance occurs
	37. Enhancement variance increasing rate (1/s)	Rate of increase of the enhancement-variance during uptake
	38. Enhancement variance decreasing rate (1/s)	Rate of decrease of the enhancement-variance during washout

2.5. Statistical Analysis of Feature Robustness

The statistical analysis was performed in Matlab (MathWorks, Natick, MA) with the software developed specifically for the study as well as with the software developed in the Giger Lab over the years. The robustness of radiomic features was analyzed in terms of 1) feature values across the two datasets, 2) feature values across the two datasets within the clinical subgroups (LN+, LN-, etc.), 3) feature classification performance in a specific clinical task (LN+ vs. LN-, etc.) across the two datasets, and 4) feature model classification performance in a specific clinical task (LN+ vs. LN-, etc.) across the two datasets.

Table 2.3. Number of cases in the subsets of Dataset A and Dataset B after matching for size and clinical distributions.

Clinical Question	Number of Cases	
	Dataset A_{LN}	Dataset B_{LN}
LN+	43 (48.3%)	86 (48.3%)
LN-	46 (51.7%)	92 (51.7%)
Total for LN	89	178
	Dataset A_{ER}	Dataset B_{ER}
ER+	74 (84.1%)	148 (84.1%)
ER-	14 (15.9%)	28 (15.9%)
Total for ER	88	176
	Dataset A_{PR}	Dataset B_{PR}
PR+	65 (77.4%)	130 (77.4%)
PR-	19 (22.6%)	38 (22.6%)
Total for PR	84	168
	Dataset A_{HER2}	Dataset B_{HER2}
HER2+	16 (22.2%)	32 (22.2%)
HER2-	72 (81.8%)	144 (81.8%)
Total for HER2	88	176

Each feature was standardized within Dataset A and within Dataset B with zero-mean and unit-variance standardization. The two datasets were matched in terms of size and clinical distributions for each of the clinical tasks. The matching process yielded subsets of Dataset A and Dataset B, notated by Datasets A_{LN} and B_{LN}, A_{ER} and B_{ER}, A_{PR} and B_{PR}, and A_{HER2} and B_{HER2}. Table 2.3 summarizes the number of cases in each subset.

2.5.1. Robustness in Feature Values

To assess the robustness in feature values, average values within the entire dataset and within each clinical subset were compared graphically and statistically using the Mann-Whitney U test.[71,72]

2.5.2. Robustness in Individual Feature Classification Performance

Robustness evaluation of individual feature classification performance was performed using superiority and non-inferiority testing.[73] The area under the receiver operating characteristic curve (AUC) served as the figure of merit for feature performance in the classification tasks.[74] The difference in the feature performance between Dataset A and Dataset B was evaluated based on the difference in the AUC values (ΔAUC).

First, superiority analysis was performed with a two-sided 95% confidence interval test on ΔAUC values to identify features that showed statistically significant difference in performance. The AUC values were estimated with Wilcoxon approximation and the confidence intervals were calculated using a bootstrap method with 500 iterations.[75,76] Since ROC analysis assumes normally distributed data and the datasets were moderately small, bootstrapping was used to make calculations independent of the data distribution. Difference in AUC was considered statistically significant if the two-sided 95% confidence interval did not include zero.[75] Next, non-inferiority

testing was conducted for the features, the performance of which, based on the superiority testing, failed to show a statistically significant difference between the two datasets. The non-inferiority test determined whether the individual feature performance in the task of differentiating the lymph node status or the hormone receptor status on one dataset was non-inferior or equivalent to its performance on the other dataset. Thus, a one-sided 90% confidence interval on ΔAUC was studied.

2.5.3. Robustness in Feature Model Classification Performance

After performing robustness analysis of individual features, we further evaluated the robustness of classification models for the four clinical tasks. The robustness was assessed by comparing the performance of the model trained on Dataset A independently tested on Dataset B to the cross-validated performance evaluated on Dataset A.

The classification models included combinations of features and were constructed using linear discriminant analysis.[77] Based on the results from previous work performed in the lab for the task of lymph node status assessment on Database A, the classification model for the task was created using the two features previously identified as having the best classification performance.[19] For the ER, PR, and HER2 assessments, stepwise feature selection was used to identify features for the classification models of ER, PR, and HER2 statuses.

First, the models were trained on the subsets of Dataset A and tested on the corresponding subsets of Dataset B. Testing was performed on ten random samples of Datasets B_{LN}, B_{ER}, B_{PR}, and B_{HER2}, to match for the number of cases in the corresponding clinical subsets of Dataset A. The final results were found by averaging the results from the individual analysis of the ten subsets. Second, leave-one-out cross-validation was conducted to assess the model performance within

Dataset A. The classification performances of the two models were compared between Dataset A and Dataset B in terms of AUC.

2.6. Results

2.6.1. Robustness in Feature Values

Figure 2.2 shows the graphical assessment of average feature values computed on all cases of Dataset A and Dataset B. Volume and surface area average feature values are multiple orders of magnitude larger than the rest of the feature values, and thus are omitted from the figure for better visualization of the remaining features. Statistical analysis showed no statistically significant differences in average values of features from the size category (Table 2.2). The rest of the features in terms of their original values were found to be statistically different between Dataset A and Dataset B.

Figures 2.3-2.6 show the comparison of feature values between datasets within positive and negative subgroups for the four clinical tasks. Larger spread in features values between the two datasets is seen for the ER-, PR-, and HER+ subsets; smaller spread is seen for the ER+, PR+, HER2-, LN+, and LN-. It is important to note the class prevalence effect on the results. The subsets with larger feature value differences have a lesser number of cases than the subsets with the smaller feature value spread. In the future, the analysis should be repeated on the datasets that are class-balanced.

Further analysis of the feature value robustness is focused only on the LN clinical task, since it is the only task that leads to balanced Dataset A_{LN} and Dataset B_{LN} in terms of positive and negative

cases (Table 2.3). Both Dataset A_{HER2} and Dataset B_{HER2} are largely skewed towards negative cases; while Dataset A_{ER} and Dataset B_{ER} as well as Dataset A_{PR} and Dataset B_{PR} have many more positive cases. Further investigation of Figure 2.3 for the LN classification is given in Table 2.4, which quantitatively summarizes the spread of feature values between the two datasets for positive and negative LN subgroups by giving the range of distances to the diagonal identity line for each feature category.

Table 2.4. The range of unsigned distances of standardized feature values to the diagonal line, as shown in Figure 2.2. The ranges are presented for LN+ and LN- subgroups for the six feature categories.

Feature Category	LN+	LN-
Size	[0.006, 0.098]	[0.013, 0.100]
Shape	[0.062, 0.104]	[0.042, 0.126]
Morphology	[0.023, 0.088]	[0.024, 0.074]
Enhancement Texture	[0.035, 0.163]	[0.025, 0.162]
Kinetic curve assessments	[0.002, 0.080]	[0.006, 0.090]
Enhancement-variance kinetics	[0.067, 0.098]	[0.060, 0.096]

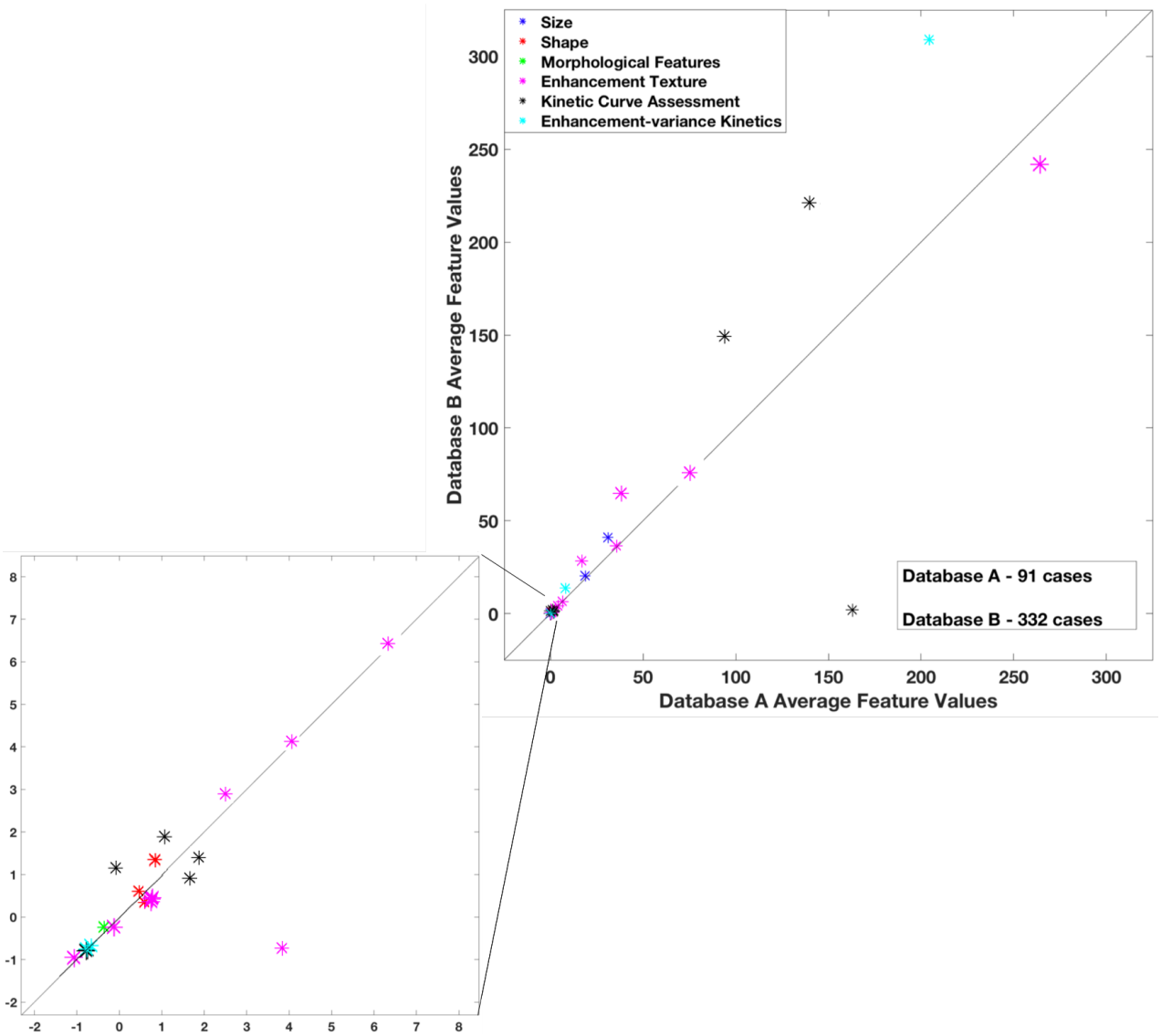


Figure 2.2. Comparison of average radiomic feature values between Dataset A and Dataset B. Volume and surface area average feature values are omitted from the figure due to their large values, compared to the rest of the features.

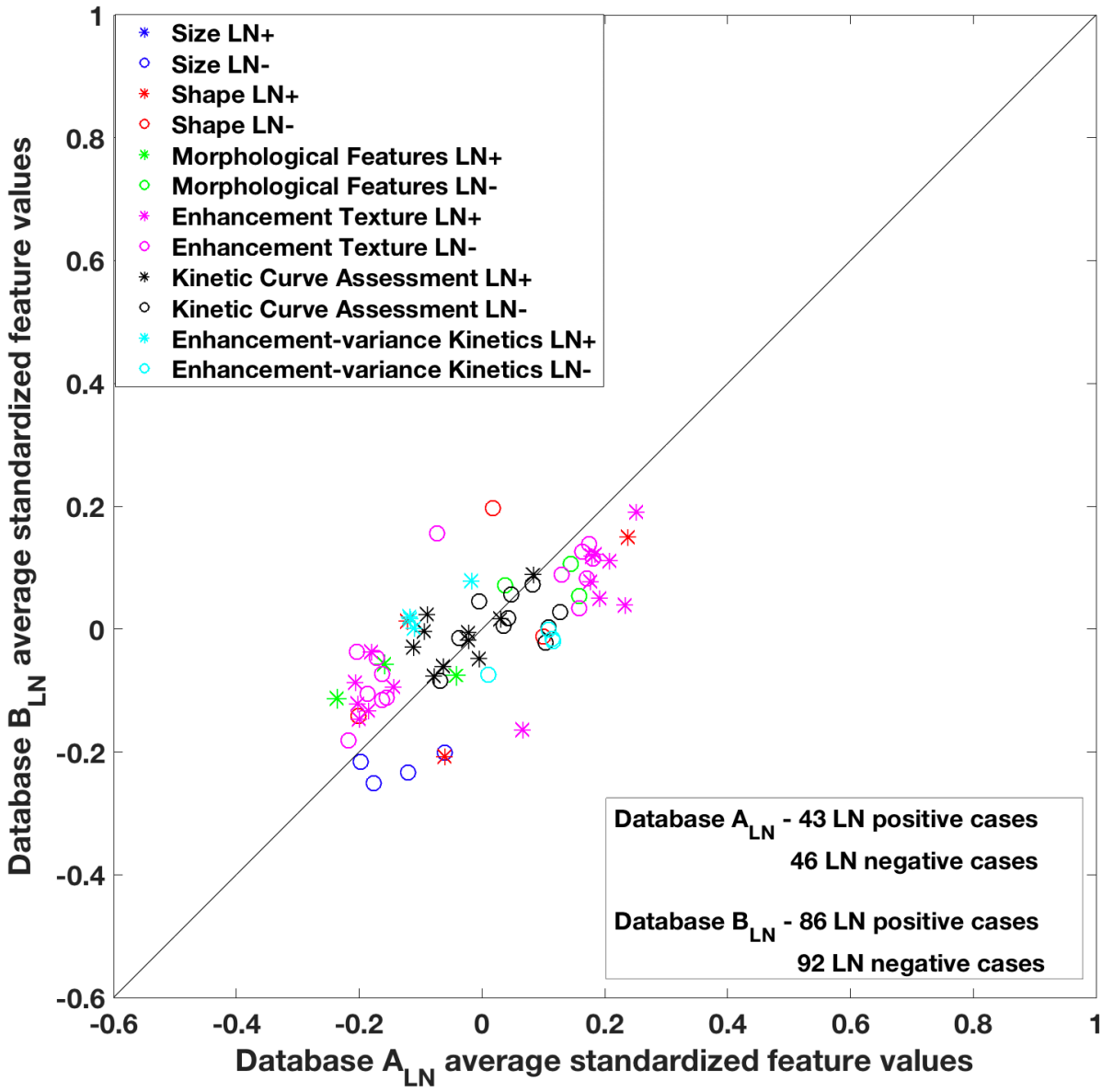


Figure 2.3. Comparison of average standardized feature values for LN clinical task. The comparison is performed between LN positive cases in Dataset A_{LN} and Dataset B_{LN} and between LN negative cases in Dataset A_{LN} and Dataset B_{LN}.

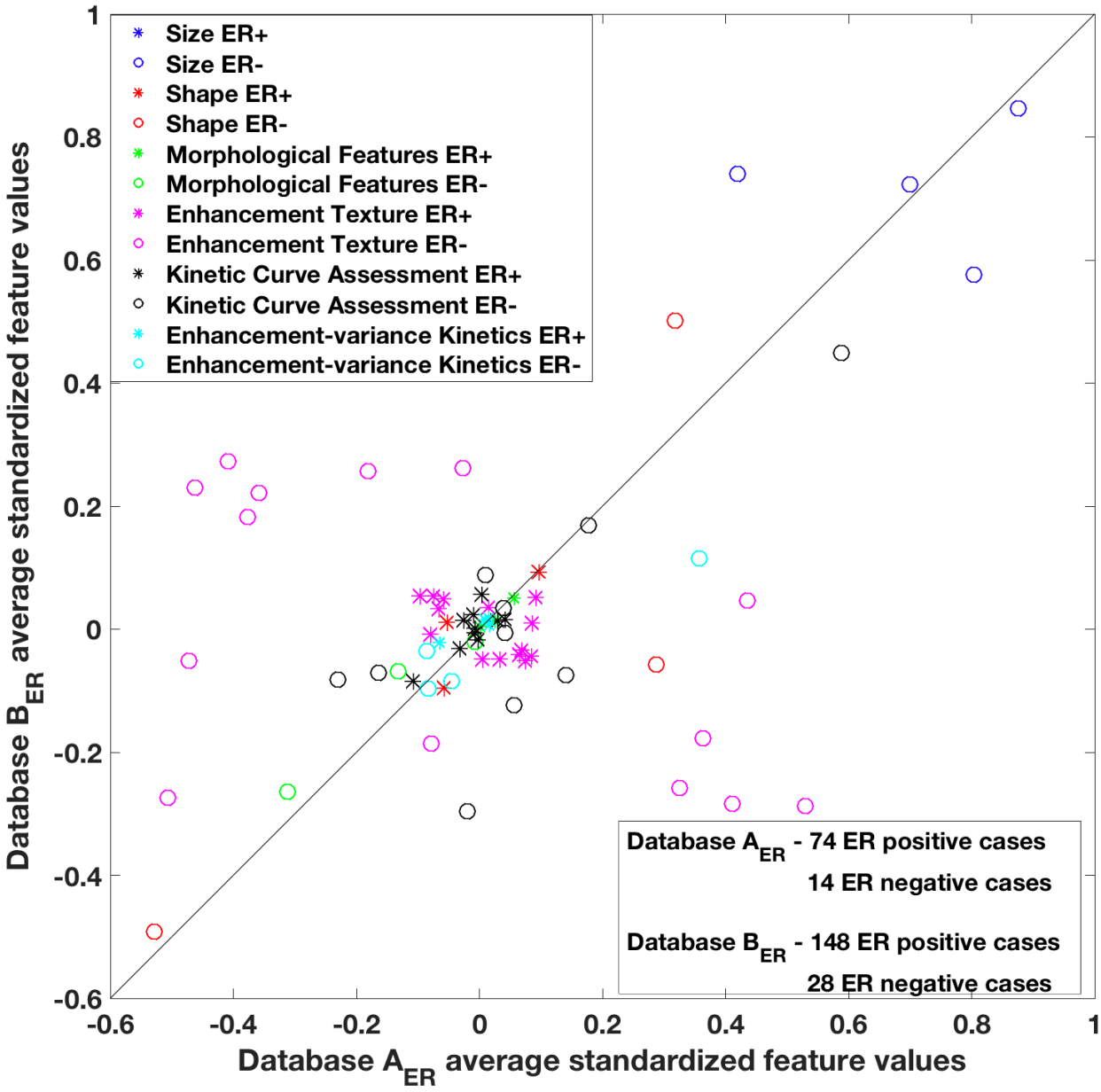


Figure 2.4. Comparison of average standardized feature values for ER clinical task. The comparison is performed between ER positive cases in Dataset A_{ER} and Dataset B_{ER} and between ER negative cases in Dataset A_{ER} and Dataset B_{ER}.

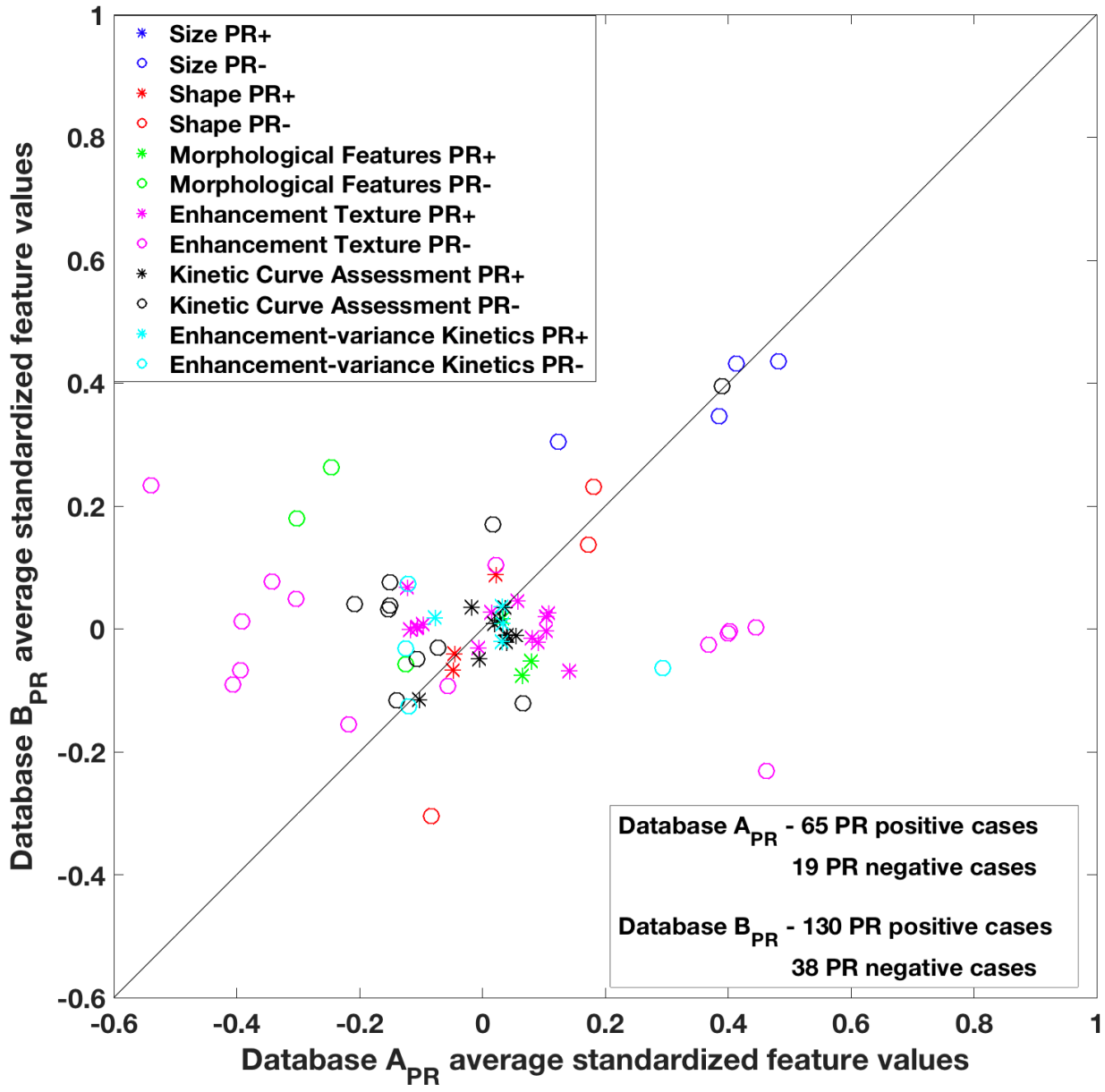


Figure 2.5. Comparison of average standardized feature values for PR clinical task. The comparison is performed between PR positive cases in Dataset A_{PR} and Dataset B_{PR} and between PR negative cases in Dataset A_{PR} and Dataset B_{PR}.

