THE UNIVERSITY OF CHICAGO

ARTIFICIAL INTELLIGENCE FOR MEDICAL IMAGE ANALYSIS FOR BREAST CANCER MULTIPARAMETRIC MRI AND COVID-19 CHEST RADIOGRAPHY

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
QIYUAN HU

CHICAGO, ILLINOIS
DECEMBER 2021

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES xv

ACKNOWLEDGMENTS xviii

ABSTRACT xx

1 INTRODUCTION 1
1.1 Artificial Intelligence in Medical Image Analysis 2
1.1.1 Human-Engineered Radiomics 3
1.1.2 Deep-Learning-Based Radiomics 3
1.2 MRI for Breast Cancer Imaging 6
1.2.1 Breast MRI in Clinical Practice 6
1.2.2 MRI Physics 10
1.2.3 The Multiparametric Breast MRI Protocol 11
1.3 Chest Radiography for COVID-19 Imaging 16
1.3.1 Chest Radiography and Computed Tomography 17
1.3.2 Use of Imaging in COVID-19 18
1.4 Machine Learning Algorithms Employed 21
1.4.1 Support Vector Machine 22
1.4.2 Convolutional Neural Networks 22
1.4.3 Multiple Instance Learning 32
1.4.4 Evaluation of Machine Learning Performance 34
1.4.5 Explainability and Interpretability 37
1.5 Research Objectives and Scope 37

2 MULTIPARAMETRIC MRI FOR BREAST CANCER DIAGNOSIS USING HUMAN-ENGINEERED RADIOMICS AND DEEP LEARNING 39
2.1 Introduction 39
2.2 Dataset 40
2.3 Methods 47
2.3.1 Human-Engineered CADx 47
2.3.2 Deep-Learning-Based CADx 50
2.3.3 Evaluation and Statistical Analysis 54
2.4 Further Investigations 55
2.4.1 ResNet Feature Extraction for mpMRI 55
2.4.2 Human-Engineered Radiomics versus Deep Learning for DWI 55
2.5 Results 57
2.5.1 Multiparametric MRI CADx 57
2.5.2 ResNet Feature Extraction for mpMRI 66
2.5.3 Human-Engineered Radiomics versus Deep Learning for DWI 67

2.6 Discussion and Conclusions 69

3 HIGH-DIMENSIONAL DEEP LEARNING IMAGE ANALYSIS FOR BREAST CANCER DIAGNOSIS ON MULTIPARAMETRIC MRI 74
3.1 Introduction 74
3.2 Datasets 75
3.2.1 UChicago Medicine DCE-MRI Dataset 75
3.2.2 Tianjin Medical University MRI Dataset 76
3.3 Methods 80
3.3.1 Volumetric and Temporal Information in DCE-MRI 80
3.3.2 Independent Validation of Feature MIP on DCE-MRI 84
3.3.3 Deep Learning for DWI 86
3.3.4 Feature MIP on Multiparametric MRI 87
3.4 Results 88
3.4.1 Volumetric and Temporal Information in DCE-MRI 88
3.4.2 Independent Validation of Feature MIP on DCE-MRI 91
3.4.3 Deep Learning for DWI 93
3.4.4 Feature MIP on Multiparametric MRI 93
3.5 Discussion and Conclusions 97

4 ARTIFICIAL INTELLIGENCE FOR COVID-19 DIAGNOSIS AND PROGNOSIS ON CHEST RADIOGRAPHY 101
4.1 Introduction 101
4.2 Datasets 103
4.3 Methods 105
4.3.1 Early Diagnosis 105
4.3.2 Prognosis 109
4.3.3 Evaluation and Statistical Analysis 109
4.4 Further Investigations: Classification and Visualization using Multiple Instance Learning 110
4.5 Results 111
4.5.1 Early Diagnosis 111
4.5.2 Prognosis 117
4.5.3 Classification and Visualization using Multiple Instance Learning . . 119
4.6 Discussion and Conclusions 120

5 SUMMARY AND FUTURE DIRECTIONS 124

REFERENCES 129

A HUMAN-ENGINEERED RADIOMIC FEATURES FOR MULTIPARAMETRIC MRI143

B COMPARATIVE RADIOMICS EVALUATION OF PAIRED CONVENTIONAL DCE-MRI AND ABBREVIATED MRI FOR BREAST CANCER DIAGNOSIS 151
B.1 Introduction 151
B.2 Methods 151
B.3 Results and Discussion 152

LIST OF PUBLICATIONS AND PRESENTATIONS 159
Peer-Reviewed Publications 159
Proceedings Papers and Extended Abstracts 160
Oral Presentations 161
Poster Presentations 162