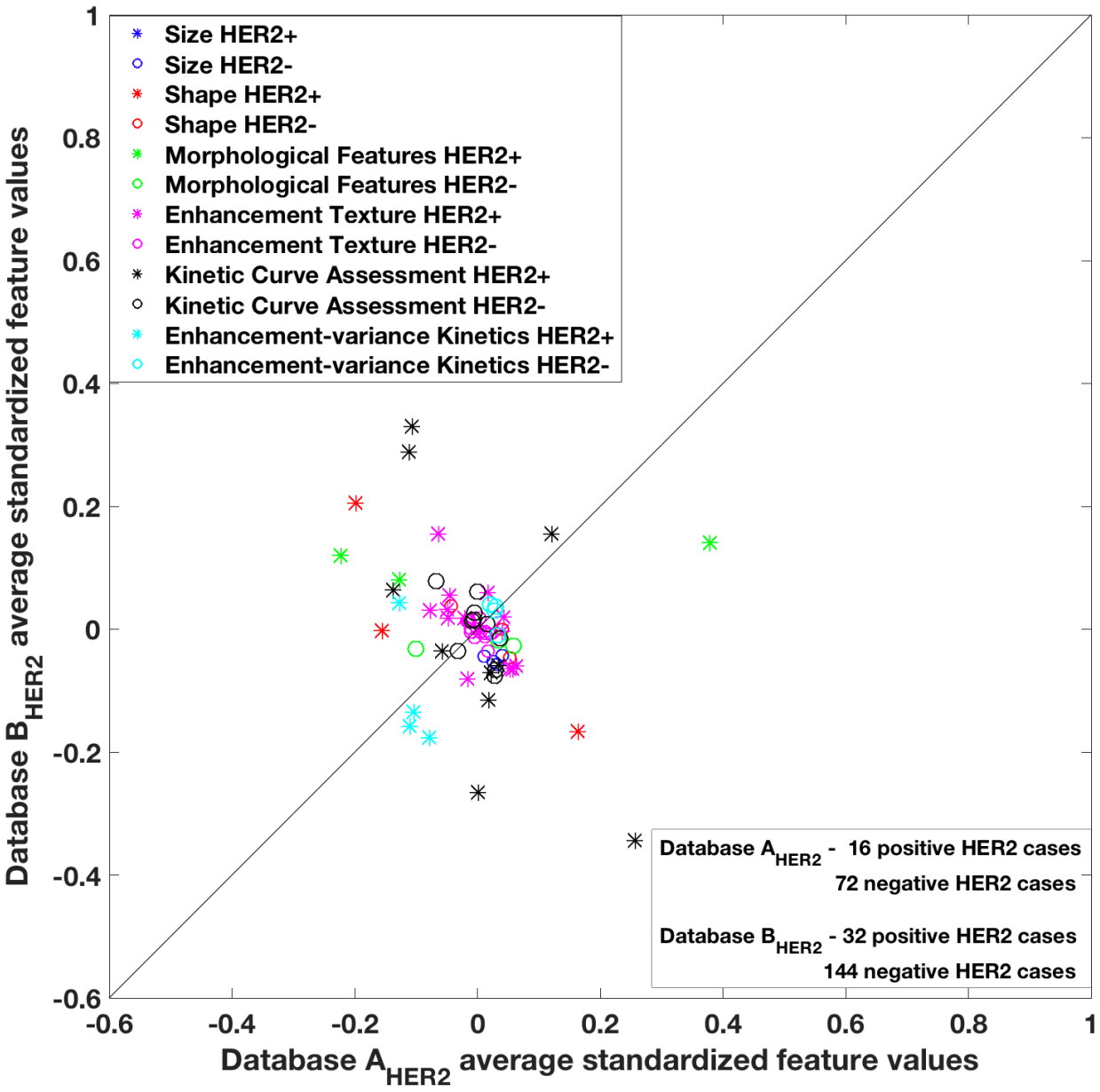


Figure 2.6. Comparison of average standardized feature values for HER2 clinical task. The comparison is performed between HER2 positive cases in Dataset A_{HER2} and Dataset B_{HER2} and between HER2 negative cases in Dataset A_{HER2} and Dataset B_{HER2}.

2.6.2. Robustness in Individual Feature Classification Performance

The superiority tests yielded seven features with statistically significant difference in performance in the task of distinguishing ER status, three features in distinguishing PR status, and one feature in distinguishing HER2 status. Possibly due to statistical skewness arising from the small number of ER-, PR-, and HER2+ cases, the non-inferiority testing was not performed for evaluation of feature performance in the task of differentiating the three hormone receptor statuses. For the lymph node classification, no features were found statistically different in their performances across the two datasets. Therefore, all of the features were used in the non-inferiaority analysis, which demonstrated varying degrees of robustness in the lymph node classification performance for different features. Figure 2.7 shows the lower bound of the 90% confidence interval for features that performed better than guessing on both datasets (AUC>0.5). Two features, one describing lesion morphology and another describing lesion homogeneity, showed best agreement in performance with absolute value of the lower bound of the 90% confidence interval for ΔAUC<0.05. Surface area and inverse difference moment resulted in absolute values of the lower bound of the 90% confidence interval for ΔAUC of 0.074 and 0.071, respectively.

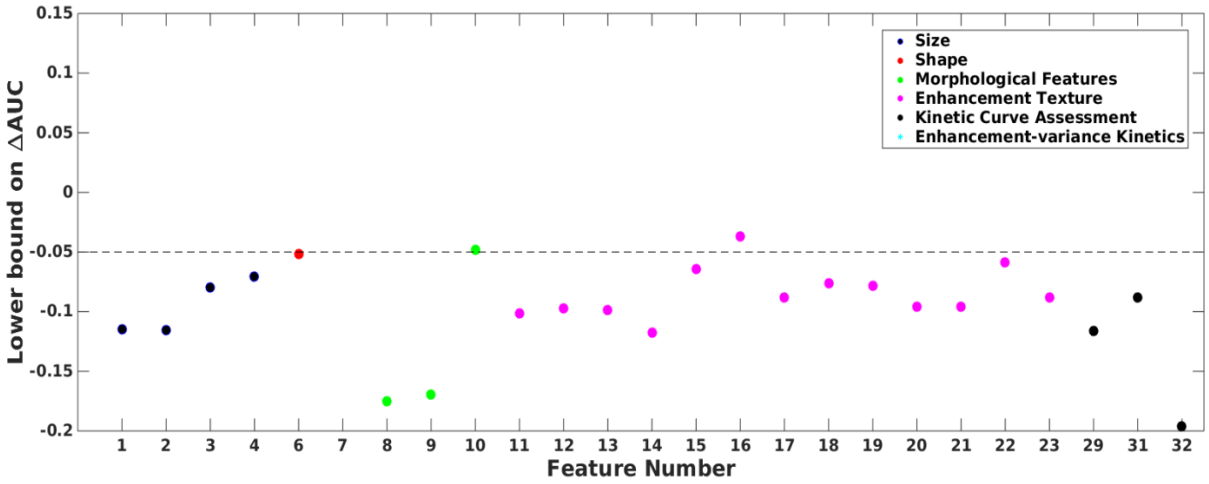


Figure 2.7. Non-inferiority testing of equivalence in performance in the task of distinguishing between lesions with positive or negative lymph nodes. Dataset A_{LN} and Dataset B_{LN} are used.

2.6.3. Robustness in Feature Model Classification Performance

For the task of LN classification, the model, trained on Dataset A_{LN}, achieved classification performance on Dataset B_{LN} with AUC value of 0.64 (standard error (se) = 0.1). Leave-one-out cross-validation within Dataset A_{LN} resulted in AUC value of 0.64 (se = 0.1). The independent testing across datasets and cross-validation within Dataset A was not performed for ER, PR, and HER2 clinical tasks due to those subsets' skewed distributions.

2.7. Discussion and Conclusions

The work presented in this chapter revealed CADx features that are robust in their value and/or classification performance across MRI scanners of two different manufacturers. In particular, we failed to show statistically significant difference in average values of the size features. The rest of the features were found to be statistically significantly different in terms of their average value. The features that visually showed the greatest variation of their average values were from kinetic

curve assessment and enhancement-variance kinetics categories. These included *time to peak,* representing the time at which maximum enhancement occurs and measured in seconds; *volume of the most enhancing voxels,* measured in mm^3 and indirectly dependent on the time rate of the scanner; *total rate variation*, representing the rate at which contrast enters and exits from the lesion and is measured in sec^{-2}; and *time to peak at maximum variance, the* time at which the maximum variance occurs. All four features are related to time and may vary depending on the time resolution of GE and Phillips MRI scanners. Further investigation of the effect of time resolution on feature values needs to be performed.

In future work, larger datasets controlled for the class prevalence, need to be acquired to eliminate the effect of small sample size on the analysis. The smallest spread in feature values was found across the two MRI scanner for the LN+ from LN- subsets. It is important to note that the LN+ and LN- subsets had large numbers of available cases. It is probable that the number of available cases affects the comparison of average feature values. This can be observed in evaluating the spread of average feature values across the two scanners for the ER, PR, and HER2 tasks. ER-, PR-, and HER+ have a smaller number of cases and show a large deviation of feature values across databases. In contrast, ER+, PR+, and HER-, having a larger number of cases, showed closer agreement.

Some features demonstrated close agreement in performance in distinguishing LN+ from LN- cases. In the analysis, we arbitrarly set the cut-off value of ΔAUC that signified feature equivalence in performance. Any feature with $\Delta AUC < 0.05$ was decided to have equivalent performance on the two datasets. However, the cut-off value for difference in AUC needs to be further explored, taking into consideration the task and the size of datasets.

In our research, we performed feature value standardization to obtain mean zero and unit variance. As seen in our analysis, time resolution of scanners has an effect on multiple feature values and their performance in lesion characterization. A more detailed study of other scanner parameters and patient population variations may be useful to understand their influence on feature values. Subsequently, a standardization technique that corrects for the variations needs to be employed to harmonize the data prior to feature extraction. As an alternative, the radiomics feature extraction scheme itself may be tailored to account for large data variations.

In summary, the work presented in this dissertation chapter evaluated the robustness of quantitative MRI radiomics in the task of distinguishing lymph node involvement and hormone receptor statuses across two MRI scanners, GE and Philips. The features were studied across two databases in terms of their average feature values computed within entire datasets and within clinical subgroups. Additionally, non-inferiority testing was used to study their performance in the task of distinguishing lymph node status, but not receptor statuses. In conclusion, MRI features showed promise in robustness in terms of their average values across MRI scanners. The conclusions need to be supported by analysis of larger datasets. The work showed that the number of cases available for evaluation had a substantive effect on the results. In general, the radiomic features appear useful in the classification within each dataset. Statistical analysis revealed robust features in cancer molecular and lymph node classification. Lesion size and enhancement texture features hold promise exhibiting equivalent prognostic performance in the task of distinguishing lymph node status across databases.

CHAPTER 3

DEEP LEARNING-BASED AND CONVENTIONAL RADIOMICS OF BREAST CANCER

The previous chapter analyzed conventional radiomic features in terms of their robustness across MRI scanners for breast lymph node and hormone receptor status analysis. This chapter develops deep learning-based methods and studies their fusion with conventional radiomics for breast DCE-MRI. Furthermore, the work of this chapter evaluates the effect of a lesion image representation input into the proposed deep learning classification pipeline. The methods are developed for two clinical tasks, distinguishing malignant and benign lesions and prediction of cancer response to therapy.

3.1. Introduction

In the 1990s, early forms of convolutional neural networks (CNNs) were introduced for CADx by learning imaging features directly from regions of interest (ROIs) without explicit manual intervention.[78,79] Recent advances in technology have led to the widespread use of deep learning methods that use deeper and more advanced CNN architectures for general computer vision tasks. Although CNNs typically rely on massive datasets for training and are thus often intractable for CADx, it has been shown that standard transfer learning techniques like fine-tuning or feature extraction based on pre-trained CNNs can be used to reduce the need for larger datasets.[80,81] As a

result, deep learning techniques have exhibited strong predictive performances on CADx tasks without requiring massive datasets.[82,83]

However, challenges remain in developing deep learning methods for characterizing medical images. Methods are still reliant on extensive image preprocessing, are hindered by heterogeneous data sources, and often suffer from long training times, leading to inefficient use of data for validation. We present a methodology that extracts and pools low-, mid-, and high-level features using a pre-trained CNN and integrates them with hand-crafted radiomic features computed using conventional CADx methods, which were also studied in Chapter 2. Figure 3.1 graphically summarizes the lesion classification pipeline proposed in this chapter, with deep learning-based and conventional hand-crafted radiomics methods. The methodology demonstrates strong performance in the task of estimating the probability of breast lesion malignancy and predicting cancer treatment response based on DCE-MRI without the need for preprocessing or long training times. Furthermore, we demonstrate that the lesion image input has a large effect on the classification performance and needs to be carefully designed prior to the image analysis.

3.2. Clinical Tasks

The DCE-MRI radiomics methods were developed for two clinical tasks: 1) classifying breast lesions as benign and malignant, and 2) predicting breast cancer response to neoadjuvant chemotherapy.

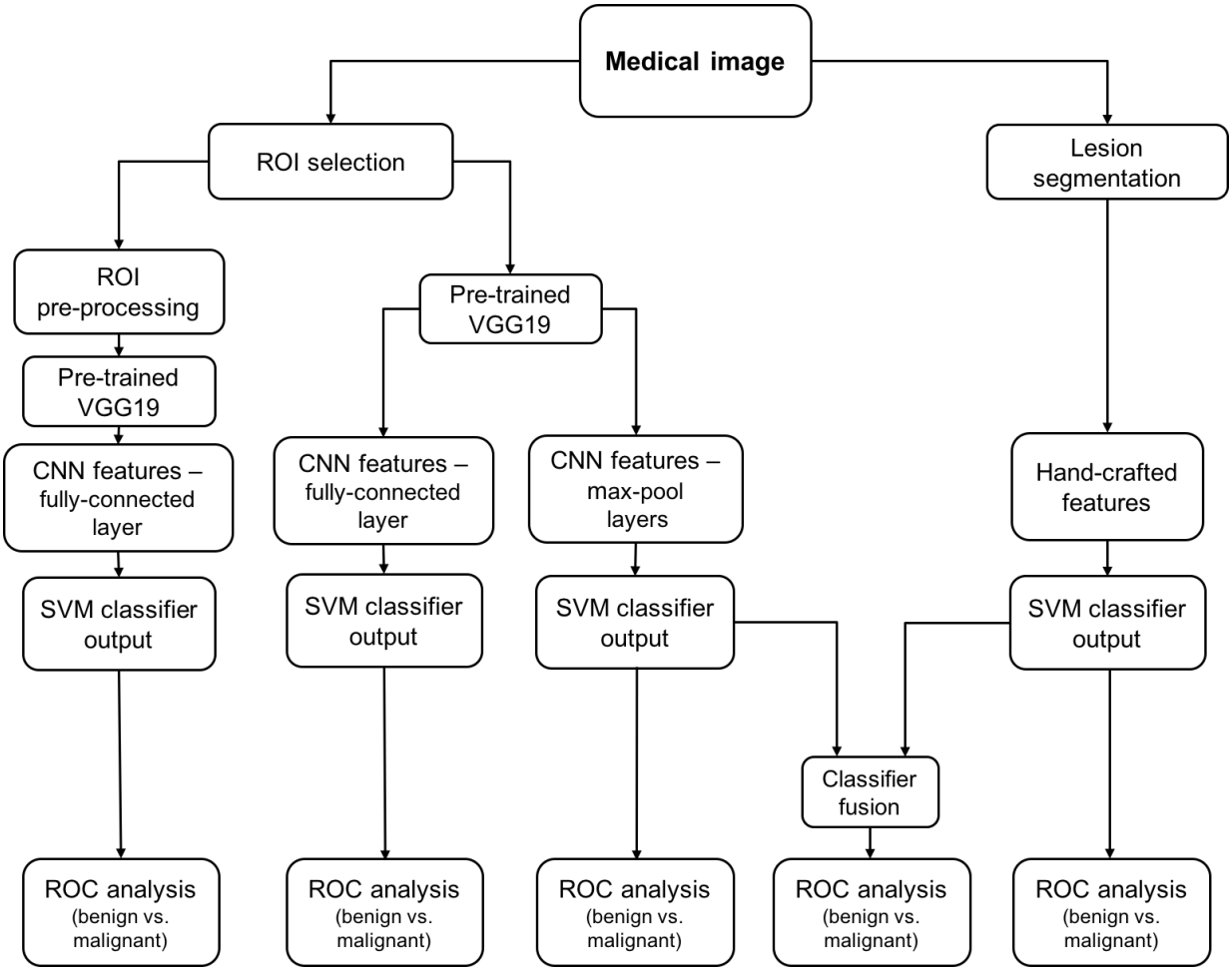


Figure 3.1. Lesion classification pipeline based on diagnostic images. Two types of features are extracted from a medical image: 1) CNN features with pre-trained CNN and 2) handcrafted features with conventional CADx. High-, mid-, and low-level features extracted by pre-trained CNN are evaluated in terms of their classification performance and preprocessing requirements. Further, the classifier outputs from the pooled CNN features and the hand-crafted features are fused in the evaluation of a combination of the two types of features.

3.3. DCE-MRI Datasets

3.3.1. Breast Lesion Malignancy

The radiomics methods for malignancy assessment were developed based on the breast DCE-MRI dataset, retrospectively collected under a HIPAA-compliant Institutional Review Board protocol. The dataset was collected at the University of Chicago over a 7-year period, 2006 - 2013, and includes 690 breast cases, annotated as benign (212 cases) or malignant (478 cases) based on pathology and radiology reports (Table 3.1). All of the lesions were clinically biopsy-confirmed. Both primary and secondary lesions were utilized in the study. Malignant cases included masses and non-mass enhancements; benign cases included masses and foci. Detailed clinical characteristics of the dataset are presented in Table 3.2.

Table 3.1. Properties of the DCE-MRI image dataset.

Number of benign lesions	212
Number of malignant lesions	478
Total number of lesions	690
Average pixel size	0.69 mm
ROI size range	48x48 – 126x126 pixels

A portion of the contrast-enhanced MR images, 454 cases, was acquired on a Philips Achieva 1.5 Tesla (T) scanner. The remaining 236 cases were acquired with a Philips Achieva-TX 3T Philips system. A T1-weighted spoiled echo gradient sequence was utilized during image acquisition. To perform dynamic imaging, patients were injected with the following gadolinium-based contrast agents: Ominscan was used for patients with GFR over 60ml/min prior to November 25, 2012 and Multihance was used for the patients with GFR less than 60ml/min and for all patients imaged after November 25, 2012. Each sequence included a pre-contrast image followed by

multiple post-contrast images, with the first post-contrast image obtained 55-60 seconds following the contrast injection. The dataset consisted of images of various slice thickness, with 2/3 of the cases having slice thickness of 2 mm and 1/3 cases having slice thickness 1.5 mm or 1.6 mm.

Table 3.2. Clinical characteristics of the DCE-MRI dataset.

Age: mean (STD)	54.9 (13.3) Unidentified cases: 103	
Benign Tumor Characteristics		
Tumor Subtypes:	Fibroadenoma	87
	Fibrocystic change	77
	Papilloma	12
	Unidentified	36
Malignant Tumor Characteristics		
Tumor Subtypes:	Invasive ductal carcinoma	135
	Ductal carcinoma in situ	19
	Invasive ductal carcinoma + ductal carcinoma in situ	263
	Invasive lobular carcinoma	18
	Invasive lobular carcinoma mixed	19
	Unidentified	24
Estrogen Receptor Status: No. of cases	Positive	328
	Negative	108
	Unidentified	42
Progesterone Receptor Status: No. of cases	Positive	274
	Negative	159
	Unidentified	45
HER2 Status: No. of cases	Positive	72
	Negative	349
	Equivocal	3
	Unidentified	54

3.3.2. Breast Cancer Treatment Response

The radiomics methods for prediction of breast cancer response to neoadjuvant chemotherapy were developed based on two DCE-MRI datasets, one collected at the University of Chicago and another

collected for the **I**nvestigation of **S**erial Studies to **P**redict **Y**our **T**herapeutic **R**esponse with **I**maging and mo**L**ecular **A**nalysis (ISPY TRIAL) breast cancer trail.[84]

The analysis was performed on a DCE-MRI dataset that was a subset of the University of Chicago DCE-MRI dataset utilized for malignancy assessment and described in Section 3.3.1. The subset included 82 breast cancer cases that had chemotherapy treatment response available. The imaging details were detailed in Section 3.3.1 and are not repeated here. The DCE-MR images utilized in this research were acquired before beginning the chemotherapy treatment. The clinical and pathologic responses, determined from radiology and pathology reports, were available for 82 and 28 cases, respectively. Out of the 28 cases with pathologic response available, 12 cancers achieved pathologic complete response (pCR) and 16 cancers did not achieve pCR. Clinical response was categorized in two ways: A) the responding group (26 cases) included only cases with complete response and the non-responding group (56) included cases with partial response, stable, and progressive disease, and B) the responding group (63) included cases with either complete or partial response and the non-responding group (19) included cases with either stable or progressive disease. The treatment regimens for these cases were unknown. Because of a limited number of cases, this dataset was utilized only for the evaluation of the conventional CADx radiomics.

The ISPY dataset is a publically available de-identified beast DCE-MRI dataset, collected for the ACRIN 6657 study.[84] It was made available on and downloaded from The Cancer Imaging Archive.[64,84] The ACRIN study included women with lesion size of 3 cm and greater and who were scheduled to receive anthracycline-based neoadjuvant chemotherapy. Within this dataset, each patient had four MRI exams at the following time points: 4 weeks before beginning

anthracycline-cyclophosphamide chemotherapy treatment (MRI_1); at least two weeks after the first cycle of anthracycline-cyclophosphamide chemotherapy (MRI_2); after receiving all of the anthracycline-cyclophosphamide and before taxane (MRI_3); and after the completion of chemotherapy and before surgery (MRI_4). Images were acquired on 1.5T scanners of the three major MRI manufacturers, Siemens, Philips, and GE. Fat-suppressed, T1-weighted gradient echo sequence was used with TR≤20 ms, TE = 4.5 ms, 16cm to 18 cm field-of-view, 64 slices, and slice thickness ≤ 2.5 mm. Biopsy-proven pathologic response was available for the dataset.

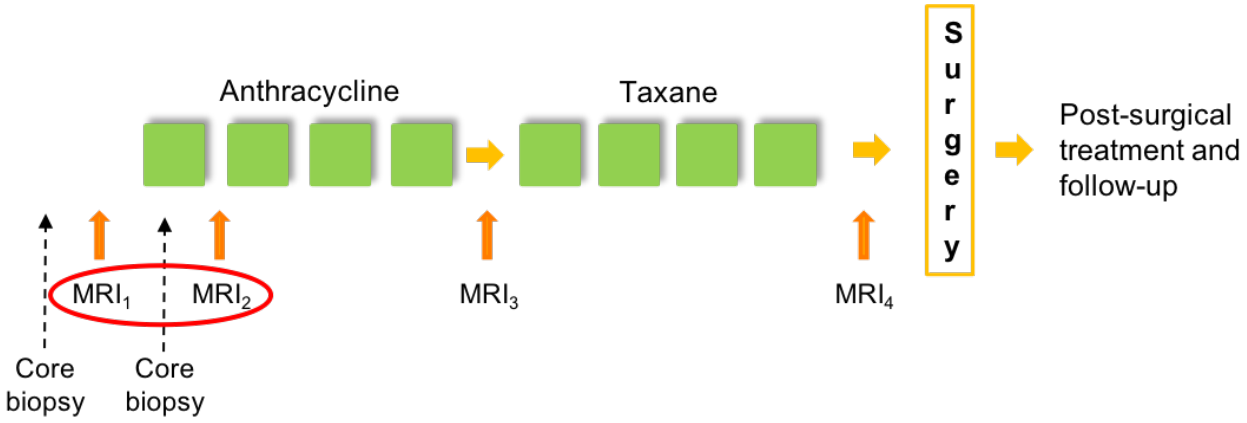


Figure 3.2. Treatment regimen and imaging schedule for the patients in the ISPY dataset.[84] DCE-MRIs were acquired prior to the start of the therapy (MRI_1), following the first cycle of anthracycline (MRI_2), following all cycles of anthracycline (MRI_3), and after the entire chemotherapy treatment (MRI_4). Our research developed radiomics methods based on the first two MRIs, i.e. MRI_1 and MRI_2.

A total of 162 women were enrolled in the ACRIN study. In our research, only a subset of the 162 cases was analyzed. We studied whether or not we can automatically predict pathologic treatment response based on the pre-treatment and the first post-treatment MR images. The subset included 126 breast cancer cases with 35 cases achieving pCR and 91 cases not achieving pCR. The criteria for the case inclusion was based on the availability of the pathologic treatment

response and of both MRI_1 and MRI_2. The clinical response was not available and therefore not evaluated for this dataset.

Images at all treatment time points were utilized by the Giger Lab in the NCI Quantitative Imaging Network (QIN) Breast MRI Metrics of Response (BMMR) challenge. The challenge was organized to develop biomarkers for assessment of breast cancer pathologic response and recurrence and was hosted on the QIN Labs website. The Giger Lab participated in the challenge and took second place. The results are for the breast cancer recurrence are partially presented in this dissertation and are fully described elsewhere.[85,86]

3.3.3. Lesion Regions of Interest (ROIs)

Referring to Figure 3.1, lesion representations needed to be selected from the 4D DCE-MRI data prior to the application of deep learning methods studied in this chapter. For all datasets, each lesion was represented by a region of interest (ROI) enclosing the lesion. Figure 3.3 illustrates examples of lesion ROIs selected from the DCE-MRI central slices for a benign and malignant case.

DCE-MRI is unique compared to other breast imaging modalities, such as mammography and ultrasound. DCE-MR images are 4D data that include volumetric and temporal components. Since the pre-trained CNNs require a 2D image input into three channels, a decision is required regarding which slice of the entire volume and which time point to use for ROI selection. In this part of the research, ROIs were selected around each lesion on a transverse slice in the area of the lesion center (some center slices had a biopsy clip and were avoided) at the pre-contrast time point (t0) and at the first (t1) and the second post-contrast time points (t2). For each lesion, the ROI size

was chosen based on the maximum dimension of the lesion and held constant across DCE time points. The smallest ROI size was set to 48x48 pixels, to match pre-trained CNN requirements on the minimal input ROI size. The chosen CNN architecture performs a few pooling operations, which reduce the spatial dimensions of the feature maps by a factor of 2. The size of 48 pixels was chosen based on the number of pooling operations, the pooling stride, and the size of the pooling filter (2x2). The minimum size of the input has to be recalculated for other CNN architectures.

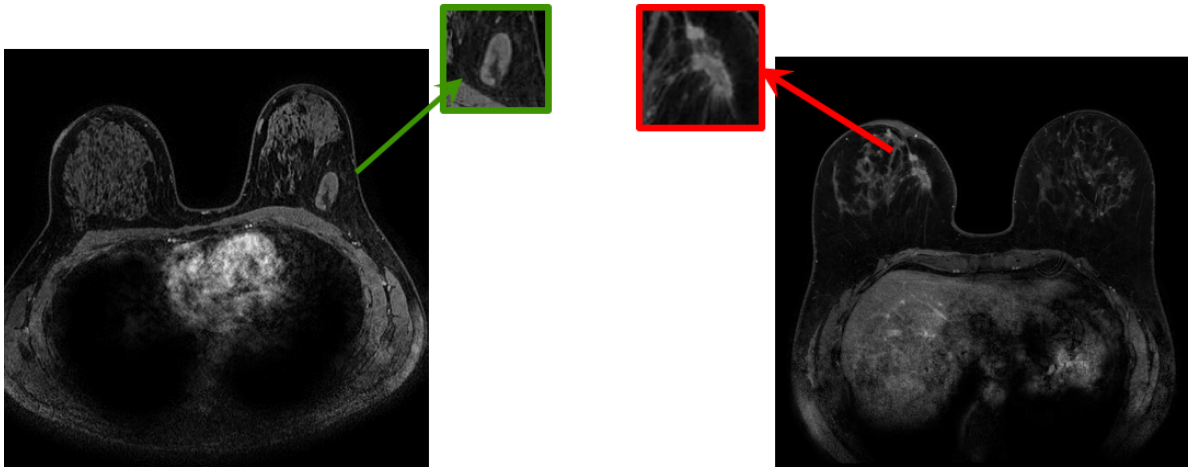


Figure 3.3. Examples of DCE-MRI transverse center slices with the corresponding ROIs extracted. On the left is a benign case and on the right is a malignant case.

3.4. Deep Convolutional Neural Networks for DCE-MRIs

CNN features were extracted from the selected lesion ROIs with the publicly-available VGG19 model, pre-trained on ImageNet.[42,46] There are several CNN architectures that demonstrate high performance with classification tasks on the natural image datasets and are worth mentioning here. AlexNet was a pioneer network that broke the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition classification accuracy record in 2012 by 16%, the largest

margin in the history of the competition before and since.[42,43] In 2014, based on the insights from extensive research of the AlexNet, VGGNet was developed with significant improvements to the AlexNet architecture. One of the improvements was that it reduced the filter sizes from AlexNet's 11x11 and 5x5 to 3x3, effectively reducing the area in the original image on which the filter is acting. This allowed the network to be grown deeper, allowing it to learn more complex higher-level features. In this research, the preliminary studies were performed with the AlexNet architecture. Then, our work compared the performances of AlexNet and VGGNet on the DCE-MRI dataset and found that VGGNet had superior performance. Thus, it was used in all further evaluations. It is important to note that because of fast progress in the field, deep learning classification models have been evolving over the course of the completion of this dissertation work. While VGGNet was applied throughout this research to be able to make comparison to the initial studies, more recent architectures appeared. These include GoogleNet and ResNet, which outperform VGGNet in the classification tasks on ImageNet.[47,87] It is left to the future research to evaluate their effectiveness for the clinical tasks based on DCE-MRI data.

The extracted CNN features were further used to train classifiers evaluated as described in Section 3.5. The architecture of VGG19 model includes five stacks - with each stack containing two or four convolutional layers and a max-pooling layer - followed by three fully-connected layers. The VGG19 architecture and CNN feature-extraction pipeline is illustrated in Figure 3.4. VGG19 model takes in an input to three RGB channels. In this part of the research, ROIs extracted at the pre-contrast and the first and the second post-contrast DCE time points were input to the three channels, as demonstrated in Figure 3.5, artificially creating an RGB lesion ROI input. This procedure allowed the pre-trained VGGNet to simultaneously extract CNN features from the ROIs

at the three time points.

The described ROIs were created with the Matlab software, written specifically for the task. The ROI selection was based on the segmentation maps, previously generated by the in-house software. The software requires the user to specify the locations of the MR image (.ima file extension) and the lesion segmentation file (.les file extension). It produces the RGB ROIs based on the assumption that three contrast time points exist for every image, in particular, the pre-contrast and the 1^{st} and 2^{nd} post-contrast time points.

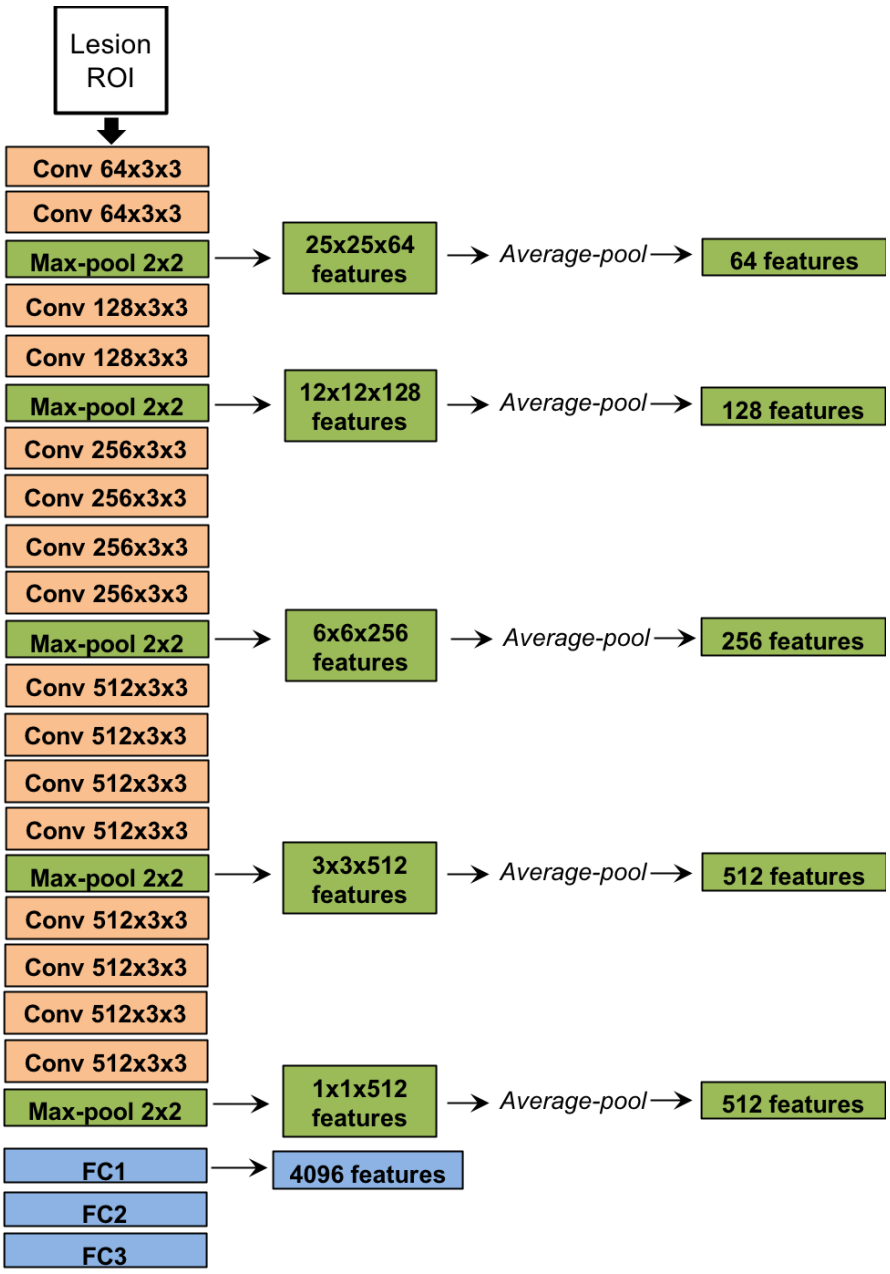


Figure 3.4. Architecture of VGG19 model. It takes in an image ROI as an input. The model comprises five blocks, each of which contains two or four convolutional layers and a max-pooling layer. The five blocks are followed by three fully connected layers. Features are extracted from the five max-pooling layers, average-pooled across the channel (third) dimension, and normalized with L2 norm. The normalized features are concatenated to form our CNN feature vector.

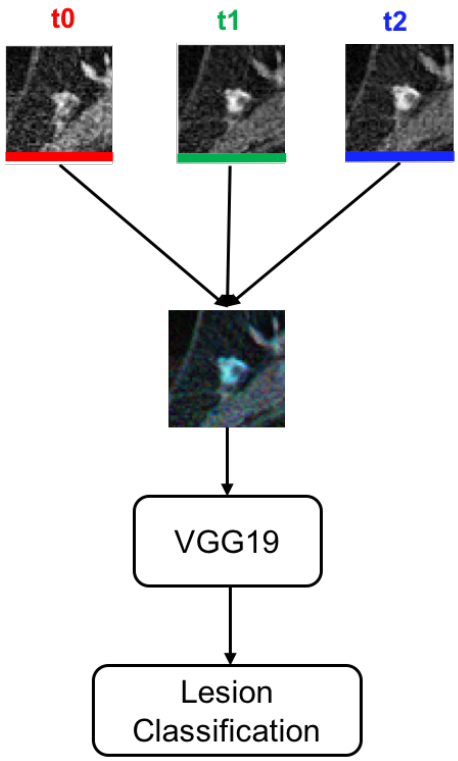


Figure 3.5. Lesion classification pipeline with RGB ROIs. ROIs extracted from the pre-contrast time point (t0) and the first two post-contrast time- points (t1, t2) are input into the three color channels of VGG19, red, green, and blue.

3.4.1. Hierarchical Pooled features

As shown in Figure 3.4, CNN features were extracted from each of five max-pool layers. Using a method similar to the one proposed by Zheng et al,[88] they were then average-pooled[89] along spatial dimensions, resulting in five feature vectors. Each of the five vectors was individually normalized with the Euclidean norm[90] and concatenated to form a final CNN feature vector, which was then normalized again.

It should be noted that the DCE-MRI dataset contained image ROIs of varying sizes. Typically, when extracting features from images of varying sizes, some form of preprocessing or

resizing is necessary in order to ensure the extracted features correspond to the same spatial information across all images. However, by average pooling across the layers, the dimensionality of the features is reduced while preserving the spatial structure of the extracted feature maps. Since a typical feature map dimension in a given layer is *height* x *width* x *depth* and the average pooling is performed across the spatial dimensions of the feature maps, the resulting dimension of pooled features corresponds to the depth dimension of the feature maps. Pooling thus removes the need for preprocessing by producing feature vectors of identical length regardless of original input dimensions. Consequently, the original ROIs of varying sizes were directly input into VGG19 without any preprocessing.

3.4.2. Fully-Connected Features

Following the left-most pipeline presented in Figure 3.1, CNN features were also extracted from the first fully-connected layer for the comparison to the hierarchical pooled features. Due to the sparsity of fully-connected features, all zero-variance features were removed prior to analysis. Since the DCE-MRI dataset had images of varying sizes, the effects of preprocessing were investigated to determine if fully-connected features required resized input ROIs. First, fully-connected features were extracted from the ROIs of original sizes with no preprocessing performed. Then for comparison, ROIs were preprocessed to have constant pixel size (224x224 pixels) prior to feature extraction. To form constant-pixel-size ROIs, DCE-MRI ROIs were enclosed in a frame with pixel values set to the average value of the enclosed ROI (Figure 3.6).

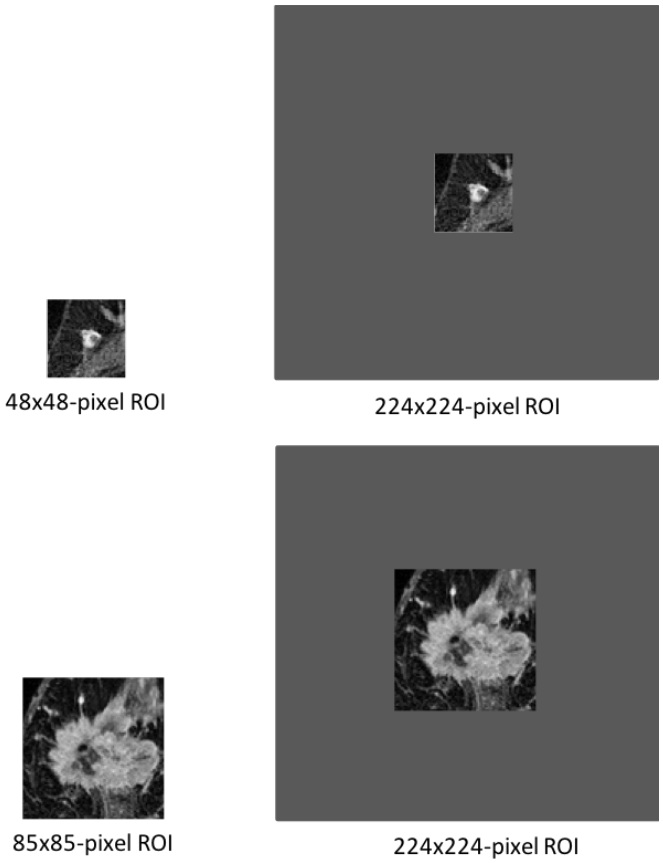


Figure 3.6. Two lesion ROIs used to study the effect of the ROI pre-processing (framing) on the CNN feature performance (Figure 3.1) in classifying breast lesions. The ROI on the left is an ROI with the original constant pixel size, i.e. without any pre-processing. The ROI on the right is created by padding the ROI on the left with the pixel values set to the average value of the surrounded ROI. The size of the padded ROI was set to 224x224 pixels.

3.5. Classification and Evaluation Methods

3.5.1. Lesion Features

The DCE-MRI datasets were utilized to extract CNN features and conventional hand-crafted features, detailed in Section 2.4. Those features were utilized for the task of classifying lesions as benign or malignant and prediction of treatment response. Figure 3.1 schematically shows the classification and evaluation process of our methodology.

3.5.2. SVM Classifiers

A nonlinear support vector machine (SVM)[91] with Gaussian radial basis function (RBF) kernel was utilized for classification using the CNN features and conventional CADx features. We refer to the classifiers based on CNN features and conventional CADx features as CNN-based classifiers and conventional CADx classifiers, respectively.

The SVM was chosen over other classification methods due to its ability to handle sparse high-dimensional data, which is an attribute of the CNN features. SVM hyperparameters were optimized on a grid search with nested five-fold cross-validation. For the nested cross-validation, each training set was further divided into training and validation sets. The nested training set was used for training of classifiers with variable hyperparameters and the nested validation set was used for validation of these hyperparameters to choose the one that gave the best classifier performance.

In addition to performance evaluation of CNN-based and conventional CADx classifiers, we assessed the performance of a fusion classifier that integrated both CNN-based and conventional CADx classifier output scores. For this task, the outputs were fused by averaging them together

$$p(malignancy) = \frac{p(malignancy)_{CNN-based} + p(malignancy)_{Conventional\ CADx}}{2} \quad (3.1)$$

where $p(malignancy)$, $p(malignancy)_{CNN\text{-}based}$, and $p(malignancy)_{Conventional\ CADx}$ are the resulting probability of lesion malignancy for the fused, CNN-based, and conventional CADx classifiers. Diagnostic discrimination performance of the CNN-based and conventional CADx classifiers was

compared to their performance in combination. In order to assess statistical significance, DeLong tests were performed.[92] Bonferroni-Holm corrections were used to account for multiple comparisons.[93]

3.5.3. Performance Evaluation Metrics

The performances of the classifiers were evaluated by patient using receiver operating characteristic (ROC) analysis.[28, 37] AUC, a metric that is independent of cancer prevalence, served as the figure of merit and was computed with five-fold cross-validation. Within the cross-validation, training folds were standardized to zero mean and unit variance. The test folds were standardized with the statistics of the corresponding training folds.

3.5.4. Implementation Details

The ROI extraction was performed with the MATLAB software, developed specifically for the tasks. Feature extraction was implemented in Python (Python Version 2.7.12, Python Software Foundation) using the Keras library with a Tensorflow backend[95] on an NVIDIA Titan X GPU. Methods were evaluated with Scikit-learn package.[96]

3.6. Further Investigations

The previous section described the main procedure of this part of the dissertation. While designing the DCE-MRI radiomics methods, many additional thorough investigations were performed. Those involved studying the effect of the lesion representation on the CNN-based classification

performance as well as improving classification with the conventional CADx features. Studied lesion representations included lesion ROIs of different sizes, lesion ROIs at different DCE time points, and lesion ROIs selected on maximum intensity projection (MIP) images of 3D MRIs. Conventional CADx classifiers were improved with multi-task learning methods.

3.6.1. Effect of the ROI size on the CNN Classification Performance

The effect of the ROI size on the CNN-based and fusion classifiers' performances was evaluated. In addition to the tight-to-the-lesion ROIs, described in the methods Section 3.3.3, ROIs were cropped from the MRI slices with a constant pixel size across the entire dataset (Figure 3.7). The constant pixel size of the ROIs corresponded to the maximum dimension of the largest lesion in the dataset. For the University of Chicago DCE-MRI dataset, utilized for the malignancy assessment, the ROI size was 148x148 pixels. This approach led to many ROIs having most of the pixels representing breast parenchyma and not the lesion itself. The selected ROIs were utilized to extract CNN features, which were further used to train classifiers as described in Section 3.5.

We compared the CNN-based and fusion classifier performances based on the constant-size ROIs to the two classifier performances based on the ROIs selected tight to the lesion (Section 3.3.3).

3.6.2. Effect of the DCE Time Point on the CNN Classification Performance

In addition to the ROI size, we evaluated the effect of the DCE time point at which the lesion ROI was selected. Besides the RGB ROIs created with t0, t1, and t2 DCE time point ROIs, as detailed in Section 3.3.3., RGB ROIs were also created with the t1, t2 and t3 and with t0, t1, and t3 ROIs. ROIs selected at a single DCE time point, i.e. t0, t1, t2, and t3, were also evaluated. The CNN

features were extracted from each of the described ROIs and used to train separate classifiers. The influence of DCE time point on the classifier performance was assessed for the task of discriminating between benign and malignant lesions, as detailed in Section 3.5.

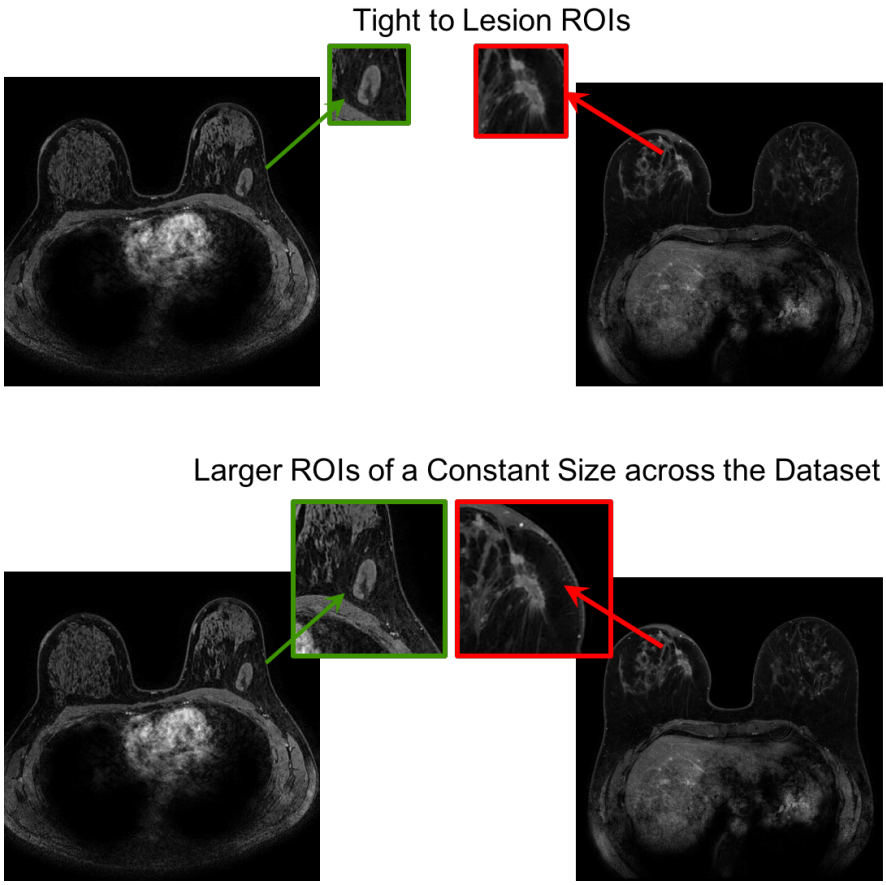


Figure 3.7. Variations of the selected ROI sizes. The results detailed above were achieved with the ROIs of size corresponding to the size of the enclosed lesion (top). Initially, the experiments were performed with the ROIs of constant 148x148 pixel size, which corresponded to the maximum dimension of the largest lesion in the dataset (bottom).

3.6.3. Maximum Intensity Projection Images for CNN-based Classification

In this part of the dissertation, the utility of MIP images (computed over all of 3D MRI volume with the extect of the tumor) was explored for the task of distinguishing benign and malignant

lesions. Clinicians often evaluate the extent of the entire lesion using maximum intensity projections of DCE-MRIs. This study utilized this idea and extended it to deep learning-based lesion evaluations.

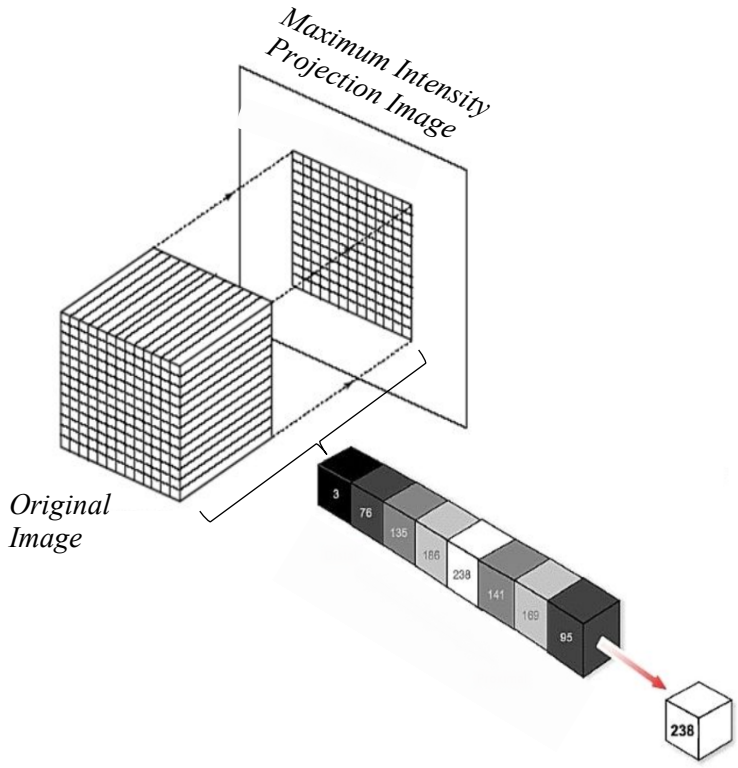


Figure 3.8. Illustration of maximum intensity projection (MIP) of a 3D image. The MIP image is obtained by taking the maximum value along the ray of projection, which is perpendicular to MIP.