LIST OF FIGURES

1.1 Schematic flowchart of a computerized tumor phenotyping system for breast cancers on DCE-MRI. The CAD radiomics pipeline includes computer segmentation of the tumor from the local parenchyma and computer-extraction of human-engineered radiomic features covering six phenotypic categories: (1) size (measuring tumor dimensions), (2) shape (quantifying the 3-D geometry), (3) morphology (characterizing tumor margin), (4) enhancement texture (describing the heterogeneity within the texture of the contrast uptake in the tumor on the first postcontrast MRIs), (5) kinetic curve assessment (describing the shape of the kinetic curve and assessing the physiologic process of the uptake and washout of the contrast agent in the tumor during the dynamic imaging series), and (6) enhancement-variance kinetics (characterizing the time course of the spatial variance of the enhancement within the tumor). CAD = computer-aided diagnosis; DCE-MRI = dynamic contrast-enhanced MRI. Reprinted from [1]. 4

1.2 Screening breast MRI detects malignancies occult on other imaging modalities. (A) craniocaudal and (B) mediolateral oblique full-field digital mammography images of the left breast demonstrate no suspicious findings. (C) Early postcontrast T1-weighted fat subtracted axial and (D) maximum intensity projection images from screening breast MRI demonstrate a 7 mm enhancing mass with spiculated margins in the left breast at 12 o'clock 10 cm from the nipple. Pathology from an MRI–guided percutaneous breast biopsy yielded invasive ductal carcinoma (grade 1). Reprinted from [2]. 8

1.3 Components of the basic multiparametric breast MRI protocol. In general, the protocol is begun with the non–contrast-enhanced acquisitions (T2-weighted and diffusion-weighted imaging [DWI]). This is followed by a native T1-weighted acquisition and subsequently the contrast-enhanced series (ultrafast [UF] imaging and regular T1-weighted imaging). For screening purposes, this protocol may be abbreviated to contain only the T1-weighted acquisitions before and directly after contrast material administration, with or without the acquisition of ultrafast images (or only ultrafast images if they are of sufficiently high resolution). For lesion discrimination, adding T2-weighted imaging and DWI is beneficial. The information from ultrafast images is in essence similar to (although somewhat more discriminative than) the delayed phase dynamics, and these can therefore both be used. After neoadjuvant chemotherapy, the delayed phase is essential to document the presence of residual ductal carcinoma in situ. Reprinted from [3]. 12

1.4 Semiquantitative breast DCE kinetics analysis approach, as defined in the ACR BI-RADS atlas [4]. The initial phase is classified based on the percent increase in signal intensity from precontrast levels, with increases of less than 50%, 50% to 100%, and greater than 100% classified as slow, medium, and fast, respectively. The delayed phase is classified by the curve type after initial peak enhancement as persistent (defined as a continuous increase in the enhancement of $> 10\%$ initial enhancement), plateau (constant signal intensity once the peak is reached $\pm 10\%$ initial enhancement), or washout (decreasing signal intensity after peak enhancement $> 10\%$ initial enhancement). Reprinted from [5]. 13
1.5 Example images obtained with DWI scan. Shown are corresponding slices from (A) S_0 with $b = 0$ s/mm^2, (B) S_D with $b = 800$ s/mm^2, (C) Apparent diffusion coefficient (ADC) map. An invasive tumor (arrow) exhibits reduced diffusivity on DWI, appearing hyperintense on S_D and hypointense on the ADC map. Reprinted from [6]. 15
1.6 (a) Standard unsubtracted image obtained in a 54-year-old woman with a history of smoking. (b) Soft-tissue-selective image more clearly demonstrates increased opacity (arrow) in the right infrahilar region. (c) Bone-selective image shows a malpositioned left central venous catheter. Arrow indicates the tip of the catheter, located in the neck. Note the artifacts along the aortic arch (white streak), the border of the left side of the heart (black streak), and the left hemidiaphragm and stomach bubble (parallel white and black streaks) due to misregistration during the subtraction process. (d) Axial CT scan shows a right infrahilar mass that proved to be lung cancer. Reprinted from [7]. 19
1.7 Example hyperplanes for discriminating the two classes (black and white circles). H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximal margin. 23
1.8 Dropout neural net model. Left: A standard neural net with two hidden layers. Right: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped. Reprinted from [8]. 25
1.9 Illustration of the network architecture of VGG-19 model, in the case of a 1000-class classification, as in the ImageNet challenge. 26
1.10 Illustration of a two-layer residual block, a building block of ResNet. 27
1.11 ResNet architectures for ImageNet classification. Building blocks are shown in brackets, with the numbers of blocks stacked. Downsampling is performed by conv3_1, conv4_1, and conv5_1 with a stride of 2. FLOPs = floating point operations per second. Reprinted from [9]. 28
1.12 (a) Illustration of a five-layer dense block. (b) Illustration of a deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling. Each layer takes all preceding feature maps as input. Adapted from [10]. 29

1.13 U-net architecture (example for 32 × 32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower-left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Reprinted from [11]. 31
1.14 Illustration of the inception block used to augment the U-Net. An inception block is also the building block of an inception network architecture. 32
1.15 Deep MIL approaches: (a) the instance-based approach, (b) the embedding-based approach, (c) the proposed approach with the attention mechanism as the MIL pooling. Red color corresponds to instance scores; blue color depicts a bag vector representation. Reprinted from [12]. 35

2.1 Distribution of (a) slice thickness and (b) in-plane resolution of the dynamic contrast-enhanced (DCE) sequences and T2-weighted (T2w) sequences in the multiparametric MRI database [13]. 41
2.2 Distribution of (a) MRI acquisition date, (b) magnet strength, and (c) lesion volume of the full dataset and the subset that contains diffusion-weighted imaging (DWI) sequence in the multiparametric MRI database. 44
2.3 Lesion classification pipeline based on diagnostic images [14]. Radiomic features were extracted from dynamic contrast-enhanced (DCE), T2-weighted (T2w), and diffusion-weighted MRI (DWI) sequences. The mpMRI information was incorporated in two different ways: feature fusion, i.e., merging radiomic features extracted from all sequences to train a support vector machine (SVM) classifier, and classifier fusion, i.e., aggregating the probability of malignancy output from all single-sequence classifiers via soft voting. Parentheses contain the numbers of features extracted from each sequence. The dashed lines for DWI indicate that the DWI sequence was only included in the classification process when it was available, while the DCE and T2w sequences were available for all lesions and thus were always included. ADC = apparent diffusion coefficient, ROC = receiver operating characteristic. 48
2.4 Lesion classification pipeline based on diagnostic images [13]. Information from dynamic contrast-enhanced (DCE) and T2-weighted (T2w) MRI sequences are incorporated in three ways: image fusion, i.e., fusing DCE and T2w images to create RGB composite image, feature fusion, i.e., merging convolutional neural network features extracted from DCE and T2w as the support vector machine (SVM) classifier input, and classifier fusion, i.e., aggregating the probability of malignancy output from the DCE and T2w classifiers via soft voting. MIP = maximum intensity projection. ROI = region of interest. ROC = receiver operating characteristic. 52

2.5 An example of the image fusion process [13]. A dynamic contrast-enhanced (DCE)-MRI transverse second post-contrast subtraction maximum intensity projection (MIP) and a T2-weighted (T2w)-MRI transverse center slice are shown with their corresponding regions of interest (ROIs) extracted. The RGB fusion ROI is created by inputting the DCE ROI into the red channel and the T2w ROI into the green channel. 54
2.6 Feature pooling from various levels of ResNet-50. 56
2.7 Diagonal classifier agreement plot between the T2-weighted (T2w) and dynamic contrast-enhanced (DCE) single-sequence classifiers trained on human-engineered radiomic features [14]. The x-axis and y-axis denote the probability of malignancy (PM) scores predicted by the classifiers using DCE and T2w features, respectively. Each point represents a lesion for which predictions were made. Points along or near the diagonal from bottom left to top right correspond to high classifier agreement; points far from the diagonal correspond to low agreement. Examples of lesions on which the two classifiers were in extreme agreement/disagreement are also included. Disagreement: lower right benign: papilloma; lower right malignant: IDC/DCIS, HER-2 enriched; upper left benign: fibroadenoma; upper left malignant: IDC/DCIS, luminal A. Agreement (both incorrect): upper right benign: hyalinized stromal fibrosis; lower left malignant: ductal carcinoma in situ. Agreement (both correct): upper right malignant: IDC/DCIS, triple negative, very large; lower left benign: fibroadenoma. 58
2.8 Bland-Altman plot illustrating classifier agreement between the single-sequence classifiers trained on human-engineered dynamic contrast-enhanced (DCE) features and T2-weighted (T2w) features [14]. The y-axis shows the difference between the support vector machine output scores of the two classifiers; the x-axis shows the mean of two classifiers' outputs. 59
2.9 Fitted binormal receiver operating characteristic (ROC) curves for single-sequence (dashed line) and multiparametric MRI (mpMRI) classifiers (solid line) based on human-engineered radiomic features, trained on the full set [14]. The three single-sequence classifiers were trained separately on (i) dynamic contrast-enhanced (DCE), (ii) T2-weighted (T2w), and (iii) diffusion-weighted imaging (DWI) features. The mpMRI models (iv) were trained on the ensemble of features extracted from all available sequences, and (v) aggregated the probabilities of malignancy from the single-sequence classifiers via soft voting. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier. 60