The preceding methods of this chapter were developed based on the lesion ROIs selected on the central slice of the image volume. The fact that the image data had 3D structure was not fully exploited in the proposed classification pipeline. One way to leverage this structure would be to use a 3D CNN for the image classification tasks. However, the network has to be trained on the domain-specific data, which requires a very large number of training examples and can be extremely computationally expensive. This problem was solved by using the pre-trained 2D VGGNet, as detailed in the Section 3.4 above, with the 2D images, created by embedding 3D information into a 2D image using MIP. The MIP images served as a representation of the original 3D MRI volume.

To study the effect of MIP images on CNN-based lesion classification, we first selected ROIs surrounding each lesion on three MRI presentations: (i) the MIP image generated on the 2nd post-contrast subtraction (2nd post-contrast – pre-contrast) MRI, (ii) the central slice image of the 2nd post-contrast MRI, and (iii) the central slice image of the 2nd post-contrast subtraction (2nd post-contrast – pre-contrast) MRI. Examples of the three DCE-MRI representations for benign and malignant lesions are provided in Figures 3.9 and 3.10 To generate a MIP image, a subtracted 3D MR image is collapsed into a single 2D image by selecting the voxel having the maximum intensity along the projection through all transverse slices containing the lesion (Figure 3.7).[25,97]

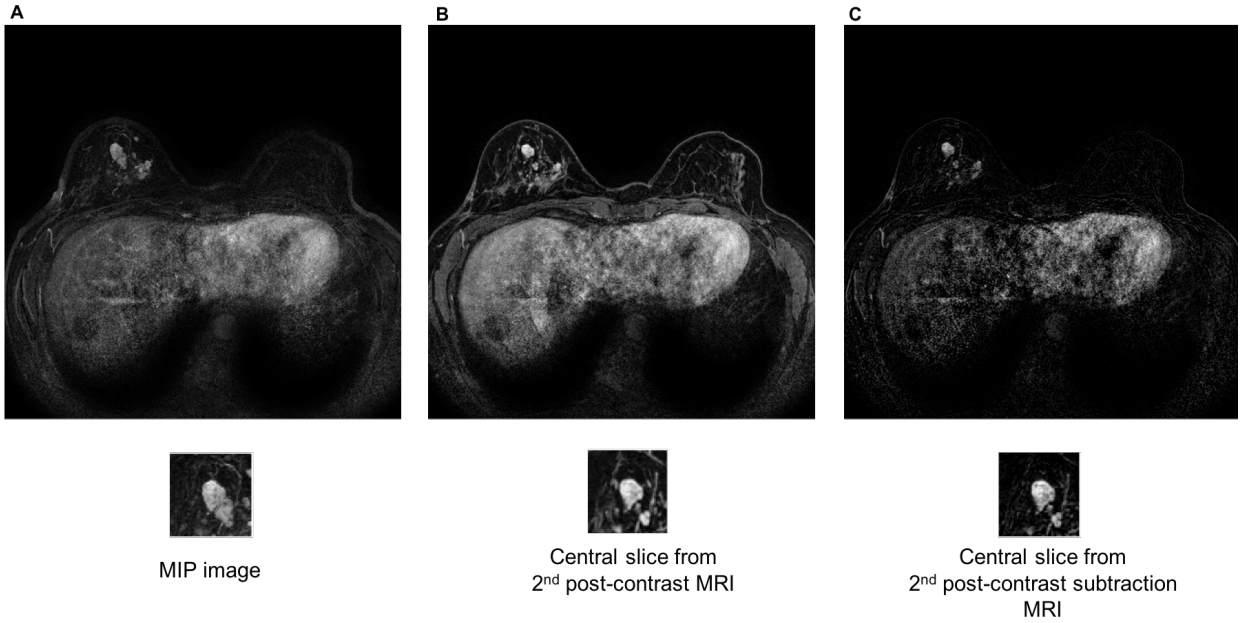


Figure 3.9. Example of a benign lesion with its three representations. Full MRI slices and ROIs for **A)** the MIP image of the 2nd post-contrast subtraction MRI, **B)** the center slice of the 2nd post-contrast MRI, and **C)** the central slice of the 2nd post-contrast subtraction MRI.

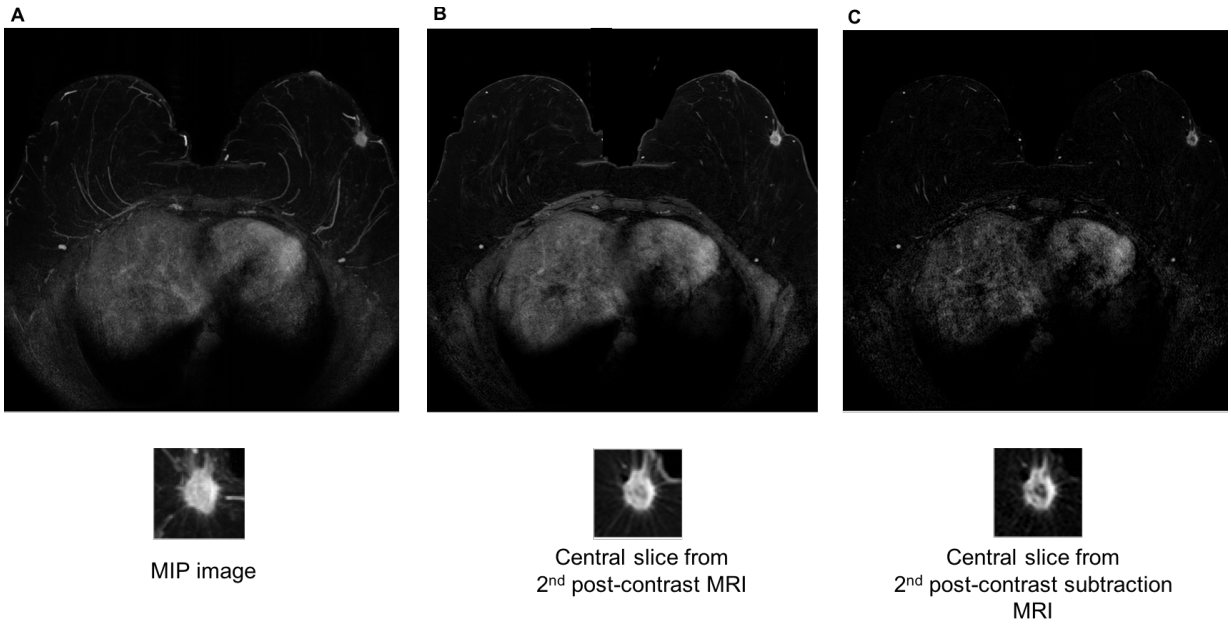


Figure 3.10. Example of a malignant lesion with its three representations. Full MRI slices and ROIs for **A)** the MIP image of the 2nd post-contrast subtraction MRI, **B)** the center slice of the 2nd post-contrast MRI, and **C)** the central slice of the 2nd post-contrast subtraction MRI.

Since the results of the previous methods (further described in Section 3.7.1) showed that the hierarchical pooled features outperformed the fully-connected features in the task of benign and malignant discrimination, we applied the pooled feature extraction to the three MRI representations (Figure 3.11). The classifications follow the methods described in Sections 3.4.1 and Section 3.5. Three separate SVM classifiers were trained and evaluated on the corresponding CNN features extracted from three types of ROIs. The significance of differences in classifier malignancy assessment performances was assessed with DeLong tests and corrected for multiple comparisons with Bonferroni-Holm corrections.[92,98]

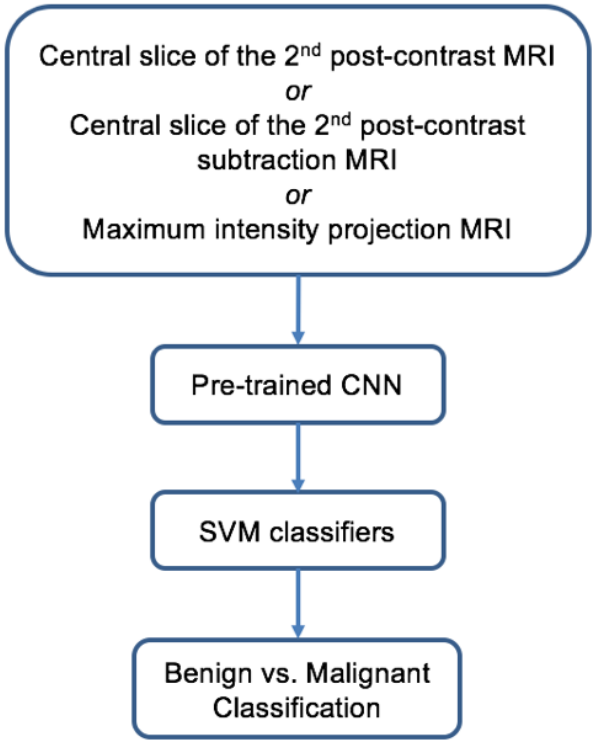


Figure 3.11. Lesion classification pipeline for the MIP image evaluation. Lesion ROIs were selected from three MRI representations: 1) central slice of the 2nd post-contrast MRI, 2) central slice of the 2nd post-contrast subtraction MRI, and 3) maximum intensity projection image of the 2nd post-contrast subtraction MRI. CNN features were extracted from the three representations and used to train separate SVM classifiers for the task of distinguishing benign and malignant lesions.

3.6.4. Multi-Task Learning for CADx Classifiers

As mentioned in Section 3.3.1, the DCE-MR images were acquired on two Philips scanners with varying magnet strength (1.5T and 3T). The classical CADx predictive models may get affected by the heterogeneity of the MRI data. Chapter 2 of the dissertation studied the robustness of the conventional radiomics across different MRI manufacturers. In this chapter, the image data are acquired on MRI scanners with different magnet strengths. The magnet strength has an effect on image quality, with lower-strength scanners producing images with higher noise and lower resolution, leading to less-resolved anatomical details.[99] Thus, radiomic features are expected to

vary with the MRI magnet strength. It is possible to train separate classifiers for the datasets acquired with different imaging conditions. Unfortunately, currently available medical image datasets are of limited size. Separating the DCE-MRI data based on magnet strength is undesirable since it would further decrease the dataset sizes.

The methods described in Section 3.5 used SVM models to perform lesion classification. Here, a methodology is proposed to enhance the SVM classifier performance by allowing heterogeneities in the datasets and using all of the data during its training. We hypothesize there exist underlying relationships between features extracted from images acquired with 1.5T and 3T scanners. Based on this assumption, we explore multi-task learning (MTL) to generalize discrimination of malignant and benign lesions over different feature domains. The approach assumes that there exists relatedness between the different-domain features, which is utilized to improve the classification performance.

We utilized a multi-task relationship learning approach described by Zhang and Yeung, which models relationships between the tasks to discriminate benign and malignant lesions.[100] The approach uses a prior on $\boldsymbol{W}$ (the matrix of the model weights), which is modeled with a matrix-variate normal distribution with the probability density function $p(\boldsymbol{W}|M, I, \Omega)$. Here, a covariance matrix Ω incorporates the relationships between the task-specific model weights. We compared classification performance of the multi-task algorithm to the conventional CADx SVM classifier (Section 3.5.2) in the task of distinguishing between benign and malignant lesions.

The MTL classifier is applied to the conventional CADx only. The deep learning-based MTL methods are left for future evaluations.

3.7. Results - Lesion Malignancy Assessment

The proposed methods were first evaluated on the DCE-MRI dataset for lesion malignancy assessment and this section presents the corresponding results. The method that showed the best classification was later applied to prediction of cancer's response to therapy. The results for the response clinical task are presented in Section 3.8.

3.7.1. Hierarchical pooled features vs. fully-connected features

Within the CNN-based methods, the classification performance of pooled features extracted from the original size ROIs was moderately stronger than that of fully-connected features extracted from preprocessed ROIs (Table 3.3). Fully-connected features extracted from the original ROIs with varying sizes resulted in much poorer classification performance. This is likely because the fully-connected features do not map to the same spatial location in the ROIs of varying sizes. Figure 3.12 shows the ROC curves for the two CNN-based classifiers. Since the hierarchical pooled features gave the best classification performance, those were used in further studies.

Table 3.3. Classification performance in terms of AUC of CNN features obtained from five max-pooling layers and from the first fully-connected layer. The methods are evaluated for the task of distinguishing benign and malignant lesions.

Fully-connected Features (no preprocessing)	Fully-connected Features (with preprocessing)	Max-pool Features
0.788 (se = 0.01)	0.809 (se = 0.01)	0.866 (se = 0.01)

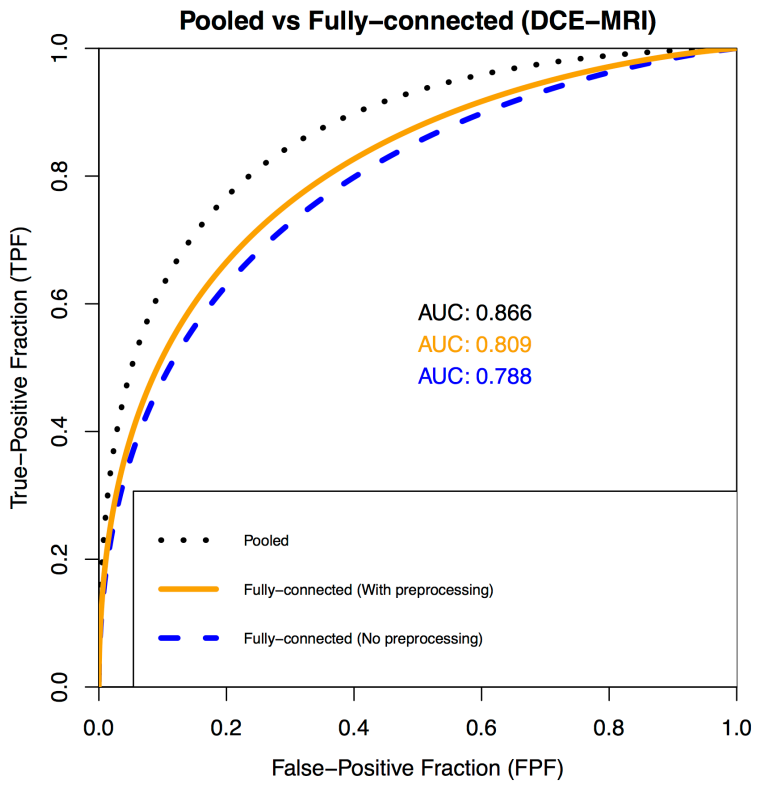


Figure 3.12. Fitted binormal ROC curves comparing the predictive performance of CNN-based classifiers.

3.7.2. Fusion of CNN-based Classifiers and Conventional CADx Classifiers

Since the CNN-based classifiers trained on pooled features performed the best in malignancy assessment, they were chosen as the final CNN-based classifiers to be fused with the conventional CADx classifiers. Fusion of the two types of classifiers outperformed any single type of classifier in the task of distinguishing benign and malignant lesions for each dataset. Table 3.4 and Figure 3.13 demonstrate the classification performances of CNN-based and conventional CADx classifiers individually and in combination. Figure 3.14 shows the classifier agreement levels for the two individual types of classifiers, as well as the potential decision boundaries for the fusion classifiers. Notably, there appears to be moderate disagreement between the CNN-based classifiers

and the conventional CADx classifiers, likely explaining why fusion improves predictive performance.

Table 3.4. AUC values for the benign vs. malignant lesion discrimination task for the CNN-based, CADx-bases, and fusion classifiers. P-values were corrected for multiple comparisons.

Conventional CADx Classifier	CNN-based Classifier	Fusion Classifier
0.86 (se = 0.01)	0.87 (se = 0.01)	0.89 (se = 0.01)
		p = 1.838e-06
		p = 0.026

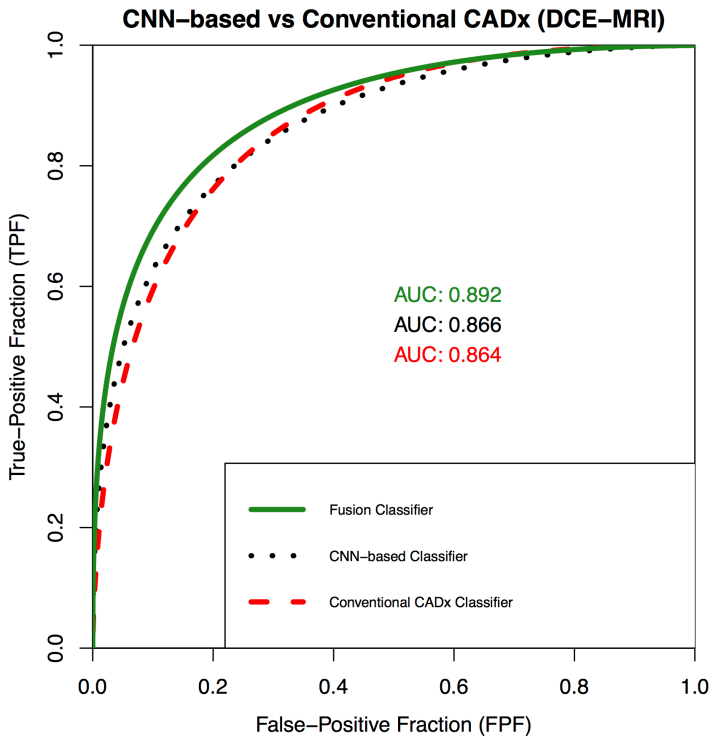


Figure 3.13. Fitted binormal ROC curves comparing the performances of CNN-based classifiers, conventional CADx classifiers, and fusion classifiers.

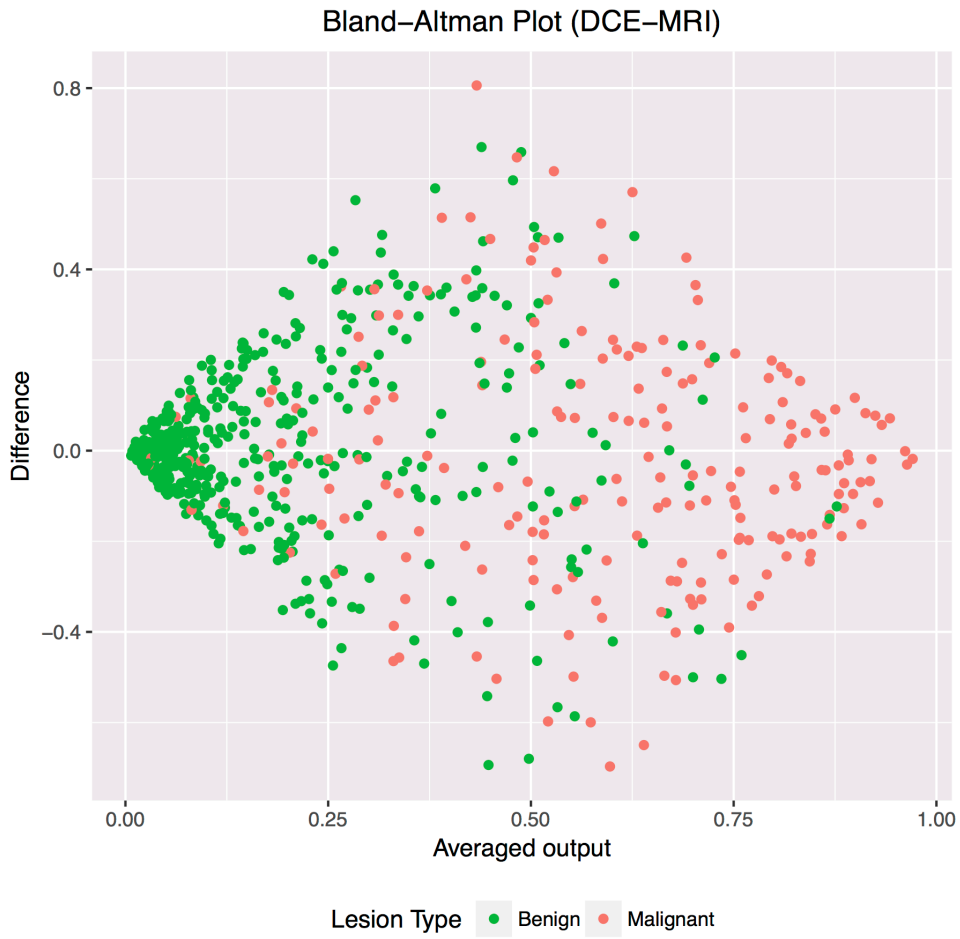


Figure 3.14. Bland-Altman plots to illustrate classifier agreement between the CNN-based classifier and the conventional CADx classifier. The y-axis shows the difference between the SVM outputs of the two classifiers; the x-axis shows the averaged output of the two classifiers. Since the averaged output is also the output of the fusion classifier, these plots also help visualize potential decision boundaries between benign and malignant classifications.

3.7.3. Effect of the ROI size on the CNN Classification Performance

Table 3.5 summarizes the performances of CNN-based, conventional CADx, and fusion (CNN+conventional CADx) classifiers in the task of distinguishing malignant and benign lesions to demonstrate the effect of the lesion ROI matrix size on the classification performance. Conventional CADx-based classifiers work with lesion ROIs, but are based on the entire DCE-

MRIs. However, we include their performances in the table for completeness. Since the results above showed that the hierarchical pooled features (extracted from the tight-to-the-lesion ROIs shown in the middle of Figure 3.15) performed better than the fully-connected features (extracted from the pre-processed framed ROI shown on the right of Figure 3.15), the CNN features were extracted from non-framed ROIs as detailed in Section 3.4.1.

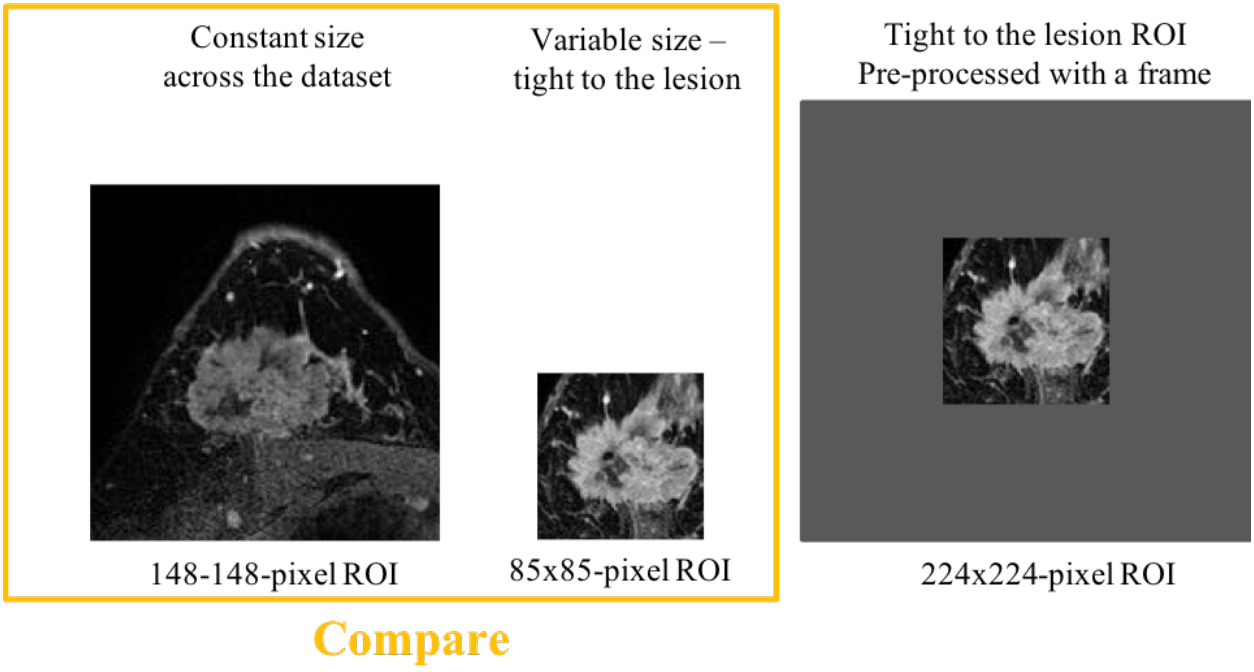


Figure 3.15. Various lesion ROIs studied in Chapter 3. Section 3.7.1 demonstrated that the hierarchical pooled features extracted from the tight to the lesion ROIs (middle) result in a significantly better performance than the fully-connected features extracted from those ROIs pre-processed with a frame (right). Thus, the effect of the ROI size on the classification performance was evaluated between the ROIs without pre-processing (left vs. middle ROIs).

For the task of distinguishing benign and malignant lesion, the classification performance of CNN-based classifier was worse when constant pixel-size ROIs were utilized, compared to tight-to-the-lesion ROIs. An AUC value of 0.72 (se = 0.02) was achieved with the CNN features extracted from the constant size ROIs, compared to the AUC value of 0.87 (se = 0.01) achieved

with the CNN features extracted from the tight-to-the-lesion ROIs (the results discussed in Section 3.7.1).

Similarly, the fusion classifier trained on the CNN features extracted from the constant size ROIs showed worse performance than the classifier trained on the CNN features extracted from the tight-to-the-lesion ROIs. For the fusion classifier, the same classification trend was observed for both types of the ROIs. Combining the CNN-based and conventional CADx classifiers improved the classification performance of each individual classifier alone. In particular, classification models constructed using the conventional CADx features achieved $AUC_{conventional\ CADx} = 0.86$ (se= 0.02); using the CNN features achieved $AUC_{CNN} = 0.72$ (se = 0.02); and using the conventional CADx + CNN features achieved $AUC_{fusion} = 0.88$ (se = 0.01). The SVM scores from conventional CADx and CNN-based models showed moderate correlation with a correlation coefficient of 0.27 with p-value<0.05 (Figure 3.16).

Table 3.5. Performances of CNN-based, conventional CADx, and fusion (CNN+conventional CADx) classifiers in the task of distinguishing malignant and benign lesions to demonstrate the effect of the lesion ROI matrix size on the classification performance. The performance is measured in terms of AUC. Note that the conventional CADx classification pipeline does not work with lesion ROIs and its classification performance is provided in the table for completeness.

ROI size	$AUC_{conventionalCADx}$	AUC_{CNN}	AUC_{fusion}
Constant across the dataset; corresponding to the largest lesion in the dataset	0.86 (se = 0.02)	0.72 (se = 0.02)	0.88 (se = 0.01)
Variable; corresponding to enclosed lesion size	0.86 (se = 0.02)	0.87 (se = 0.02)	0.89 (se = 0.01)

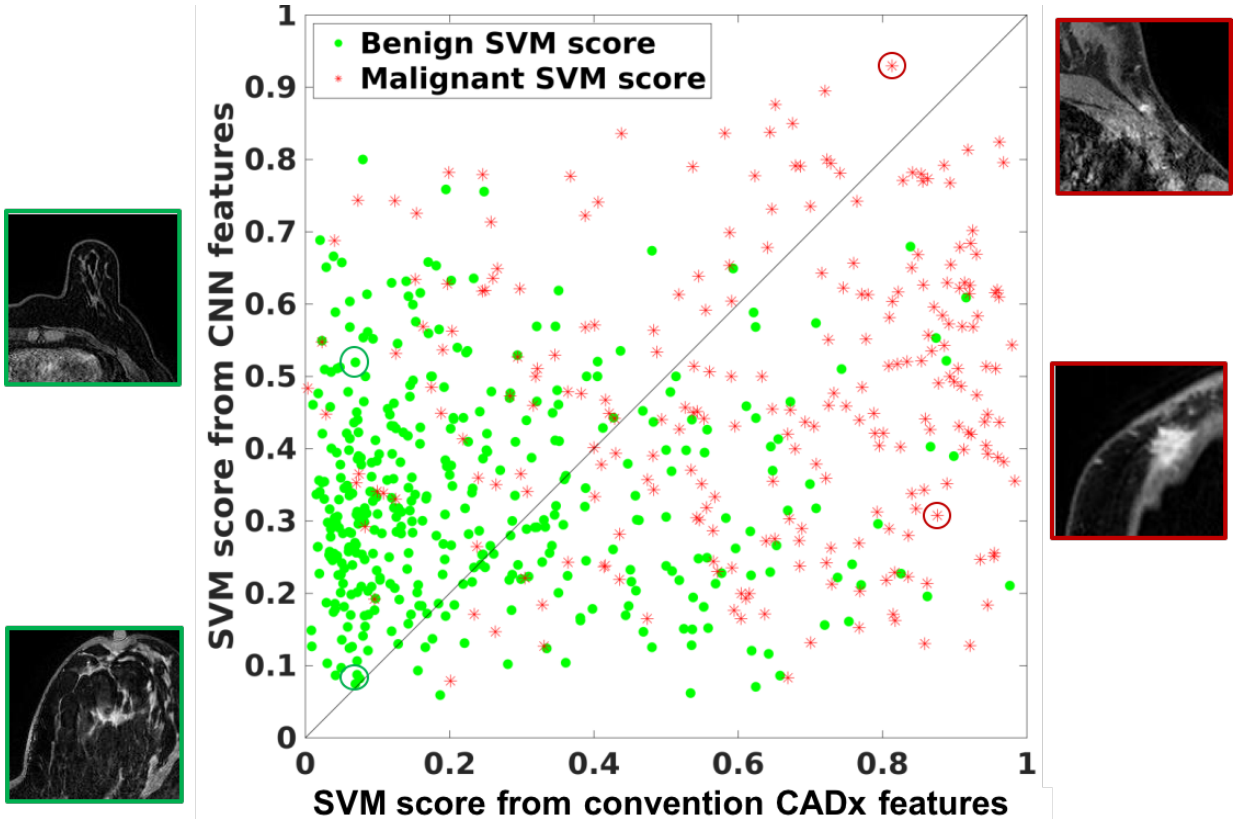


Figure 3.16. SVM scores for benign and malignant lesions for the conventional CADx vs CNN-based classifiers. The CNN-based classifier was trained on CNN features extracted from 5 max-pooling layers of VGGNet from the ROIs having a constant pixel size across the DCE-MRI dataset. Moderate correlation of scores is observed (r=0.27).

3.7.4. Effect of DCE Time Point on the CNN Classification Performance

Referring to Figure 3.17, the experiments showed that RGB ROIs resulted in superior classification performance to the single time point ROIs. The three RGB ROIs resulted in similar classification performance to each other, with AUC values ranging between 0.85-0.86 (se = 0.01). Since the RGB performances were not significantly different and since the dependency between pre-contrast and post-contrast images are critical for clinical evaluations, we chose the DCE time point

sequence that included the pre-contrast and 1st and 2nd post-contrast time points (t0+t1+t2) to form the RGB ROIs in all of our experiments.

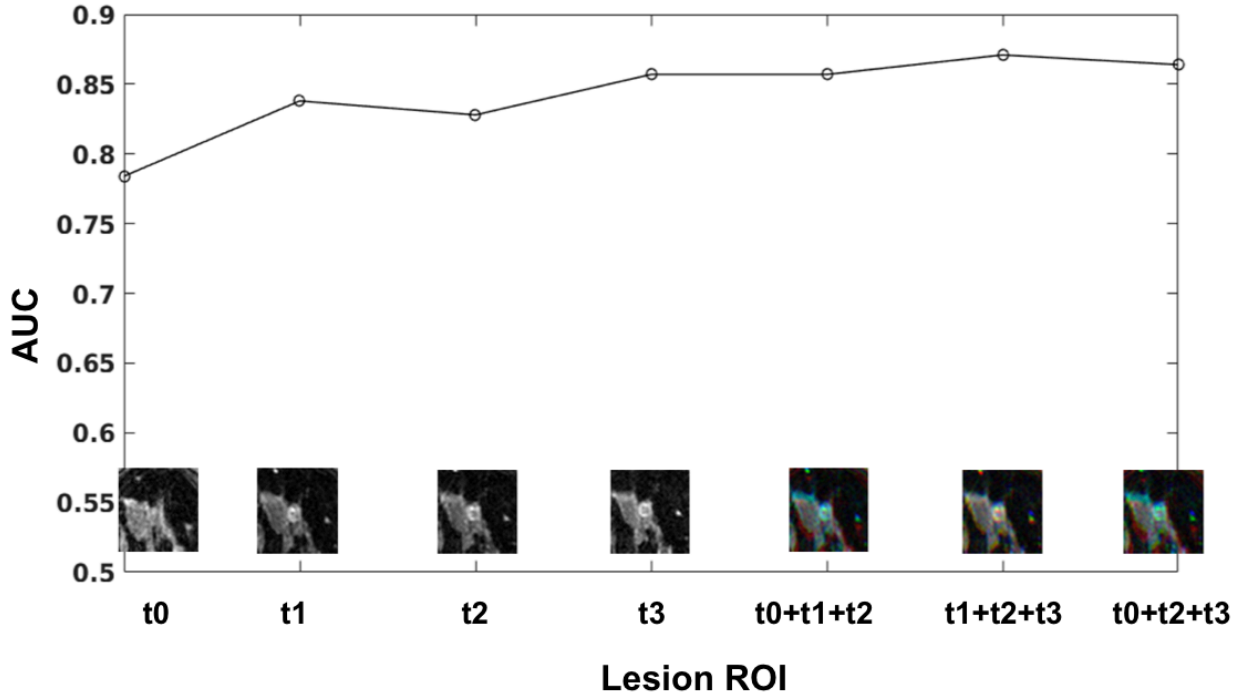


Figure 3.17. Classification performance of SVM classifiers trained using CNN features extracted from lesion ROIs selected at various DCE time points, pre-contrast (t0) and three post-contrast (t1, t2, t3), and from lesion RGB ROIs, formed by different combinations of single time point ROIs. The errors bars are omitted from the images, since all of the AUC values had se = 0.01.

3.7.5. Use of DCE-MRI Maximum Intensity Projection Images

Our results demonstrate that presenting MIP images, instead of single slices, to a pre-trained CNN leads to superior classification of lesion malignancy. Table 3.6 summarizes AUC values for the performances of the three classifiers in the task of distinguishing benign and malignant lesions corresponding to the three MRI representations: (i) the MIP image generated on the 2nd post-contrast subtraction (2nd post-contrast – pre-contrast) MRI, (ii) the central slice image of the 2nd

post-contrast MRI, and (iii) the central slice image of the 2nd post-contrast subtraction (2nd post-contrast – pre-contrast) MRI. SVMs trained on CNN features extracted from MIP ROIs significantly outperformed classifiers trained on CNN features extracted from ROIs from either the central slice of the 2nd post-contrast MRI or from the central slice of the 2nd post-contrast subtraction MRI. Figure 3.18 shows the ROC curves corresponding to the classification performances of the three classifiers. Our results suggest that the diagnostically useful information within the MRI volume can be successfully represented by maximum intensity projection images, further utilized in CNN-based classification methods.

Table 3.6. Classification performance of classifiers trained on CNN features extracted from three types of ROIs in the task of distinguishing malignant and benign lesions. P-values are computed with respect to MIP classifiers and are corrected for multiple comparisons with Bonferroni-Holm corrections.

ROI type	AUC	p-value
Central slice of 2nd post-contrast	0.80 (se=0.02)	0.00058
Central slice of 2nd post-contrast subtracted	0.83 (se=0.02)	0.048
MIP	0.88 (se=0.01)	-

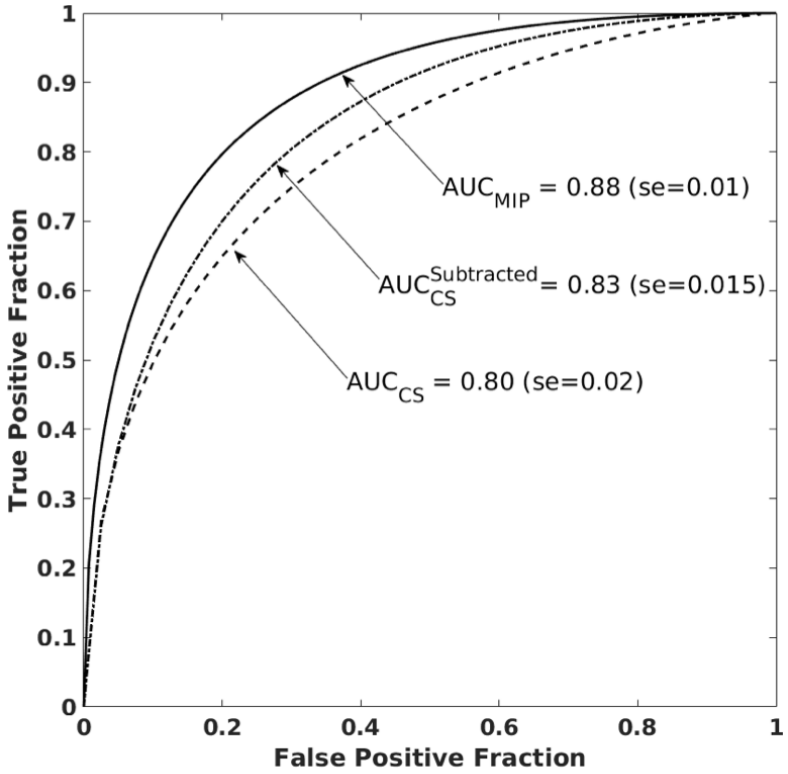


Figure 3.18. ROC curves showing the performance of three classifiers. The classifiers were trained on CNN features extracted from ROIs selected on: 1) the MIP images of 2nd post-contrast subtraction MRIs, AUC_{MIP}; 2) the central slices of the 2nd post-contrast MRIs, AUC_{CS}; and 3) the central slices of 2nd post-contrast subtraction MRIs, $AUC_{CS}^{Subtracted}$.

3.7.6. Multi-task Learning for CADx Classifiers

The performance of the MTL classifier was compared to that of the conventional CADx classifier trained on the 38 conventional CADx features. Table 3.7 summarizes performance metrics for the MTL and SVM classifiers while varying the decision boundary. Compared to the SVM classification, the MTL model results in higher positive (PPV) and negative (NPV) predictive values, and, therefore, it has more correct than incorrect identifications of benign and malignant lesions. Specificity values are higher for the MTL model, indicating reduction of false positive predictions. The reduction of false positives with the automated methods is important to note, since human DCE-MRIs evaluations result in a high false-positive number. We conclude that the MTL

classifier has higher performance than the SVM classifier with the resulting cross-validated AUC values of AUC_{MTL}= 0.90 (se = 0.01) and AUC_{SVM} = 0.87 (se = 0.01) for MTL and SVM classifiers, respectively.[101]

For the future work, the MTL classifier can be fused with the CNN-based classifier. Furthermore, a CNN-based multi-task learning pipeline can be designed to take into account variability in the images.[102,103] Multi-task CNNs are left for future work.

Table 3.7. Performance metrics values for classification of the merged dataset (1.5T and 3T) by MTL and SVM classifiers with the conventional CADx features. For a given sensitivity value, the MTL method outperforms the SVM method.

Sensitivity	Specificity		Positive Predictive Value (PPV)		Negative Predictive Value (NPV)	
	MTL	*SVM*	*MTL*	*SVM*	*MTL*	*SVM*
0.900	0.689	0.652	0.552	0.524	0.942	0.939
0.910	0.667	0.629	0.537	0.511	0.946	0.943
0.921	0.642	0.607	0.522	0.499	0.950	0.947
0.930	0.617	0.583	0.508	0.487	0.954	0.953
0.940	0.587	0.555	0.492	0.473	0.958	0.958
0.950	0.551	0.522	0.474	0.457	0.963	0.959
0.960	0.507	0.481	0.453	0.440	0.967	0.966
0.970	0.455	0.434	0.431	0.421	0.973	0.971
0.980	0.386	0.370	0.404	0.398	0.978	0.977
0.990	0.286	0.279	0.371	0.369	0.985	0.984

3.8. Results – Predicting Cancer Treatment Response

The response to therapy datasets were more challenging for the analysis than the dataset for the malignancy assessment task. The response to therapy datasets had an insufficient number of cases, limiting the deep learning-based analysis. Furthermore, the ISPY dataset had the additional challenge of comprising cancers with a wide range of sizes. These limitations gave us interesting

insights into the application of the conventional and deep learning-based radiomics and are described in details below. Figure 3.18 provides an overview of which dataset was utilized for the conventional and deep learning-based radiomics analysis.

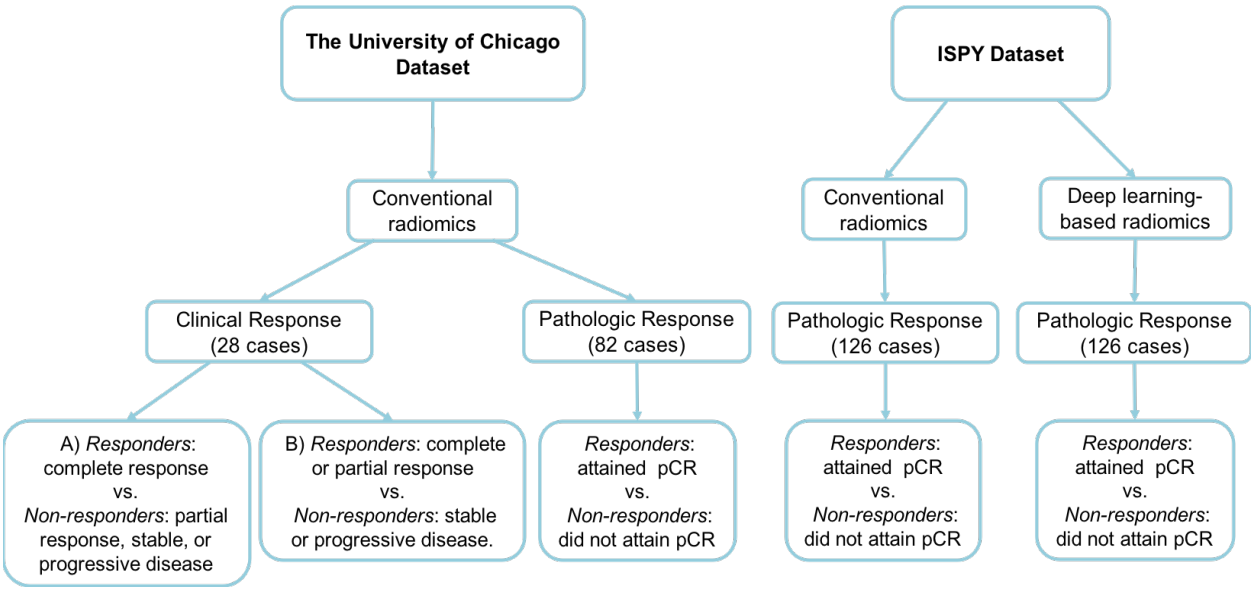
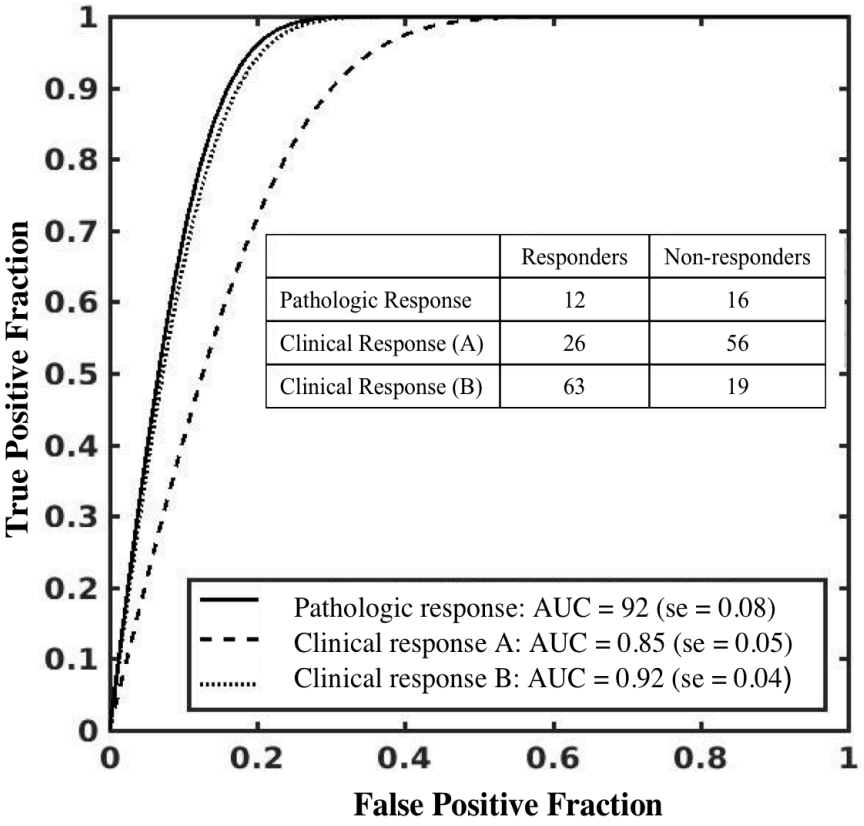


Figure 3.19. Response to therapy analysis flowchart. Two datasets, The University of Chicago and ISPY, were used in the development of radiomics methods for breast cancer response to neoadjuvant chemotherapy. Given a limited size of the University of Chicago dataset, only conventional radiomics methods were applied to it. Both conventional and deep learning-based radiomics were applied to the ISPY dataset.

3.8.1. Conventional CADx Classification

The University of Chicago dataset had a limited number of cases, with only 82 cases available with clinical treatment response and only 28 cases having pathologic treatment response. Therefore, only the conventional CADx was applied to evaluate its ability to predict both a cancer's pathologic and clinical response to neoadjuvant chemotherapy. For the pathologic response assessment, an AUC value of 0.92 (se = 0.09) was obtained in the task of distinguishing between responders and non-responders (Figure 3.20). For the clinical response assessment, AUC values were 0.85 (se = 0.05) for categorization A and 0.92 (se = 0.04) for categorization B. Our analyses

demonstrated that the most significant radiomic features for the response classification task described tumor shape and kinetics.



	Responders	Non-responders
Pathologic Response	12	16
Clinical Response (A)	26	56
Clinical Response (B)	63	19

Figure 3.20. The ROC curves demonstrating the performance of conventional radiomics in determination of breast cancer response to neoadjuvant chemotherapy for the University of Chicago dataset. The results are shown for the pathologic and clinical responses.

For the ISPY dataset, only pathologic, and not clinical, response to neoadjuvant chemotherapy was available and analyzed. The predictive models were built first with the CADx features extracted from the pre-treatment MR scans, then with the CADx features extracted from the first post-treatment MR scans, and finally with the percent change in CADx features from the two exams. The best predictive model of pCR was found to be a model with only one features selected, the percent change in the volume of most enhancing voxels feature (feature 32 in the

Table 2.2). The corresponding ROC curve for the distinguishing pCR from no pCR is shown in Figure 3.21. As mentioned previously in the methods section, the ISPY dataset was also analyzed for the task of predicting recurrence-free survival in breast cancer neoadjuvant chemotherapy. Interestingly, the most enhancing tumor volume was found to be the best predictor in the recurrence task. The C-statistics for the association of the most enhancing tumor volume with recurrence-free survival were 0.69 with 95% confidence interval of [0.58; 0.80] at pre-treatment and 0.72 [0.60; 0.84] at early treatment.