2.10 A diagonal classifier agreement plot between the T2-weighted (T2w) and dynamic contrast-enhanced (DCE) single-sequence classifiers trained on features extracted using VGG-19 [13]. The x-axis and y-axis denote the probability of malignancy (PM) scores predicted by the DCE classifier and the T2w classifier, respectively. Each point represents a lesion for which predictions were made. Points along or near the diagonal from bottom left to top right indicate high classifier agreement; points far from the diagonal indicate low agreement. Examples of lesions on which the two classifiers were in extreme agreement/disagreement are also included. Disagreement: lower right benign: fibroadenoma; lower right malignant: IDC/DCIS; upper left benign: unknown; upper left malignant: IDC/DCIS. Agreement (both incorrect): upper right benign: fat necrosis; lower left malignant: IDC. 62
2.11 Bland-Altman plot illustrating classifier agreement between the dynamic contrast-enhanced (DCE) maximum intensity projection and T2-weighted (T2w)-based single-sequence classifiers trained on features extracted using VGG-19 [13]. The y-axis shows the difference between the support vector machine output scores (predicted posterior probabilities of malignancy) of the two classifiers; the x-axis shows the mean of two classifiers' outputs, which is also the probability of malignancy scores calculated in the classifier fusion method. 63
2.12 Fitted binormal receiver operating characteristic (ROC) curves for two single-sequence and three mpMRI classifiers trained on features extracted using VGG-19 [13]. The classifiers used (i) convolutional neural network (CNN) features extracted from dynamic contrast-enhanced (DCE) subtraction maximum intensity projections (MIPs), (ii) CNN features extracted from T2-weighted (T2w) center slices, (iii) CNN features extracted from DCE and T2w fusion images, (iv) ensemble of features extracted from DCE and T2w images, and (v) probability of malignancy outputs from the DCE MIP and T2w classifiers aggregated via soft voting. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier scheme. T2w images were rescaled to match the in-plane resolution of their corresponding DCE sequences, but image registration was not performed. 64
2.13 Fitted binormal receiver operating characteristic (ROC) curves for two single-sequence and three mpMRI classifiers trained on features extracted using ResNet-50 [15]. The classifiers used (i) convolutional neural network (CNN) features extracted from dynamic contrast-enhanced (DCE) subtraction maximum intensity projections (MIPs), (ii) CNN features extracted from T2-weighted (T2w) center slices, (iii) CNN features extracted from DCE and T2w fusion images, (iv) ensemble of features extracted from DCE and T2w images, and (v) probability of malignancy outputs from the DCE MIP and T2w classifiers aggregated via soft voting. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier scheme. T2w images were rescaled to match the in-plane resolution of their corresponding DCE sequences, but image registration was not performed. 67

2.14 Fitted binormal receiver operating characteristic (ROC) curves for the ADC-based classifier, the CNN-based classifier, and the fusion classifier for DWI [16]. 69

3.1 Flowchart of study participants enrollment [17]. 77
3.2 Illustration of the processes to construct the second post-contrast subtraction images, the subtraction maximum intensity projection (MIP) images, and region of interest (ROI) [18]. 80
3.3 Illustration of the RGB region of interest (ROI) construction process. 81
3.4 Lesion classification pipelines based on diagnostic images. Three-dimensional volumetric lesion information from dynamic contrast-enhanced (DCE)-MRI is collapsed into 2D by maximum intensity projection (MIP) at the image level (left) or at the feature level (middle and right) along the axial dimension. Temporal information is incorporated via either subtraction images (left and middle) or inputting different time points in a DCE sequence into the RGB channels of the CNN input (right). 83
3.5 Lesion classification pipelines for image maximum intensity projection (MIP) and feature MIP [17]. The top portion illustrates the construction of the region of interest (ROI) that incorporates volumetric and temporal information from the four-dimensional dynamic contrast-enhanced MRI sequence. The same ROI was cropped from the first, second, and third post-contrast subtraction images and combined in the red, green, and blue (RGB) channels to form a three-dimensional (3D) RGB ROI. For image MIP (left branch of the bottom portion), the MIP RGB ROI was generated from the 3D RGB ROI, collapsing volumetric lesion information at the image level. For feature MIP (right branch of the bottom portion), volumetric lesion information was integrated at the feature level by max pooling feature extracted from all slices. SVM = support vector machine. . 85
3.6 Lesion classification pipelines for multiparametric MRI. Feature MIP was applied to the lesion volume in three MRI sequences, and features from these sequences were concatenated and input to a multilayer perceptron (MLP). 88
3.7 Architecture of the modified VGG-19 and the multilayer perceptron (MLP) of which the overall model was composed. 89
3.8 Fitted binormal receiver operating characteristic (ROC) curves for two classifiers that utilize the volumetric and temporal information from dynamic contrast-enhanced (DCE)-MRI. The legend gives the area under the ROC curve (AUC) with standard error (SE) for each classifier scheme. 90
3.9 Fitted binormal receiver operating characteristic (ROC) curves for two classifiers that use the four-dimensional volumetric and temporal information from dynamic contrast-enhanced MRI [17]. The legend gives the area under the ROC curve (AUC) with the 95% CI for each classifier. 92