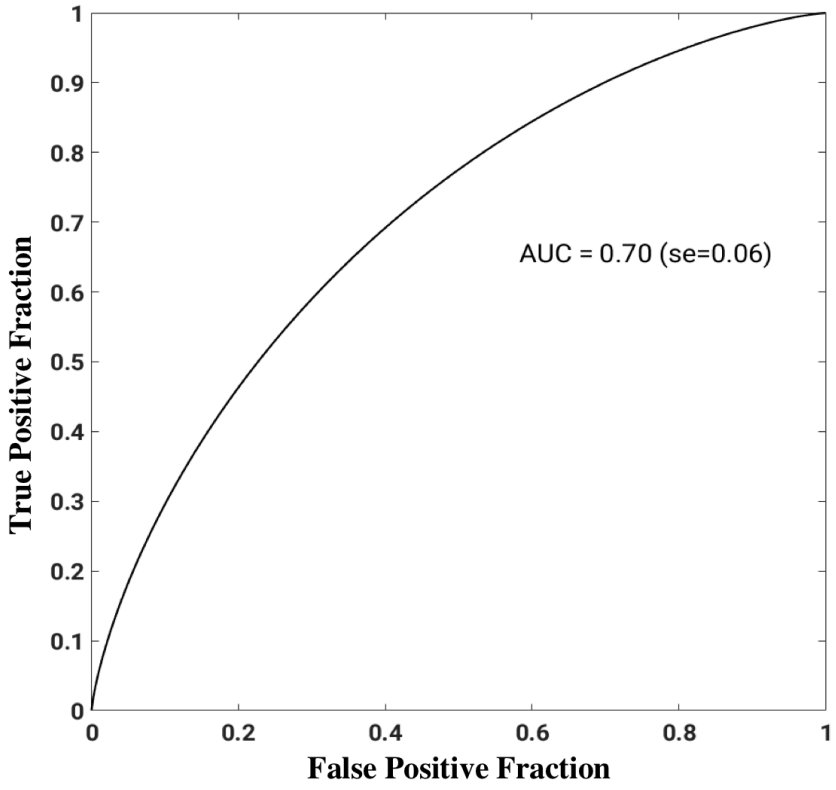


Figure 3.21. The ROC curve demonstrating the performance of conventional radiomics in determination of breast cancer response to neoadjuvant chemotherapy for the ISPY dataset. The results are shown for the pathologic. The most predictive feature was found to be the percent change in the volume of most enhancing voxels.

3.8.2. CNN-based Classification

Deep learning radiomics was evaluated only for the ISPY dataset collected under the ACRIN study. The University of Chicago response dataset was not utilized, since it had a small number of cases available.

Based on the results for the lesion malignancy assessment, the lesion representation that gave the best classification performance was achieved with the RGB ROIs as well as the ROIs selected on the MIP images of the second post-contrast subtraction MRI. These ROI selection techniques were applied to the ISPY MRI dataset, including both pre-treatment and first post-treatment MRIs. SVM classifiers trained on the CNN features extracted from the RBG ROIs achieved classification performance of 0.59 (se = 0.05) and 0.61 (se = 0.06) for pre-treatment and post-treatment exams, respectively. For the ROIs selected on the MIP images, the AUC values were 0.52 (se = 0.05) and 0.56 (se = 0.05) for the pre-treatment and post-treatment exams, respectively.

The ISPY dataset is challenging for deep learning methods. First of all, the images are old and cancers are of extremely variable sizes (Figure 3.22 and Figure 3.23). Some of the cancer cases are very large, resulting in large ROIs, which include skin line and image background. Figure 3.22 shows examples of three lesion ROIs selected around one small, one intermediate-sized, and one large lesion. Based on the results from Section 3.8.1 on the effect of the ROI size on CNN-based lesion classification, larger ROIs show worse classification performance. However, the large ROI size was unavoidable for many of the ISPY cases, since many of the lesions extended over the entire breast.

Given a poor CNN feature performance, the fusion of CNN-based and conventional CADx classifiers was not studied for the task of prediction of cancer's response to therapy.

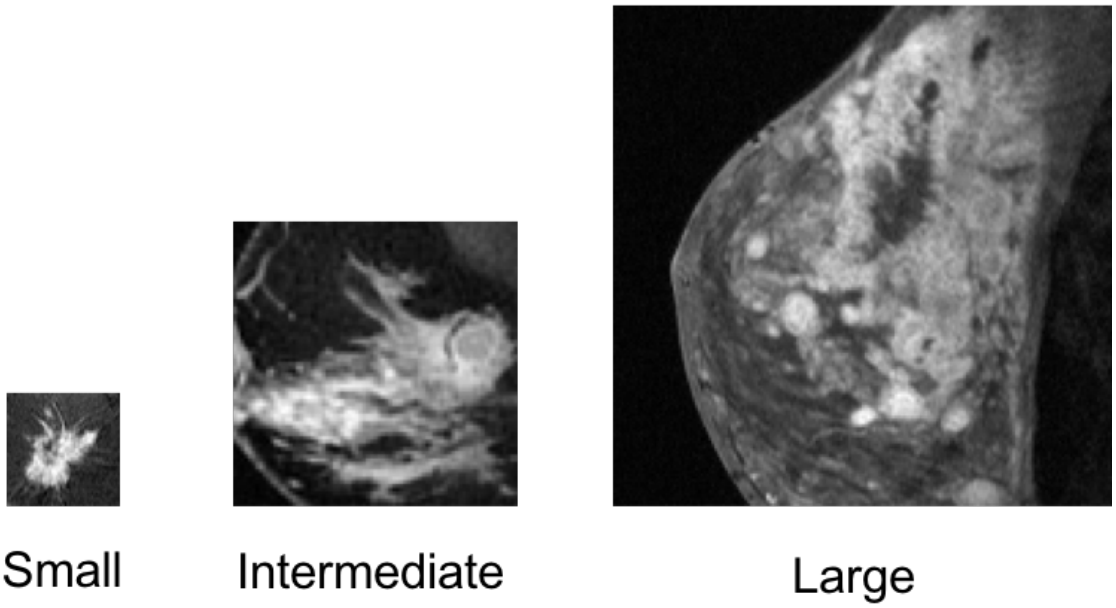


Figure 3.22. Examples of three lesion ROIs selected around small, intermediate-sized, and large lesions. The small ROI includes the lesion and a small part of the breast parenchyma; the intermediate-sized ROI includes the lesion and a larger amount of breast parenchyma as well as parts of the skin; the large ROI contains almost the entire MRI slice.

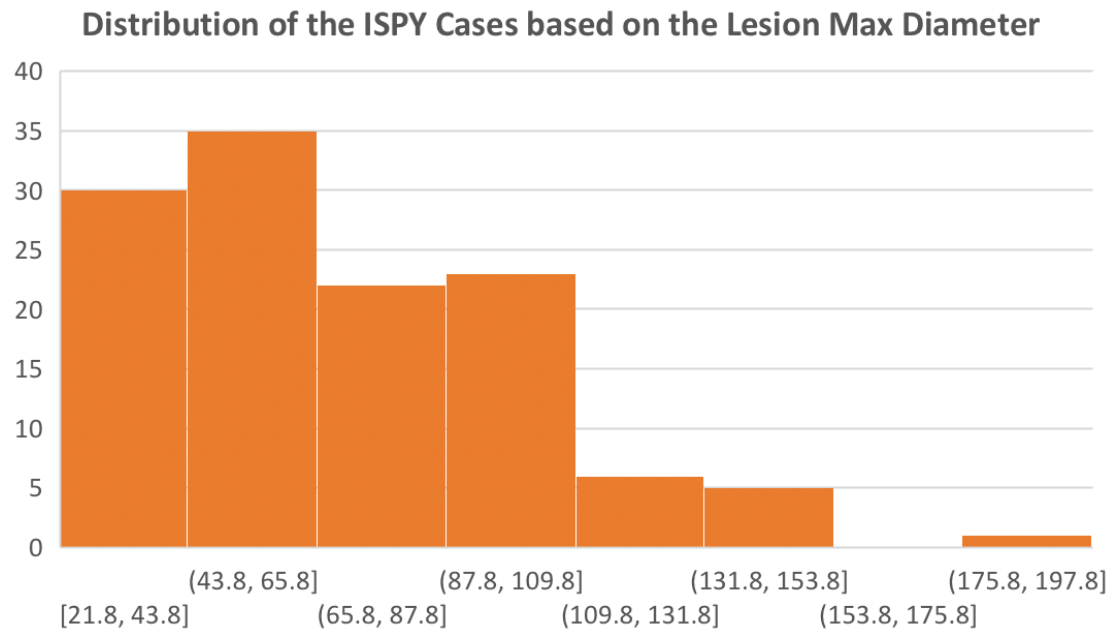


Figure 3.23. Distribution of the IPSY cases based on lesion's maximum diameter. The maximum diameter is calculated by the Giger Lab quantitative radiomics workstation based on the lesion segmentations and is measured in mm. The diagram shows that the range of the lesion sizes in the ISPY dataset is wide, with the presence of many extremely large lesions.

3.9. Discussion and Conclusions

We have shown that classifiers trained on deep features and existing conventional CADx features can be fused to significantly improve predictive performance in the task of breast lesion diagnosis based on DCE-MRIs. Further, we demonstrated that our multi-layer feature extraction methodology outperforms the commonly used approaches to deep feature extraction in addition to not requiring image preprocessing. We found that when extracting fully-connected deep features from images, different *ad hoc* preprocessing techniques required evaluation to maximize performance: DCE-MR images worked best with average pixel padding. By circumventing the need to resize images, our methodology is more generalizable across different datasets, while also

commanding stronger predictive performance, less sparsity, and lower dimensionality.

To our best knowledge, this is the first development of a hybrid technique involving hierarchical deep feature extraction and conventional CADx methods. A previous paper from our laboratory[82] investigated the feasibility of using pre-trained CNNs and how they compare with conventional CADx methods, but only on a FFDM dataset. In this work, we used a novel method of feature extraction inspired by the work of Zheng et al.,[88] involving average pooling of extracted features from multiple CNN layers in order to considerably reduce dimensionality while preserving spatial structure and occlusion invariance. Our method avoided using higher-level layers as Zheng et al. did, since higher-level layers are more specific to the original ImageNet task of general object recognition. For the purpose of CADx and medical images, lower-level layers appear to be of greater importance, and our moderately sized datasets restricted us from freely including extra layers and parameters. Other papers have used CNNs for computer-aided diagnosis with success,[82,83,104–106] but did not provide baseline comparisons with conventional CADx methods.

It is important to note that we used CNNs as fixed feature extractors instead of training them from scratch or fine-tuning them.[80] Our motivation for doing so is threefold: (1) *Computational time*: we evaluated the methods' computation times on an ultrasound dataset, used in adjacent studies. Using an NVIDIA GeForce GTX 970, feature extraction for the ultrasound dataset of 2393 images took approximately 3 minutes; fine-tuning a CNN on the dataset takes between 8-24 hours of training time. In an applied setting, models need to be retrained upon receiving new data, so the training time is nontrivial. (2) *Validation*: Our preliminary results involving fine-tuning underperformed our feature extraction methods, but we are aware that fine-tuning can often

outperform generic feature extraction given the proper circumstances (e.g. ad hoc optimizations of hyperparameters/architectures, augmentation techniques, sufficient sample sizes). However, the slow training time of fine-tuning limits the efficiency of data usage: standard validation procedures when training CNNs typically only separate the data into a single training set and a single test set. With feature extraction, we are able to use more rigorous validation procedures like k-fold cross validation or bootstrapping, resulting in a more precise and reliable model. (3) *Generalizability*: Medical images vary dramatically based on institution and manufacturer. Consequently, it is important to have a method that quickly generalizes across these differences without overfitting to trivial nuances unique to single institutions or manufacturers. Our method only uses generic CNN features and radiomic features, eliminating the need to retrain a new CNN and use ad hoc hyperparameter optimizations for every new dataset. Shin et al.[105] reported that fine-tuning substantially outperformed feature extraction in the task of computer-aided detection, but they only extracted features from the final layer of AlexNet. Other works have shown that the final layer of AlexNet is significantly inferior to earlier layers for the task of medical image analysis.[82,107] Our method employs a more advanced feature extraction technique by hierarchically integrating multiple layers from VGGNet in order to incorporate low-, mid-, and high-level information from images. It therefore remains unclear how fine-tuning and feature extraction perform in comparison with each other.

There were several limitations to our study. While we used VGGNet, other networks, such as deep residual networks,[47] have shown greater performance and promise for transfer learning, but their depth (upwards of 1000 layers) and complexity make investigating their potential for CADx out of the scope of this study, especially due to the moderate sizes of our datasets. Further, our

datasets all came from one medical center. Due to the heterogeneity resulting from different imaging manufacturers and facility protocols, it is unknown whether our classifiers would test well on images from another institution. Additionally, the selection of contrast time points for the DCE-MRI was suboptimal. In some of our preliminary experiments, we found that other combinations of contrast time points may perform better than the one we chose (t0, t1, and t2), warranting further investigation. The work of Chapter 4 addresses this problem by developing a deep learning classification pipeline that allows for the use all DCE time points.

Our results also demonstrated that deep learning radiomics is reliant on the data input. We proposed a method to incorporate volumetric and partially temporal components of DCE-MRI for classifying lesions as benign or malignant using MIP images with pre-trained CNNs. Specifically, the method involves subtracting a pre-contrast MRI from the 2nd post-contrast MRI, and then calculating the maximum intensity projection of the subtraction image. As a result, the MIP image contains information about enhancement changes throughout the lesion volume. This method can be easily adopted in clinical practice, since MIP images are already commonly used in the evaluation of breast tumors.

The analysis of the DCE-MRI datasets for the task of response to therapy assessment allowed us to draw important conclusions. We saw that the deep learning methods are heavily reliant on the number of cases available and on the image data itself. The deep learning-based methods with CNN, used either as a feature extractor or fine-tuned, were not able to predict the breast cancer response to neoadjuvant chemotherapy. We hypothesize, that the reason for it is the combination of the limited dataset size and the heterogeneity in the image data. On the other hand, the conventional radiomic features were able to predict the response to therapy with moderate

accuracy (AUC = 0.70). These results demonstrated that the conventional radiomics might be more useful, when the deep learning models fail.

In summary, we demonstrated the feasibility of using deep feature extraction techniques in CADx for breast DCE-MRI. Moreover, we developed a system incorporating both deep learning and conventional CADx methods that performed statistically significantly better than either one separately. Our methodology is computationally efficient and does not require intensive image preprocessing. Given the rapid progress of deep learning, our intent is not that our exact methodology be incorporated in clinical practice, but that our proposed solutions to the challenges of efficiency, precision, and preprocessing help pave the way towards more effective CADx methods.

CHAPTER 4

INCORPORATION OF DCE-MRI TEMPORAL COMPONENT INTO DEEP LEARNING-BASED RADIOMICS

4.1. Introduction

Chapter 3 developed deep learning methodology for lesion classification based on DCE-MRIs based on partial DCE-MRI data. In this chapter, we propose an automated deep learning-based methodology that captures not only tumor morphological characteristics from 2D images, but also the temporal enhancement changes presented in dynamic MRI sequence. Compared to MIPs, the methodology explicitly utilized the sequential dependencies present in the DCE-MRI sequence. Furthermore, it can work with the DCE-MRI sequences of any temporal length. The combination of the two components of DCE-MR images allows for more accurate breast cancer diagnostic decision-making.

Over the course of the dissertation research, we observed deep learning approaches, specifically deep convolutional neural networks (CNNs), becoming state-of-the-art methods in many computer vision tasks, including medical image classification and segmentation.[108] CNNs consist of multiple transformation layers (e.g. convolutional, pooling, fully-connected), which extract features from pixel-level data, generating new image representations in their respective feature spaces. Image features extracted from earlier layers of CNNs are more general and are related to local image structures, such as edges and shapes.[55] On the other hand, later layers, such as fully-connected layers, are more class-specific and responsible for representing increasingly more abstract features, hierarchically composed of lower-level features. CNNs have shown great

success in standard image classification tasks[46,87] and have been adapted in medical image analysis to improve accuracy and speed of image-based diagnosis and prognosis.[44] Chapter 3 of this thesis demonstrated their utility for breast cancer assessment based on DCE-MRIs. Training an accurate and generalizable CNN requires large amounts of data. Due to the lack of large-scale medical image datasets, medical analyses have been frequently performed with CNNs pre-trained on a natural image dataset, such as ImageNet.[82,109–111] The typical approach to using pre-trained CNN is to fine-tune the last fully-connected layers of the network for the medical classification tasks. Fine-tuning is limited in dimensionality and is performed on 2D images, making it difficult to apply it to imaging exams that have temporal or volumetric components.

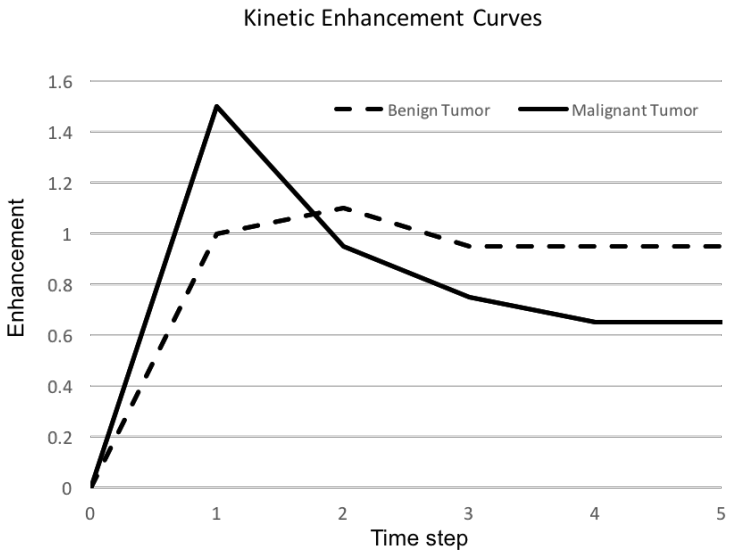


Figure 4.1. Lesion contrast enhancement curves. Benign and malignant lesions tend to have different enhancement patterns.

In this work, we propose a deep learning-based methodology that enables incorporation of the entire temporal component of DCE-MRI sequences. Many breast lesion diagnostic and prognostic decisions often rely on lesion enhancement over time, as shown in DCE-MRIs. The sequential imaging yields enhancement patterns that carry clinically useful information (Figure 4.1). For example, in lesion malignancy assessment, benign lesions typically have moderate uptake and slow washout of the ejected contrast agent, while malignant lesions have both rapid uptake and washout. Therefore, the focus of this work is to incorporate the

temporal component of the dynamic contrast-enhanced MRIs into breast lesion classification by using deep learning methods.

We achieve our goal by training a recurrent neural network (RNN), specifically long short-term memory (LSTM), which exploits the temporal correlations. We train the LSTM on sequences of feature vectors extracted from dynamic MRIs with a pre-trained CNN.[112,113] Higher-level CNN features represent information important for class discrimination. On the other hand, lower-level CNN features possess local pattern information valuable for further differentiating within a given class. [88,114] RNNs perform classifications based on sequences of input data (image feature vectors in our case) and rely on the fact that a sequence itself carries useful information for a given task. Thus, in order to capture the lesion enhancement changes presented in MRI images of a given DCE-MRI sequence, we form each image feature vector by concatenating features from various levels of pre-trained CNN. We compare LSTM method's performance to that of CNN fine-tuned on 2D MRIs. The results suggest that incorporation of enhancement patterns observed over the dynamic MRI sequence into lesion classification with deep learning methods improves malignancy assessment for breast cancer.

4.2. Methods

4.2.1. DCE-MRI Dataset

The proposed method was demonstrated for the task of discriminating malignant and benign lesions on a dataset of 703 DCE-MRI cases. The DCE-MRI dataset contains the 690 cases studied in Chapter 3. The dataset is continuously growing. The Giger lab has acquired more cases since

the competition of the research of Chapter 3. Therefore, Chapter 4 performs experiments on a slightly larger dataset. The dataset was retrospectively collected under a HIPAA-compliant Institutional Review Board protocol and annotated as benign and malignant based on pathology or radiology reports. Other clinical characteristics of the dataset are detailed in Table 4.1.

Table 4.1. Clinical characteristics of the DCE-MRI dataset studied for benign vs. malignant lesion discrimination with long short-term memory networks. Compared to the dataset utilized in Chapter 3, the DCE-MRI data has 703 breast cases.

Benign/Malignant Prevalence: # of cases (%)	Benign: 221 (31.4%) Malignant: 482 (68.6%) Total: 703
Age: mean (STD)	54.5 (13.2) Unknown: 103
Benign Tumor Characteristics	
Tumor Sub-types:	Fibroadenoma: 92 Fibrocystic change: 79 Papilloma: 14 Unknown: 36
Malignant Tumor Characteristics	
Tumor Sub-types:	Invasive ductal carcinoma: 135 Ductal carcinoma in situ: 20 Invasive ductal carcinoma + ductal carcinoma in situ: 264 Invasive lobular carcinoma: 20 Invasive lobular carcinoma mixed: 19 Unknown: 24
Estrogen Receptor Status: # of cases	Positive: 328 Negative: 108 Unknown: 46
Progesterone Receptor Status: # of cases	Positive: 274 Negative: 159 Unknown: 49
HER2 Status: # of cases	Positive: 72 Negative: 349 Equivocal: 3 Unknown: 58

DCE-MR images were acquired on 1.5 Tesla and 3 Tesla Philips Achieva scanners with T1-weighted spoiled gradient-echo sequence over the period of ten years, 2006-2016. Image slice thickness varied across the dataset, with 469 cases (66.7%) having slice thickness of 2 mm and 234 cases (33.3%) having slice thickness 1.5 mm or 1.6 mm. The image sequence included one image (pre-contrast) acquired prior to and multiple images (post-contrast) acquired after contrast injection.

Prior to CNN fine-tuning and image feature extraction, we selected regions of interest (ROIs) around each lesion for each DCE-MRI slice and time point (Figure 4.2). The ROIs were selected based on lesion segmentations from the 4D MRIs, performed prior to this research with Fuzzy C-means algorithm.[40] These ROIs around a lesion were delineated from each transverse slice of the 3D lesion image and from each DCE time point (those include pre-contrast, t_0, and multiple post-contrast time points, $t_1 ... t_n$). The ROI coordinates were unchanged across DCE time points. The number of ROIs for an individual lesion varied based on the number of slices containing the lesion and on the dynamic sequence length.

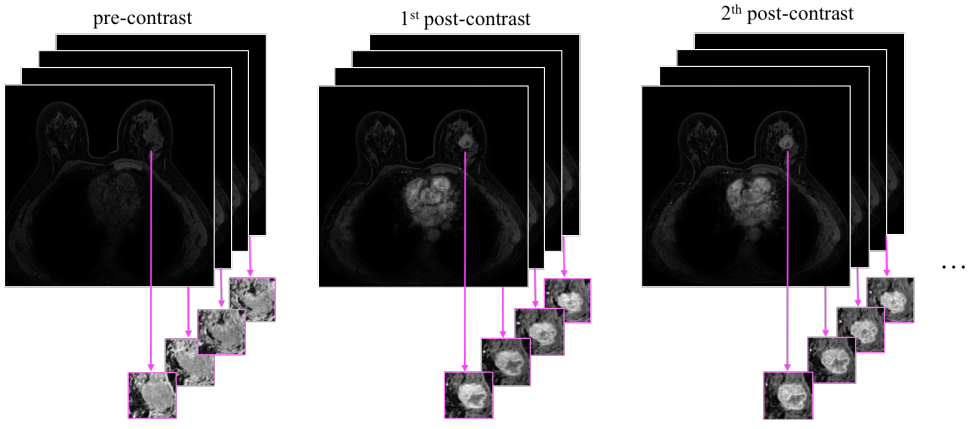


Figure 4.2. Example of a DCE-MRI sequence with ROIs selected around the lesion at each DCE time point and each slice containing the lesion.

4.2.2. CNN as a Feature Extractor vs. CNN Fine-tuning

As in the work of Chapter 3, a 19-layer VGGNet, pre-trained on ImageNet, was used as a base model.[42,46] It was applied as a feature extractor to the lesion ROIs selected from DCE-MR images (Figure 3.4). However, in Chapter 3, we did not evaluate whether the proposed feature extraction method would outperform VGGNet fine-tuned on the DCE-MRI dataset. In this part of the research we compared these two approaches (Figure 3.4 vs. Figure 4.3).

To perform fine-tuning, we utilized the convolutional base of the VGGNet and added a new fully-connected top. Since our dataset consisted of ROIs of various sizes, we adapted a global average-pooling layer after the last convolutional block of VGGNet.[115,116] The global average pooling layer is an extreme version of typical average-pooling layer, where it performs average pooling across the spatial dimension of the feature maps. Compared to the typical average-pooling layer, which reduces the spatial dimension of the input tensor by a given factor *n*, the global average-pooling layer takes an input tensor of size *h* X *w* X *d* and outputs a tensor of a fixed size of *1* X *1* X *d*. In our architecture, the average-pooling layer was followed by two fully-connected layers, with dropout applied after each of the two layers. All layers, prior to and including the 4th max-pooling layer, were frozen and the rest were fine-tuned. The schematic in Figure 4.3 demonstrates the VGGNet fine-tuning for lesion classification.

Fine-tuning was simultaneously performed on pre-contrast, 1st and 2nd post-contrast ROIs. VGGNet requires an image input consisting of three channels, red (R), green (G), and blue (B). Since pre-contrast and 1st and 2nd post-contrast ROIs are grayscale, we made use of the network's color channels and input these ROIs into the R, G, and B channels, respectively. Thus, we fine-

tuned the network on these artificially made RGB lesion ROIs, formation of which was described in the detail in Section 3.3.3.[110]

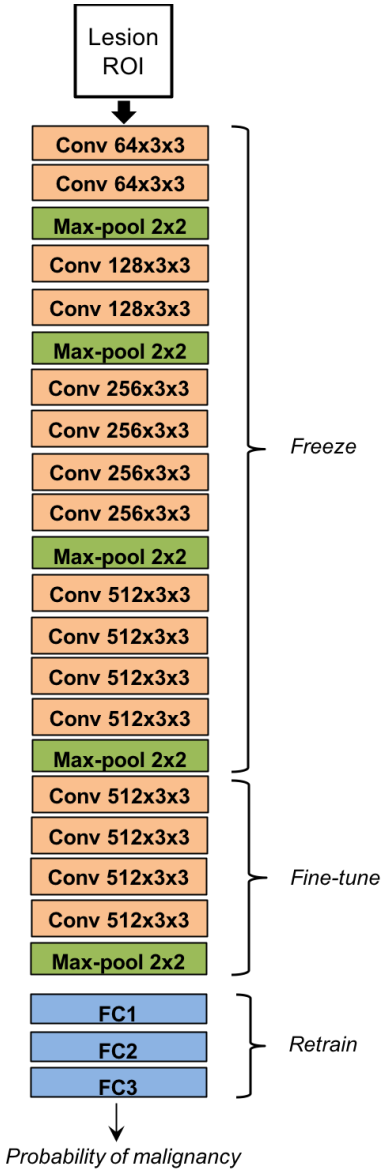


Figure 4.3. Lesion classification methodology. VGGNet was fine-tuned on RGB ROIs (RGB ROI is formed by ROIs at the pre-contrast, first and second post-contrast DCE time points). Its performance was compared to the methodology proposed in Section 3.4.1, where VGGNet was used as a feature extractor from the lesion ROIs (Figure 3.4).

4.2.3. Multi-level Image Features

The proposed LSTM model was trained on the sequences of the feature vectors extracted from the sequences of lesion ROIs with the fine-tuned VGGNet. From each ROI, CNN features were extracted by the method detailed in Section 3.4.1. In order to capture intra-class changes, i.e. contrast-enhancement changes of one lesion, the feature vectors were extracted at various network depths from the five max-pooling layers of VGGNet. These features from each level were average-pooled and normalized with Euclidian distance. The pooled features were further concatenated to form a CNN feature vector for a given ROI.[110]

For a given slice of a 3D MRI, the image feature vectors were extracted at each DCE time point. The resulting sequence of feature vectors was used as an input into LSTM network. Figure 4.4 demonstrates the lesion classification pipeline with LSTM network based on DCE-MRIs.

4.2.4. Long Short-Term Memory Network

The multi-level feature vector sequences were utilized to train an LSTM network. During its training the model captures the changes presented in a given sequences. Let x_0, x_1, x_2... x_n represent a sequence of n inputs, where each x_i is an input at time-step t=i. In our work, x_i represents an image feature vector obtained from lesion ROI at DCE time point i. An RNN has an internal hidden state at time t, h_t, which gets updated based on the current input x_t and its previous hidden state h_{t-1} (Figure 4.5).

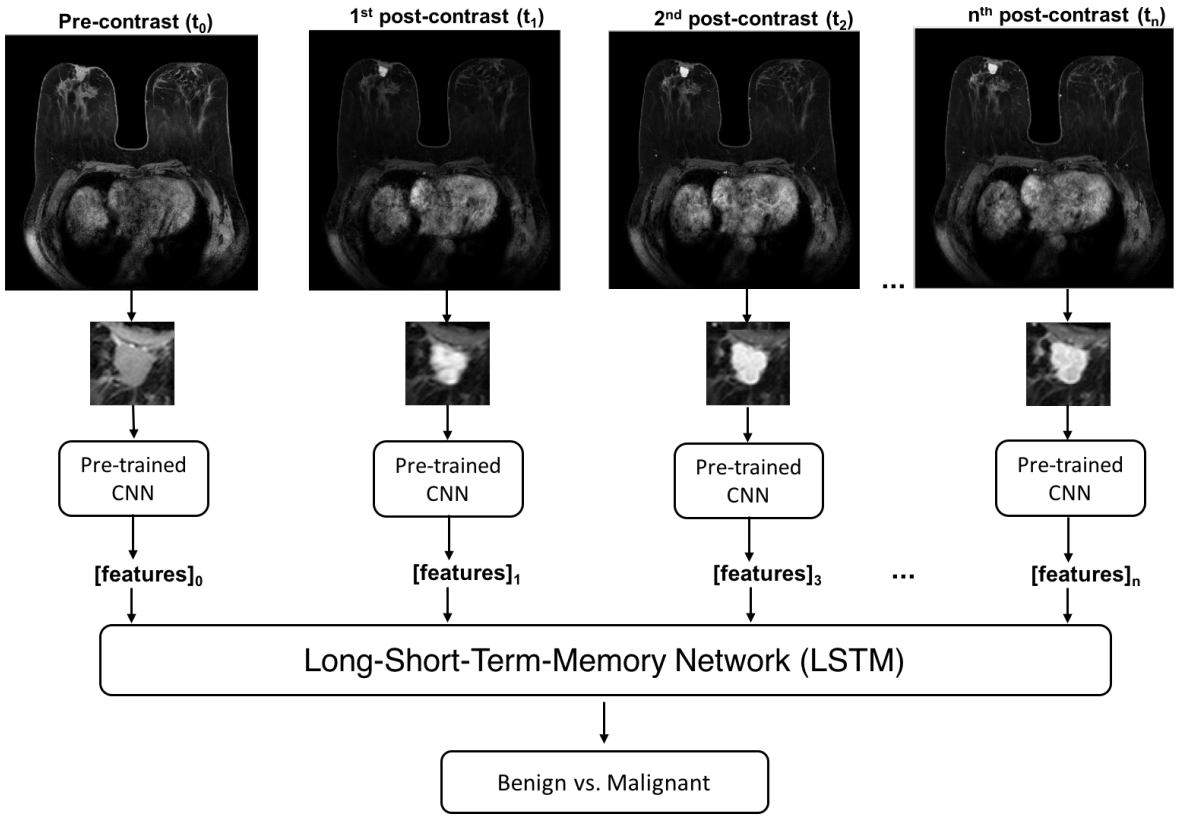


Figure 4.4. Lesion classification methodology. Image features were extracted from various levels of VGGNet from the lesion ROIs at each DCE time point and utilized for LSTM network training.

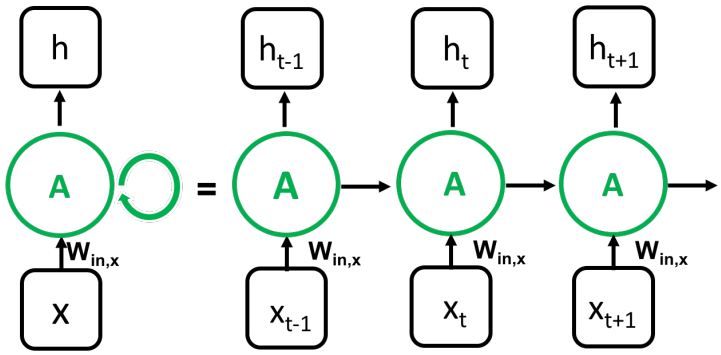


Figure 4.5. General structure of a recurrent neural network. The network recurrently computes its hidden state $\boldsymbol{h_t}$ based on its previous hidden state $\boldsymbol{h_{t-1}}$ and the current input $\boldsymbol{x_t}$. The final classification output is computed based on the hidden state of the network, which depends on the previous steps.

An LSTM, a type of RNN, takes this idea further by maintaining an additional distinctive feature, a 'memory cell.' Along with the hidden state h_t, a memory cell's state c_t is updated as the network steps through the sequence of the inputs. This update is based on the previous step's hidden state $\boldsymbol{h_{t-1}}$ and the current input $\boldsymbol{x_t}$,, and is performed by mechanisms called gates, i.e., the 'input gate,' the 'forget gate,' and the 'output gate.' Each gate has its own responsibility in information retention from h_{t-1} and x_t and regulates it with a sigmoid activation function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.1}$$

which takes values from 0 to 1, where x is a linear combination of $\boldsymbol{h_{t-1}}$ and $\boldsymbol{x_t}$.

The hidden state update involves multiple steps. First, an LSTM cell receives two inputs, the current input x_t and previous hidden state $\boldsymbol{h_{t-1}}$, and transforms them into candidate values to be added to the cell state, c_t^{in}, by

$$c_t^{in} = \tanh\left(W_{in,\ x}\,\boldsymbol{x_t} + W_{in,\ h}\,\boldsymbol{h_{t-1}} + b_{in}\right). \tag{4.2}$$

Simultaneously, the three gates, described above, monitor the flow of information into and out of the memory cell state:

1. The 'input gate' chooses the values of the network to be updated:

$$i_t = \sigma\left(W_{in,\ x}\,\boldsymbol{x_t} + W_{in,\ h}\,\boldsymbol{h_{t-1}} + b_i\right). \tag{4.3}$$

2. The 'forget gate' decides on which information from the past to keep and which to

discard:

$$f_t = \sigma\,(W_{f,\ x}\,\boldsymbol{x_t} + W_{f,\ h}\,\boldsymbol{h_{t-1}} + b_f\,). \tag{4.4}$$

3. The 'output gate' controls which information to let through to the hidden state update:

$$o_t = \sigma\,(W_{o,\ x}\,\boldsymbol{x_t} + W_{o,\ h}\,\boldsymbol{h_{t-1}} + b_o\,). \tag{4.5}$$

$W_{in,\ x}$, $W_{f,\ x}$, $W_{o,\ x}$, $W_{in,\ h}$, $W_{f,\ h}$, and $W_{o,\ h}$ are the weight matrices, responsible for updating the current input vectors, $\boldsymbol{x_t}$, and the previous hidden state of the cell, $\boldsymbol{h_{t-1}}$; and b_{in}, b_i , b_f, and b_o represent the bias terms for the corresponding update operation.

The memory cell state is updated based on the transformed input from (4.2) and the 'input gate' and 'forget gate' decisions from (4.3) and (4.4):

$$c_t = \sigma\,(f_t\,\boldsymbol{c_{t-1}} + i_t\,\boldsymbol{c_t^{in}}\,). \tag{4.6}$$

Finally, the hidden state is updated based on memory cell state from (4.6) and the 'output gate' decision (4.5):

$$h_t = o_t \tanh(\boldsymbol{c_t}). \tag{4.6}$$

4.2.5. Model Training

The prediction errors of our models were evaluated with binary cross-entropy loss. As we iterate though model training, the loss function calculates the amount of penalty the algorithm receives

for making a wrong prediction and is used to evaluate the algorithm's performance. For N training examples, the binary cross-entropy loss function, L, is defined as

$$L(y, \hat{y}) = -\sum_{i}^{N} y_i \log(\hat{y}_i) \tag{4.7}$$

where y_i and $\hat{y}_i$ are the true and predicted label for the case *i*.

We utilized stochastic gradient descent (SGD) as an optimizer and set the batch size to 64 for VGGNet fine-tuning and LSTM training. The hyperparameters of the LSTM network were optimized using a validation set. To avoid overfitting, early stopping was used to stop network training when validation loss started increasing.

The models were trained and evaluated on all slices of the 3D MRIs, which totaled ~12,000 slices. The dataset was separated into training + validation (80%) and testing (20%) sets by lesion, with cancer prevalence among the cases being constant across the sets. To avoid bias, image slices from the same lesion were retained in either the training subset or the testing subset, but not shared across both.

4.2.6. Performance Evaluation Metrics

Receiver operating characteristic (ROC) analysis was applied to evaluate binary classification performance of the models in the task of distinguishing benign and malignant lesions.[117] We measured their ability to discriminate the two classes using area under the ROC curve (AUC). The statistical difference in AUC values was evaluated using Delong tests.[92] Furthermore, specificity, positive predictive value (PPV), and negative predictive value (NPV)

were compared between the models for the same sensitivity threshold. Specificity measures the fraction of negative cases correctly identified. PPV and NPV measure the probabilities of a positive classification actually being positive and of a negative classification actually being negative.

4.2.7. Implementation Details

The ROI extraction was performed with the MATLAB software, developed specifically for the tasks. The deep learning-based methods were implemented in Python using the Keras library with Tensorflow backend.[87] Training and evaluation of the models were performed on an NVIDIA Titan X GPU.

4.3. Experiments and Results

Fine-tuning outperformed the method of using CNN as a feature extractor and therefore was used in the analysis. All of the lesions in the study had undergone biopsy. For the testing set, Table 4.2 summarizes the performance metrics, specificity, PPV, and NPV, for varying the decision thresholds for the two deep learning methods studied, i.e., the fine-tuned VGGNet and the LSTM. For a sensitivity of 0.92 and below, LSTM results in higher specificity and positive and negative predictive values as compared to the performance of VGGNet. These results demonstrate that the LSTM method achieves reduced number of false positives and calls a higher number of benign lesions as benign and malignant lesions as malignant. Above a sensitivity of 0.92, VGGNet shows slightly better specificity and predictive values.

Table 4.2. The performance metrics for fine-tuned VGGNet and LSTM network on the DCE-MRI test subset. For a given sensitivity value, we compare specificity, positive predictive value (PPV), and negative predictive value (NPV) for the two methods.

Sensitivity	**Specificity**		**PPV**		**NPV**	
	LSTM	***Fine-tuned VGGNet***	***LSTM***	***Fine-tuned VGGNet***	***LSTM***	***Fine-tuned VGGNet***
0.80	0.82	0.73	0.94	0.91	0.53	0.51
0.82	0.78	0.71	0.93	0.91	0.55	0.53
0.84	0.75	0.68	0.92	0.90	0.57	0.54
0.86	0.70	0.64	0.91	0.90	0.58	0.57
0.88	0.64	0.61	0.90	0.89	0.60	0.59
0.90	0.58	0.56	0.88	0.88	0.62	0.61
0.92	0.50	0.51	0.87	0.87	0.64	0.64
0.94	0.40	0.45	0.85	0.86	0.65	0.68
0.96	0.29	0.37	0.83	0.84	0.67	0.72
0.98	0.15	0.26	0.80	0.82	0.67	0.78

Even though both positive and negative predictive values are useful metrics in performance evaluation, class prevalence directly influences them. When holding all other variables constant and increasing just the class prevalence, PPV will increase and NPV will decrease. Our work was performed on an unbalanced dataset, with 68.6% malignant and 31.4% benign lesions. Given that, we conducted ROC analysis, which yields AUC values, a metric independent of class prevalence. Figure 4.6 shows the ROC curves for the lesion classification performance of the two models. The figure demonstrates that LSTM significantly outperformed the fine-tuned VGGNet, resulting in

AUC_{LSTM} = 0.88 (se = 0.01) and $AUC_{fine\text{-}tuned}$ = 0.84 (se = 0.01), with p = 0.00085, in the task of distinguishing benign and malignant lesions.

Note that the ROC curves cross, showing some ambiguity in the performances. Thus, we calculated the partial AUCs for the sensitivity range from 1 to 0.9 and for the specificity range of 1 to 0.9.[118] LSTM yielded improved partial AUCs of 0.064 and 0.037 as compared to those of the fine-tuned VGGNet, 0.041 and 0.025, for sensitivity and specificity ranges, respectively.

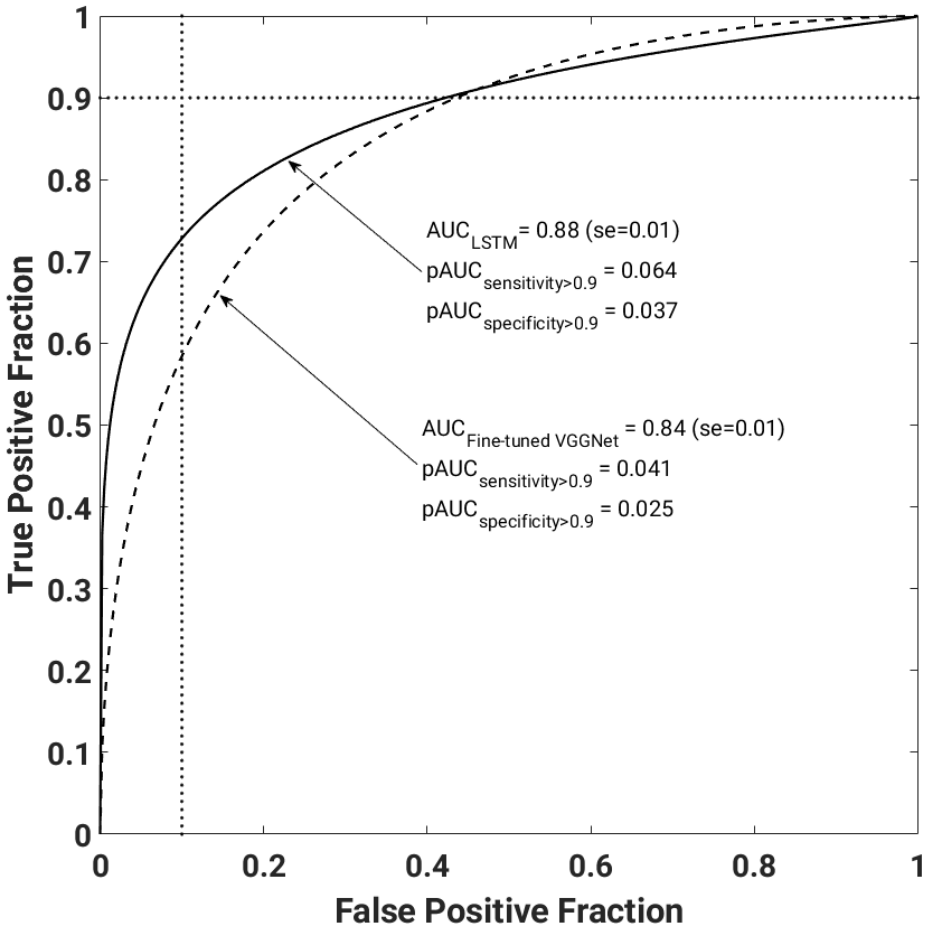


Figure 4.6. ROC curves corresponding to fine-tuned VGGNet and LSTM model performances in discriminating benign and malignant lesions. Solid line represents LSTM model and dashed line represents fine-tuned VGGNet. LSTM significantly outperformed the fine-tuned VGGNet.

4.4. Discussion and Conclusions

We present a breast lesion classification pipeline that captures simultaneously both morphological and temporal information about a lesion presented on DCE-MRIs in the task of distinguishing malignant and benign lesions. Compared to previous works, our method enables incorporation of the entire DCE-MRI sequence into deep learning-based lesion classification. For a given lesion, we extract a sequence of image feature vectors corresponding to the DCE-MRI sequence, which is further used to train an LSTM network. The image feature vectors are obtained from 5 max-pooling layers of pre-trained VGGNet in order to capture the local changes in lesion enhancement.

The LSTM method significantly outperformed the VGGNet, fine-tuned on RGB MRIs. In order to make sure that the results are not just due to higher-feature dimensionality, we also trained a simple two-layer, fully-connected neural network on multi-level image features extracted from RGB MRIs. This method incorporates the features extracted from various levels of VGGNet, but not the full DCE-MRI sequence. The classification performance of the network resulted in an AUC value of 0.82, which is lower than the performance of the fine-tuned VGGNet as well as the LSTM network.

It is important to note the differences in the datasets and the evaluation procedures of Chapter 3 and Chapter 4. Chapter 3 developed deep learning methods on a dataset of 690 cases. The dataset expanded to 703 cases by the time Chapter 3 research was performed. In Chapter 3 classifiers were evaluated with an AUC computed with five-fold cross-validation. This is a regular performance evaluation method that the Giger lab has been utilizing. Over the course of research, we began using a different approach of evaluating the developed methods, where the data is split

into training, validation, and testing subsets only once. This has been a standard evaluation method for deep learning algorithms in the research community and thus, it was adapted in our research. Given the differences in the datasets and evaluation methods, we cannot directly compare the results of the two chapters. However, both of them carry valuable results. Furthermore, Chapter 4 separately evaluated whether the fine-tuning approach is superior in performance to the CNN features extraction method developed in Chapter 3 (Figure 3.4).

LSTMs have achieved superior results for machine translation, language modeling, and image captioning tasks, outperforming other recurrent architectures. The main benefit of LSTMs is that they prevent vanishing or exploding gradients during error backpropagation with long input sequences. Therefore, the network can retain useful classification information from the beginning of the sequence. Furthermore, LSTMs are well suited for working with sequences of various lengths as well as time lags between the sequence elements, which is characteristic of our DCE-MRI data. DCE sequences contain one image taken prior to contrast injection and two to ten images taken after contrast injection. We also studied another RNN type, gated recurrent units (GRUs). GRUs have a similar but simpler architecture than LSTMs, resulting in fewer parameters and more efficient computation.[119] However, compared to LSTMs, GRUs do not control their hidden state with a memory unit. After investigating both architectures, we observed higher classification performance with the LSTM network.

The proposed method is inspired by the fact that human experts base various breast diagnostic and prognostic decisions on temporal changes in lesion enhancement observed in DCE-MRIs. Specifically, kinetic enhancement curve patterns are often visually analyzed during benign vs. malignant discrimination. Benign lesions tend to demonstrate moderate uptake and slow

washout of a contrast agent, while malignant lesions tend to have both rapid uptake and washout. Therefore, the sequences of image features extracted from the DCE-MRIs should be different for benign and malignant lesions. Among other clinical questions, radiologists' evaluation of breast cancer response to therapy is also guided by temporal changes of lesion enhancement. Lesion enhancement patterns and DCE-MRI quantitative pharmacokinetic parameters are used to assess breast cancer's response to primary and neoadjuvant chemotherapies.[32,120,121] These clinical questions are left for future work due to lack of availability of sufficient datasets.

Deep learning methods enable capturing of data patterns that have been previously unexploited, leading to more accurate, rapid, and accessible medical decision making. In this work, we demonstrate a deep learning method that captures clinically useful information presented in DCE-MRI sequence for breast lesion malignancy assessment.

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

This chapter presents the summary of the main contributions of this dissertation and provides thoughts on the future research directions. In this research we investigated robustness of conventional radiomic features across two MRI manufacturers for breast cancer lymph node and hormone receptor status assessments. Furthermore, we developed a deep learning-based methodology for breast lesion classification based on 4D DCE-MRI data for the task of lesion malignancy assessment and cancer treatment response prediction. The work also investigated whether the two types of radiomics, conventional and deep learning-based, are complementary to each other for the classification tasks and uncovered the limitations associated with each method individually.