3.10 A diagonal classifier agreement plot between the image maximum intensity projection (MIP) and feature MIP methods [17]. The x-axis and y-axis denote the probability of malignancy (PM) scores predicted by the image MIP classifier and feature MIP classifier, respectively. Each point represents a lesion for which predictions were made. Points along or near the diagonal from bottom left to top right indicate high classifier agreement; points far from the diagonal indicate low agreement. The insets are the MIP regions of interest (ROIs) and three-dimensional (3D) ROIs, which served as convolutional neural network (CNN) inputs for the image MIP and feature MIP methods, respectively, of extreme examples on which using feature MIP resulted in more accurate predictions than using image MIP (lesion 1-2), on which using image MIP resulted in more accurate predictions than using feature MIP (lesion 3), and on which the two methods both predicted accurately (lesion 4-5). Lesion 1: invasive micropapillary carcinoma; lesion 2: fibromatosis; lesion 3: invasive ductal carcinoma, grade II; lesion 4: invasive ductal carcinoma, grade II; lesion 5: non-mass enhancement, fibroadenoma. 94
3.11 Bland-Altman plot for the image maximum intensity projection (MIP) and feature MIP classifiers [17]. The x-axis and y-axis show the mean and difference between the support vector machine output scores (i.e., predicted posterior probabilities of malignancy [PMs]) of the two classifiers, respectively. 95
3.12 Fitted binormal receiver operating characteristic (ROC) curves for two diffusion-weighted imaging classifiers. Examples of the two classifiers' inputs are shown on the right. The legend gives the area under the ROC curve (AUC) with the 95% CI for each classifier. 96
3.13 Fitted binormal receiver operating characteristic (ROC) curves for single-sequence classifiers of three MRI sequences and a fusion mpMRI classifier, all of which utilize the high-dimensional information in the images. The legend gives the area under the ROC curve (AUC) with the 95% CI for each classifier. 97
3.14 Example input ROI (top row) and their Grad-CAM heatmap overlays (bottom row) of (a) a benign breast lesion and (b) a malignant breast lesion. The probability of malignancy (PM) predicted by each single-sequence classifiers is shown above its corresponding heatmap overlay, and the PM predicted by the mpMRI classifier is shown on the left. 98

4.1 Distribution of the patient visit status in which chest radiography exams were acquired among COVID-19 positive patients in the dataset. ICU = intensive care unit, ED = emergency department. 104
4.2 The sequential transfer learning curriculum for the diagnosis of COVID-19, and information on the dataset for each phase of training [19]. 106
4.3 Illustration of Phase 3 in the sequential training process, fine-tuning the model on the pandemic-era CXR dataset to distinguish between COVID-19 positive and negative patients. The model architectures shown are for illustration purposes and are not the precise or complete architectures of the modified U-Net and the DenseNet models. 108

4.4 Fitted proper binormal ROC curves for classification tasks in the first two phases of training [19]. The legend gives the AUC with 95% CI for each classification task. 111
4.5 Example standard chest radiographs (CXR) and their Grad-CAM heatmaps overlays of (a) a COVID-19 positive case and (b) a COVID-19 negative case. The model prediction scores ($P_{COVID\text{-}19}$) are noted. Both examples show influence on model predictions from irrelevant areas outside the lungs when the full images were used, which was reduced when the cropped images derived from automatic lung segmentation were used. 112
4.6 Fitted proper binormal ROC curves for the COVID-19 classification task for the held-out test set in the third phase when using cropped standard CXR and/or cropped soft-tissue CXR images. 113
4.7 Standard and soft-tissue chest radiographs (CXR) of four example cases (post cropping) and their Grad-CAM heatmap overlays. The model prediction scores ($P_{COVID\text{-}19}$) are noted. In all four cases, model predictions and/or heatmaps show differences when the two types of CXR images are used. 116
4.8 (a) Bland-Altman plot for the model predictions based on standard and soft-tissue CXR images. The patient visit status of COVID-19 positive patients are indicated by different colors. (b) ROC curves for COVID-19 classification using both standard and soft-tissue CXR combined by feature fusion, presented by patient visit status. ICU = intensive care unit, ED = emergency department. . 117
4.9 Fitted binormal ROC curves for classification tasks requiring intensive care or not within 24–96 hours from image acquisition. The legend gives the AUC with 95% CI for each task. 118
4.10 Two examples of portable chest radiography images overlaid with their Grad-CAM heatmaps for intensive care need prediction within 24, 48, 72, and 96 hours, respectively. The patient in the top example was admitted into ICU 4 hours after the image was acquired. The patient in the bottom example has not been hospitalized since the image was obtained. 119
4.11 Example heatmaps created by attention-based deep multiple instance learning algorithm in the task of COVID-19 classification on initial CXR exam. 120