Chapter 2 studied the robustness of the conventional radiomics features across two datasets acquired with MRI scanners of two different manufacturers, GE and Philips. All features, except the features from the size category, showed statistically significant differences in their average values across the two datasets. We failed to show significant differences in the average values of features from the size category. However, we avoid making conclusions on the robustness of the feature average values due to the limited size of the datasets on which the analysis was performed. During the analysis, we saw that the number of cases and class prevalence strongly affect the statistics. Furthermore, the work demonstrated that lesion size and enhancement texture features hold promise exhibiting equivalent prognostic performance in the task of distinguishing lymph node status across databases. In particular, the entropy feature, which quantifies randomness of

pixel values of the lesion image, was robust in its classification performance in multiple clinical tasks.

Chapter 3 developed and evaluated deep learning-based lesion classification methods for breast lesion malignancy assessment and prediction of cancer's response to neoadjuvant chemotherapy based on DCE-MRIs. The developed methodology is computationally efficient and does not require intensive image preprocessing. Moreover, we developed a system incorporating both deep learning and conventional CADx methods. We demonstrated that the fusion of the two types of radiomics is statistically significantly better than either one separately for the task of distinguishing benign and malignant lesions. Furthermore, our experiments demonstrated that the size of the selected ROI and the DCE time point on which the ROI was selected significantly affects the CNN-based performance. We also proposed a method to incorporate volumetric and partially temporal components of DCE-MRI using maximum intensity projection images for further lesion classification with pre-trained CNNs. DCE-MRI maximum intensity projection images incorporated clinically useful information about the entire lesion volume and partly about the contrast enhancement, which improved deep learning-based lesion classification compared to the classification based on a single MRI slice. MIP images are already commonly used in the evaluation of breast tumors and it would be natural to adopt MIP+CNN methods in clinical practice.

Chapter 4 solved one of the limitations of the methods proposed in Chapter 3. The deep learning model applied to MIP images did not fully capture the sequential dependencies present in a DCE image sequence. Therefore, Chapter 4 develops a deep learning methodology with a recurrent neural network that captures both morphological and physiological breast lesion

characteristics presented in DCE-MRI sequences. The method learns the image data patterns that have been previously unexploited, leading to more accurate, rapid, and accessible medical decision making.

Our studies have a few shortcomings, which can be improved in future research. Based on the research performed, a few suggestions are proposed:

1) Robustness of the radiomic features needs to be supported with analysis performed on larger and class-balanced datasets. It would also be valuable to perform statistical studies on datasets containing the same cases imaged on different MRI scanners.
2) For the deep learning-based research, many variations of deep learning models can be explored in the future. For example, exploring various CNN architectures and training them from scratch on the DCE-MRI dataset would be valuable and may improve breast lesion classification.
3) Our work utilized a transfer learning method between two very different image domains, where the CNN was originally trained on a natural image dataset and applied to a medical image dataset. In the future, a different form of transfer learning can be explored, where a CNN is trained on a medical dataset of another modality or disease and is then applied to the breast DCE-MRI dataset.
4) Further explorations are necessary to understand the utility of MIP images in the breast lesion classification with pre-trained neural networks. One of the directions is controlling for the slice thickness of MRIs. Our DCE-MRI dataset contains two thirds of the images with a slice thickness of 2 mm and one third with a slice thickness of 1.5 mm or 1.6 mm. Therefore, the maximum intensity projections are taken over variable

depth resolutions. In future work, larger databases need to be collected to perform studies on the effect of the slice thickness.

5) For the response to therapy analysis, larger datasets need to be acquired for more robust conclusions. Once larger datasets are acquired, the cases need to be separated based on the cancer subtypes and based on different treatment regimens are applied.
6) It would be valuable to evaluate the proposed LSTM method applied to MIP image sequences. Since LSTM models require large amounts of data for training, this can be evaluated once larger datasets are acquired.
7) Other models that are appropriate for DCE-MRI data and that should be explored in the future are 3D CNNs and CNN-LSTM models. The LSTM model proposed in this research performs well on classifying breast lesions; however, the method requires features extraction prior to training a classification model. CNN-LSTM architecture would not require the intermediate step of feature extraction and would process the image data starting from the 4D lesion ROIs and ending with the lesion classifications.
8) Since many of the proposed CNN architectures need more data and are hindered by the limited number of available medical cases, larger and more structured datasets need to be collected and data augmentation methods should be applied.

Computerized image analysis (radiomics) has the strong potential to lead clinicians towards more accurate and rapid image interpretation. Furthermore, it can serve as a "virtual digital biopsy," allowing for discovery of relationships between radiomics and pathology/genomics based on actual biopsies for ultimate use when biopsies are not practical, such as in screening and in repeated assessments during treatment monitoring. The thorough investigations of deep learning

methods and their combination with conventional radiomics methods showed the potential of the automated methods to improve breast cancer diagnosis and prognosis.

REFERENCES

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018: Cancer Statistics, 2018. *CA: A Cancer Journal for Clinicians* **68,** 7–30 (2018).
2. DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A. & Jemal, A. Breast cancer statistics, 2017, racial disparity in mortality by state: Breast Cancer Statistics, 2017. *CA: A Cancer Journal for Clinicians* **67,** 439–448 (2017).
3. Euhus, D., Di Carlo, P. A. & Khouri, N. F. Breast Cancer Screening. *Surg. Clin. North Am.* **95,** 991–1011 (2015).
4. Humphrey, L. L., Helfand, M., Chan, B. K. S. & Woolf, S. H. Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine* **137,** 347–360 (2002).
5. Tyne, K. & Nygren, P. Screening for Breast Cancer : Systematic Evidence Review Update for the U . S . Preventive Services Task Force. *Science* **151,** 727–W242. (2009).
6. Smith, R. A. *et al.* American Cancer Society Guidelines for Breast Cancer Screening: Update 2003. *CA: A Cancer Journal for Clinicians* **53,** 141–169 (2003).
7. Giger, M. L., Karssemeijer, N. & Schnabel, J. A. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annu Rev Biomed Eng* **15,** 327–357 (2013).
8. Giger, M. L., Chan, H.-P. & Boone, J. Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. *Med Phys* **35,** 5799–5820 (2008).

9. França, L. K. L. *et al.* Role of magnetic resonance imaging in the planning of breast cancer treatment strategies: comparison with conventional imaging techniques. *Radiologia Brasileira* **50,** 76–81 (2017).
10. The Canadian Task Force on Preventive Health Care. Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *Canadian Medical Association Journal* **183,** 1991–2001 (2011).
11. Steliarova-Foucher, E. *et al.* The European Cancer Observatory: A new data resource. *European Journal of Cancer* **51,** 1131–1143 (2015).
12. Saslow, D. *et al.* American Cancer Society Guidelines for Breast Screening with MRI as an Adjunct to Mammography. *CA: A Cancer Journal for Clinicians* **57,** 75–89 (2007).
13. Kuhl, C. K. *et al.* Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *J. Clin. Oncol.* **23,** 8469–8476 (2005).
14. Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer: a prospective multicentre cohort study (MARIBS). *The Lancet* **365,** 1769–1778 (2005).
15. Kriege, M. *et al.* Efficacy of MRI and Mammography for Breast-Cancer Screening in Women with a Familial or Genetic Predisposition. *New England Journal of Medicine* **351,** 427–437 (2004).
16. Lee, C. H. *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* **7,** 18–27 (2010).

17. Orel, S. G. & Schnall, M. D. MR Imaging of the Breast for the Detection, Diagnosis, and Staging of Breast Cancer. *Radiology* **220,** 13–30 (2001).
18. Sharma, U., Danishad, K. K. A., Seenu, V. & Jagannathan, N. R. Longitudinal study of the assessment by MRI and diffusion-weighted imaging of tumor response in patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. *NMR in Biomedicine* **22,** 104–113 (2009).
19. Guo, Y. *et al.* Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging. *Journal of Magnetic Resonance Imaging* **16,** 172–178 (2002).
20. *Encyclopedia of spectroscopy and spectrometry*. (Elsevier/AP, Academic Press is an imprint of Elsevier, 2017).
21. El Khouli, R. H. *et al.* Dynamic Contrast-Enhanced MRI of the Breast: Quantitative Method for Kinetic Curve Type Assessment. *American Journal of Roentgenology* **193,** W295–W300 (2009).
22. Moon, M., Cornfeld, D. & Weinreb, J. Dynamic Contrast-Enhanced Breast MR Imaging. *Magnetic Resonance Imaging Clinics of North America* **17,** 351–362 (2009).
23. Elsamaloty, H., Elzawawi, M. S., Mohammad, S. & Herial, N. Increasing Accuracy of Detection of Breast Cancer with 3-T MRI. *American Journal of Roentgenology* **192,** 1142–1148 (2009).
24. Turnbull, L. W. Dynamic contrast-enhanced MRI in the diagnosis and management of breast cancer. *NMR Biomed* **22,** 28–39 (2009).

25. Kuhl, C. K. *et al.* Abbreviated Breast Magnetic Resonance Imaging (MRI): First Postcontrast Subtracted Images and Maximum-Intensity Projection—A Novel Approach to Breast Cancer Screening With MRI. *Journal of Clinical Oncology* **32,** 2304–2310 (2014).

26. Jain, M., Jain, A., Hyzy, M. D. & Werth, G. FAST MRI breast screening revisited. *Journal of Medical Imaging and Radiation Oncology* **61,** 24–28 (2017).

27. Pineda, F. D. *et al.* Ultrafast Bilateral DCE-MRI of the Breast with Conventional Fourier Sampling. *Academic Radiology* **23,** 1137–1144 (2016).

28. Abe, H. *et al.* Kinetic Analysis of Benign and Malignant Breast Lesions With Ultrafast Dynamic Contrast-Enhanced MRI: Comparison With Standard Kinetic Assessment. *American Journal of Roentgenology* **207,** 1159–1166 (2016).

29. Esserman, L. *et al.* Utility of Magnetic Resonance Imaging in the Management of Breast Cancer: Evidence for Improved Preoperative Staging. *Journal of Clinical Oncology* **17,** 110–110 (1999).

30. Kuhl, C. K. Current Status of Breast MR Imaging Part 2. Clinical Applications. *Radiology* **244,** 672–691 (2007).

31. Hylton, N. M. *et al.* Locally Advanced Breast Cancer: MR Imaging for Prediction of Response to Neoadjuvant Chemotherapy—Results from ACRIN 6657/I-SPY TRIAL. *Radiology* **263,** 663–672 (2012).

32. Pickles, M. D., Lowry, M., Manton, D. J., Gibbs, P. & Turnbull, L. W. Role of dynamic contrast enhanced MRI in monitoring early response of locally advanced breast cancer to neoadjuvant chemotherapy. *Breast Cancer Research and Treatment* **91,** 1–10 (2005).

33. Valdora, F., Houssami, N., Rossi, F., Calabrese, M. & Tagliafico, A. S. Rapid review: radiomics and breast cancer. *Breast Cancer Research and Treatment* (2018). doi:10.1007/s10549-018-4675-4

34. Scrivener, M. *et al.* Radiomics applied to lung cancer: a review. *Translational Cancer Research* **5,** 398–409 (2016).

35. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Physics in Medicine and Biology* **61,** R150–R166 (2016).

36. Kotrotsou, A., Zinn, P. O. & Colen, R. R. Radiomics in Brain Tumors. *Magnetic Resonance Imaging Clinics of North America* **24,** 719–729 (2016).

37. Chen, W., Giger, M. L., Li, H., Bick, U. & Newstead, G. M. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magn Reson Med* **58,** 562–571 (2007).

38. Bhooshan, N. *et al.* Cancerous breast lesions on dynamic contrast-enhanced MR images: computerized characterization for image-based prognostic markers. *Radiology* **254,** 680–690 (2010).

39. Chen, W., Giger, M. L., Bick, U. & Newstead, G. M. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Med Phys* **33,** 2878–2887 (2006).

40. Chen, W., Giger, M. L. & Bick, U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiol* **13,** 63–72 (2006).

41. Chen, W., Giger, M. L., Bick, U. & Newstead, G. M. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI: Characteristic kinetic curves of breast lesions on DCE-MRI. *Medical Physics* **33,** 2878–2887 (2006).

42. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115,** 211–252 (2015).

43. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25* 1097–1105 (2012).

44. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *arXiv preprint arXiv:1702.05747* (2017).

45. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (The MIT Press, 2016).

46. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

47. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]* (2015).

48. Goodfellow, I. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160 [cs]* (2016).

49. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in 3431–3440 (IEEE, 2015). doi:10.1109/CVPR.2015.7298965

50. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (2015).

51. Sutskever, I., Vinyals, O. & Le V., Q. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* (2014).

52. Chen, X. & Zitnick, C. L. Mind's eye: A recurrent visual representation for image caption generation. in 2422–2431 (IEEE, 2015). doi:10.1109/CVPR.2015.7298856

53. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. 8

54. LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1,** 541–551 (1989).

55. Karpathy, A., Li, F.-F. & Johnson, J. CS231n: Convolutional Neural Networks for Visual Recognition.

56. Li, H. *et al.* MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays. *Radiology* **281,** 382–391 (2016).

57. Li, H. *et al.* Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *npj Breast Cancer* **2,** 16012 (2016).

58. Bhooshan, N. *et al.* Combined use of T2-weighted MRI and T1-weighted dynamic contrast-enhanced MRI in the automated analysis of breast lesions. *Magn Reson Med* **66,** 555–564 (2011).

59. Schnitt, S. J. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod. Pathol.* **23 Suppl 2,** S60-64 (2010).

60. *Devita, Hellman, and Rosenberg's cancer: principles & practice of oncology*. (Wolters Kluwer, 2015).

61. Mitri, Z., Constantine, T. & O'Regan, R. The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemotherapy Research and Practice* **2012,** 1–7 (2012).

62. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).

63. Zhu, Y., Qiu, P. & Ji, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods* **11,** 599–600 (2014).

64. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* **26,** 1045–1057 (2013).

65. Burnside, E. S. *et al.* Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage. *Cancer* (2015). doi:10.1002/cncr.29791

66. Guo, W. *et al.* Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. *J Med Imaging (Bellingham)* **2,** 041007 (2015).

67. Zhu, Y. *et al.* Deciphering Genomic Underpinnings of Quantitative MRI-based Radiomic Phenotypes of Invasive Breast Carcinoma. *Sci Rep* **5,** 17787 (2015).

68. *AJCC cancer staging manual*. (American Joint Committee on Cancer, Springer, 2017).

69. Gilhuijs, K. G., Giger, M. L. & Bick, U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Med Phys* **25,** 1647–1654 (1998).

70. Gibbs, P. & Turnbull, L. W. Textural analysis of contrast-enhanced MR images of the breast. *Magn Reson Med* **50,** 92–98 (2003).

71. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18,** 50–60 (1947).

72. Fay, M. P. & Proschan, M. A. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys* **4,** 1–39 (2010).

73. Walker, E. & Nowacki, A. S. Understanding equivalence and noninferiority testing. *J Gen Intern Med* **26,** 192–196 (2011).

74. Metz, C. & Pan, X. 'Proper' Binormal ROC Curves: Theory and Maximum-Likelihood Estimation. *J Math Psychol* **43,** 1–33 (1999).

75. Gruszauskas, N. P., Drukker, K., Giger, M. L., Sennett, C. A. & Pesce, L. L. Performance of breast ultrasound computer-aided diagnosis: dependence on image selection. *Acad Radiol* **15,** 1234–1245 (2008).

76. Mossman, D. Resampling techniques in the analysis of non-binormal ROC data. *Med Decis Making* **15,** 358–366 (1995).

77. McLachlan, G. J. *Discriminant analysis and statistical pattern recognition*. (2004).

78. Lo, S.-C. B. *et al.* Artificial convolution neural network for medical image pattern recognition. *Neural Networks* **8,** 1201–1214 (1995).

79. Zhang, W. *et al.* Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Medical Physics* **21,** 517–524 (1994).

80. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 3320–3328 (Curran Associates, Inc., 2014).

81. Donahue, J. *et al.* DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. 9

82. Huynh, B. Q., Li, H. & Giger, M. L. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging (Bellingham)* **3,** 034501 (2016).

83. Bar, Y., Diamant, I., Wolf, L. & Greenspan, H. Deep learning with non-medical training used for chest pathology identification. in (eds. Hadjiiski, L. M. & Tourassi, G. D.) 94140V (2015). doi:10.1117/12.2083124

84. Newitt, D. & Hylton, N. on behalf of the I-SPY 1 Network and ACRIN 6657 Trial Team. (2016). Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials. *The cancer imaging archive.* (2016).

85. Huynh, B. Q., Antropova, N. & Giger, M. L. Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning. in (eds. Armato, S. G. & Petrick, N. A.) 101340U (2017). doi:10.1117/12.2255316

86. Drukker, K. *et al.* Most-enhancing tumor volume by MRI radiomics predicts recurrence-free survival 'early on' in neoadjuvant treatment of breast cancer. *Cancer Imaging*

87. Szegedy, C., Liu, W., Jia, Y. & Sermanet, P. Going Deeper with Convolutions. in 1–9 (2014).

88. Zheng, L., Zhao, Y., Wang, S., Wang, J. & Tian, Q. Good Practice in CNN Feature Transfer. *arXiv:1604.00133 [cs]* (2016).

89. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).

90. Deza, M. & Deza, E. *Encyclopedia of distances*. (Springer Verlag, 2009).

91. Shawe-Taylor, J. & Sun, S. A review of optimization methodologies in support vector machines. *Neurocomputing* **74,** 3609–3618 (2011).

92. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44,** 837 (1988).

93. Hold, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6,** 65–70 (1979).

94. Pan, X. & Metz, C. E. The 'proper' binormal model: parametric receiver operating characteristic curve estimation with degenerate data. *Acad Radiol* **4,** 380–389 (1997).

95. Chollet, F. & others. Keras. *GitHub* (2016). Available at: https://github.com/fchollet/keras.

96. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825--2830 (2011).

97. Maximum Intensity Projection - MIPAV. Available at: https://mipav.cit.nih.gov/pubwiki/index.php/Maximum_Intensity_Projection. (Accessed: 18th September 2017)

98. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 65–70 (1979).

99. Rahbar, H., Partridge, S. C., DeMartini, W. B., Thursten, B. & Lehman, C. D. Clinical and technical considerations for high quality breast MRI at 3 tesla. *Journal of Magnetic Resonance Imaging* **37,** 778–790 (2013).

100. Zhang, Y. & Yeung, D.-Y. A Convex Formulation for Learning Task Relationships in Multi-Task Learning. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificail Intelligence (UA1)* 733–742 (2010).

101. Antropova, N. Multi-task Learning in the Computerized Diagnosis of Breast Cancer on DCE-MRIs.

102. Yang, Y. & Hospedales, T. M. DEEP MULTI-TASK REPRESENTATION LEARNING: A TENSOR FACTORISATION APPROACH. 12 (2017).

103. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv:1706.05098 [cs, stat]* (2017).

104. Greenspan, H., van Ginneken, B. & Summers, R. M. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging* **35,** 1153–1159 (2016).

105. Shin, H.-C. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging* **35,** 1285–1298 (2016).

106. Tajbakhsh, N. *et al.* Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging* **35,** 1299–1312 (2016).

107. Bar, Y., Diamant, I., Wolf, L. & Greenspan, H. Deep learning with non-medical training used for chest pathology identification. *SPIE Proceedings* **9414,** (2015).

108. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61,** 85–117 (2015).

109. Cheng, J.-Z. *et al.* Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports* **6,** (2016).

110. Antropova, N., Huynh, B. Q. & Giger, M. L. A Deep Feature Fusion Methodology for Breast Cancer Diagnosis Demonstrated on Three Imaging Modality Datasets. *Medical Physics* (2017). doi:10.1002/mp.12453

111. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017).

112. Goodfellow, I., Bengio, Y. & Courville, A. Sequence Modeling: Recurrent and Recursive Nets. in *Deep Learning* (MIT Press, 2016).

113. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780 (1997).

114. Ng, J. Y.-H., Yang, F. & Davis, L. S. Exploiting Local Features from Deep Networks for Image Retrieval. *arXiv:1504.05133 [cs]* (2015).

115. Lin, M., Chen, Q. & Yan, S. Network In Network. *arXiv:1312.4400 [cs]* (2013).

116. He, K., Zhang, X., Ren, S. & Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37,** 1904–1916 (2015).

117. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol* **56,** 45–50 (2008).

118. Jiang, Y., Metz, C. E. & Nishikawa, R. M. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **201,** 745–750 (1996).

119. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555 [cs]* (2014).

120. Martincich, L. *et al.* Monitoring Response to Primary Chemotherapy in Breast Cancer using Dynamic Contrast-enhanced Magnetic Resonance Imaging. *Breast Cancer Research and Treatment* **83,** 67–76 (2004).

121. Turkbey, B., Thomasson, D., Pang, Y., Bernardo, M. & Choyke, P. L. The role of dynamic contrast enhanced MR imaging in cancer diagnosis and treatment. *Diagnostic and Interventional Radiology* (2009). doi:10.4261/1305-3825.DIR.2537-08.1

LIST OF PUBLICATIONS AND PRESENTATIONS

Research Papers

N Antropova, M Giger, B Huynh, Hui Li, "Long short-term memory networks for efficient breast DCE-MRI classification." *JMI: Journal of Medical Imaging* (under review).

N Antropova, H Abe, M Giger, "Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep CNNs." *JMI: Journal of Medical Imaging* (2018).

K Drukker, N Antropova, H Li, M Giger, "Most-enhancing tumor volume by MRI radiomics predicts recurrence- free survival 'early on' in neoadjuvant treatment of breast cancer. *Cancer Imaging* (2018).

N Antropova, B Huynh, M Giger, "A deep fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. " *Medical Physics* (2017). – ***Editor's Choice***

N Antropova, B Huynh, M Giger, "Multi-task Learning in the computerized diagnosis of Breast Cancer on DCE-MRIs." *arXiv* preprint arXiv:1701.03882 (2017).

H Li, B Huynh, M Giger, N Antropova, "Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a large clinical dataset of FFDMs." *Journal of Med Imaging* (2017).

N Antropova, B Huynh, M Giger, "Performance comparison of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignant breast lesions on DCE-MRI." In *SPIE Medical Imaging Proceedings* (2017).

B Huynh, N Antropova, M Giger, "Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning." *SPIE Medical Imaging Proceedings* (2017).

H Li, B Huynh, M Giger, N Antropova, "Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms." *SPIE Medical Imaging Proceedings* (2017).

N Antropova, A Sanchez, I Reiser, E Sigky, J Boone, X Pan, "Efficient iterative image reconstruction algorithm for dedicated breast CT." In *SPIE Medical Imaging Proceedings* (2016).

Oral Presentations

N Antropova, H Abe**,** M Giger, "Maximum Intensity Projections for Incorporation of 4-Dimensional DCE-MR Images into Breast Lesion Classification with Deep CNN." *Radiological Society of North America* (2017).

N Antropova, B Huynh**,** M Giger, "Performance comparison of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignant breast lesions on DCE-MRIs." *SPIE Medical Imaging* (2017).

N Antropova, B Huynh, M Giger, "Predicting breast cancer malignancy using pre-trained convolutional neural networks on DCE-MRI data." *American Association of Physicists in Medicine* (2016).

H Li, B Huynh, M Giger, N Antropova, L Lan, "Use of deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a large clinical dataset of FFDMs." *Radiological Society of North America* (2016).

N Antropova, M Giger, H Li, K Drukker, L Lan, "Radiomics of breast cancer: A robustness study." *American Association of Physicists in Medicine* (2015).

Poster Presentations

H Whitney, N Antropova, M Giger, "Use of Deep Learning in the Classification of Benign Lesions, Luminal A Cancers, and Other Molecular Cancer Subtypes in Breast Magnetic Resonance Imaging." *American Association of Physicists in Medicine* (2018).

N Antropova, B Huynh, M Giger, "Recurrent neural networks for breast lesion classification based on DCE-MRIs." *SPIE Medical Imaging* (2018).

N Antropova, B Huynh, M Giger, "Long short-term memory networks for efficient breast DCE-MRI classification." *NIPS: Neural Information Processing Systems*, *Medical Imaging Meets NIPS* (2017).

N Antropova, B Huynh, M Giger, "Multi-task learning in the computerized diagnosis of breast cancer on DCE-MRIs." *NIPS: Neural Information Processing Systems*, *Machine Learning in Health Care* (2016).