B.1 Flowchart for the comparative radiomic analysis of paired DCE-MRI and abbreviated MRI. 153
B.2 Distribution of Dice coefficient comparing segmentation on DCE-MRI and abbreviated MRI. 154
B.3 Examples of segmentation results based on DCE-MRI (yellow) and abbreviated MRI (magenta). 155
B.4 Fitted binormal receiver operating characteristic (ROC) curves for the classification task breast lesions using DCE-MRI, abbreviated MRI, and the hybrid analysis. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier. 156
B.5 Bland-Altman plot illustrating classifier agreement between the classifiers trained on dynamic contrast-enhanced (DCE) features and abbreviated MRI features. PM = Probability of malignancy. 157

B.6 Difference in the two classifiers' PMs in various Dice coefficient ranges. 158

LIST OF TABLES

2.1 Clinical characteristics of the dataset [13]. The number of lesions is shown, along with the percentage of the total. Patient age is summarized on a patient basis, and lesion information (malignancy status and subtypes) is summarized on a lesion basis. 42

2.2 Clinical characteristics of the dataset [14]. Patient age is summarized on a patient basis, and lesion information (malignancy status and subtypes) is summarized on a lesion basis. The full set is a mixture of cases imaged using either two or three sequences, and the diffusion-weighted imaging sequence (DWI) subset contains cases imaged using three sequences. 45

2.3 Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) along with the 95% confidence interval (CI) of AUC for each classifier based on human-engineered radiomic features, trained on the full set [14]. Sensitivity and specificity presented are for the optimal operating point determined using a metric for cut-off value that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$. Because all lesions were referred for biopsy, the sensitivity and specificity of the data set were not calculated for clinical assessment. 59

2.4 Performance comparison for the five classification methods based on human-engineered radiomic features, when classifiers were trained on the full set [14]. The classifier names are shown in the first column (single-sequence) and first row (multiparametric). P-value and 95% confidence interval (CI) of the difference in the area under the receiver operating characteristic curves (AUCs) are presented for each multiparametric classifier compared with each single-sequence classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections. 61

2.5 Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic curve (AUC) along with the 95% confidence interval (CI) of AUC for each classifier trained on features extracted using VGG-19 [13]. Sensitivity, specificity, PPV, and NPV presented are for the optimal operating point determined using a metric for a cut-off value that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$. Because all lesions were referred for biopsy, the sensitivity and specificity of the data set were not calculated for clinical assessment. 65

2.6 Performance comparison for the five classification methods based on features extracted using VGG-19 [13]. The classifier names are shown in the first row (single-sequence) and first column (multiparametric). P-value and 95% confidence interval (CI) of the difference in area under the receiver operating characteristic curves (AUCs) are presented for each multiparametric classifier compared with each single-sequence classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections. 65

2.7 Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) along with the 95% confidence interval (CI) of AUC for each classifier trained on features extracted using ResNet-50 [15]. Sensitivity and specificity presented are for the optimal operating point determined using a metric for a cut-off value that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$. Because all lesions were referred for biopsy, the sensitivity and specificity of the data set were not calculated for clinical assessment. 68
2.8 Performance comparison for the five classification methods based on features extracted using ResNet-50 [15]. The classifier names are shown in the first row (single-sequence) and first column (multiparametric). *P*-value and 95% confidence interval (CI) of the difference in the area under the receiver operating characteristic curves (AUCs) are presented for each multiparametric classifier compared with each single-sequence classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections. 68

3.1 Clinicopathological characteristics of the lesions from patients in the Tianjin breast MRI dataset [17]. 78
3.2 Classification performances and comparisons between the three classification schemes using the DeLong test. The *p*-value and 95% confidence interval (CI) of the difference in the areas under the receiver operating characteristic curves (AUCs) were computed with respect to the classifier using maximum intensity projection (MIP) images. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections. 89
3.3 Performance metrics comparison between image maximum intensity projection (MIP) and feature MIP models [17]. The area under the receiver operating characteristic curve (AUC), along with the standard error and the 95% CI, as well as the sensitivity and specificity (in percentage and ratio of cases) for each method. The 95% CI and *p*-value for the difference (Δ) between the two methods are also presented for each metric. The AUCs were compared using the DeLong test, and the sensitivities and specificities were compared using the McNemar test. 91
3.4 Area under the receiver operating characteristic curve (AUC) for single-sequence classifiers of three MRI sequences and a fusion mpMRI classifier. *P*-value and 95% confidence interval (CI) of the difference in AUCs are presented for the comparison between each single-sequence classifier and the multiparametric classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections. 96

4.1 Dataset statistics by patient and total images. 103
4.2 Dataset statistics and the prevalence of cases for initial CXR exams. 104
4.3 Dataset statistics for patients who required intensive care within 24, 48, 72, and 96 hours after chest radiography exams. The numbers of patients and images (in parentheses) in each subset are listed. 105

4.4 AUC values for using full and cropped standard CXR, and the p-value and 95% CI of the difference in AUC values. Asterisks denote statistical significance. . . . 112
4.5 Comparisons of classification performances using the DeLong test when standard CXR images, soft-tissue images, or fusion of both were used. The p-value and CIs of the difference in AUCs are presented for each comparison. The significance levels (α) and the widths of the confidence intervals are adjusted based on Bonferroni-Holm corrections. Asterisks denote statistical significance after correcting for multiple comparisons. 114
4.6 Additional evaluation metrics for the COVID-19 classification task for the held-out test set in the third phase when using cropped standard CXR and/or cropped soft-tissue CXR images. The metrics are calculated at two sensitivity levels. The 95% CIs are shown in brackets. 114
4.7 COVID-19 classification performance by CXR exam type (portable or dual-energy subtraction [DES] exam). The 95% CIs are shown in brackets. 117

A.1 Radiomic features extracted from dynamic contrast-enhanced (DCE) sequence and their descriptions. 143
A.2 Radiomic features extracted from T2-weighted sequence and their descriptions. . 148
A.3 Radiomic features extracted from the apparent diffusion coefficient (ADC) map derived from diffusion-weighted imaging (DWI) sequence and their descriptions. 150

B.1 AUC values for classifiers using radiomic features from DCE-MRI, abbreviated MRI, and hybrid analysis, as well as the p-value and 95% CI of the difference in AUC values. Asterisks denote statistical significance after multiple comparison corrections. 155

ACKNOWLEDGMENTS

This PhD dissertation and my journey through graduate school were made possible by the support and guidance from many people. My advisor, Dr. Maryellen Giger, played a significant role in my development as a scientist in all aspects. Her constant mentorship not only helped me grow into a better researcher, communicator, and collaborator, but also instilled in me the value of curiosity, persistence, and stepping out of my comfort zone.

The members of my thesis committee immensely contributed to the rigor of this dissertation. Drs. Hiroyuki Abe and Deepa Sheth generously shared their clinical expertise in breast radiology. Drs. Samuel Armato and Patrick La Riviere contributed their expertise in imaging physics and image analysis. I appreciate all the discussions with the committee, which helped form research questions, shape methodologies, and ensure the quality of this work.

The members and collaborators of the Giger lab also contributed to this dissertation and my experience as a graduate student. Drs. Karen Drukker, Hui Li, and Heather Whitney were wonderful mentors to me from day one and provided me with valuable guidance throughout my graduate study. Dr. Feng Li's clinical and research expertise was crucial for this work. Dr. Madeleine Durkee's encouragement and advice were a constant source of support. Sasha Edwards, John Papaioannou, and Dr. Nick Gruszauskas, along with others at the Human Imaging Research Office, contributed tremendously to the databases involved in my research. Chun-Wai Chan and Li Lan were always available to provide technical support.

All members of the Graduate Program in Medical Physics supported me in a variety of ways during my time here. The faculty, administrative staff, and my fellow students fostered a collaborative environment and a close-knit community, and I was lucky to be a part of it. I want to especially thank the program Chair Dr. Samuel Armato for his leadership, the professors who taught my courses, the senior students who offered helpful guidance, and my classmates who accompanied me along this journey. Moreover, I am grateful for the

opportunity to mentor several wonderful students and appreciate their contribution to my research.

This dissertation would not have been possible without the support of the following grants and fellowships: the National Institutes of Health (NIH) grant U01 CA195564, the National Institute of Biomedical Imaging and Bioengineering COVID-19 Contract 75N92020D00021, the C3.ai Digital Transformation Institute grant award, the NIH grant S10 OD025081, the Radiological Society of North America/American Association of Physicists in Medicine (AAPM) Graduate Fellowship, the AAPM Expanding Horizons Travel Grant, the Conference on Neural Information Processing Systems Machine Learning for Health Workshop Travel Grant Award, the Graduate Council Travel Fund Award, and the Biological Sciences Division Travel Award.

Last, and certainly not least, I dedicate this dissertation to my parents and grandparents, who always support and believe in me unconditionally and are my role models. I am fortunate to have parents who made sure that I received the best education and gave me the strength to forge my own path. I also want to thank my boyfriend, Sawyer, for his unwavering support through the ups and downs. Finally, my late grandmother who was taken by cancer years ago has been and will always be my inspiration to advance the capacity of healthcare with science and technology.

ABSTRACT

This dissertation studies AI-assisted medical image analysis in two applications: 1) breast cancer diagnosis on multiparametric magnetic resonance imaging (mpMRI) and 2) COVID-19 diagnosis and prognosis on chest radiography (CXR). Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death among women worldwide. MRI has become indispensable for breast imaging clinical practice and has evolved to mpMRI that includes multiple sequences, including a T1-weighted dynamic contrast-enhanced (DCE) sequence, a T2-weighted (T2w) sequence, and a diffusion-weighted imaging (DWI) sequence, to improve the specificity of breast MRI while preserving its sensitivity. Computer-aided diagnosis (CADx) systems based on human-engineered radiomics and deep learning have been developed to help improve diagnostic performance and reduce reading time. While previous CADx research has been primarily focused on the DCE sequence, the first aim of this dissertation investigates CADx methods, based on both human-engineered radiomic features and deep learning, that integrate complementary information provided by the various sequences in mpMRI in the task of distinguishing benign and malignant breast lesions, with the goal of leveraging the advancements in MRI technology to improve the performance of differential breast lesion classification compared with using DCE sequence alone. Three mpMRI fusion approaches are investigated: image fusion, i.e., fusing images from multiple MRI sequences into an RGB image as the input; feature fusion, i.e., concatenating features extracted from mpMRI sequences prior to classification; and classifier fusion, i.e., aggregating the probability of malignancy output scores from single-sequence classifiers via soft voting.

Although deep learning methods have demonstrated success in computer-aided medical imaging analysis, high dimensionality and data scarcity are unique challenges in medical imaging applications of deep learning. Transfer learning techniques with pretraining on two-dimensional images are often employed to circumvent the need for massive datasets, which have led to an underutilization of the high-dimensional, clinically valuable information

in MRI. In order to utilize the rich clinical information in breast MRI without sacrificing computational efficiency or classification performance, the second aim of this dissertation proposes and evaluates two-dimensional CNN transfer learning methods that incorporate the high-dimensional breast MRI in distinguishing benign and malignant breast lesions. In particular, these methods utilize the three-dimensional volumes in all MRI sequences, the temporal dimension in DCE-MRI, and the diffusion weighting information in DWI, which are then applied along with the multiparametric fusion methods to improve the classification performance of breast lesion differential diagnosis.

As COVID-19 emerged as a novel disease and developed into a pandemic, AI-assisted medical image analysis also holds promise to help optimize patient management and alleviate strains put on the healthcare system. The third aim of this dissertation is dedicated to investigating computer-aided methods that can potentially assist in the early diagnosis and accurate prognosis of COVID-19 using CXR images. Using a large CXR database curated during the COVID-19 pandemic for this study, a sequential transfer learning strategy follows a learning curriculum designed to pretrain and fine-tune a model on increasingly specific and complex tasks, and finally 1) distinguishes COVID-19 positive and negative patients using their initial CXR exam within two days of their initial RT-PCR test for COVID-19 and 2) predicts if a COVID-19 positive patient will potentially need intensive care in the next one to four days. Automatic lung segmentation and cropping are incorporated in the classification pipeline to reduce the influence of irrelevant regions of the images on model predictions. The role of soft tissue CXR images is studied in addition to the standard CXR images. A weakly supervised learning technique, attention-based deep multiple instance learning, is also investigated for classifying and localizing COVID-19 involvement on CXR images.

This dissertation presents the following results. First, when human-engineered features are used, feature fusion and classifier fusion methods achieve significantly higher classification performance using any MRI sequence alone. When CNN features are used, the feature

fusion method significantly outperforms using the DCE sequence alone, and all fusion methods significantly outperform using the T2w sequence alone. Overall, the findings suggest that leveraging the complementary information provided by various mpMRI sequences in CADx can improve the diagnostic performance in the task of distinguishing between benign and malignant breast lesions. Furthermore, the feature maximum intensity projection method, which globally max pools the features extracted from a lesion volume along the lesion's axial dimension within a CNN, demonstrates the ability to effectively utilize volumetric information in MRI exams when using two-dimensional CNNs with transfer learning to differentiate benign and malignant breast lesions. The RGB channels of CNNs pretrained on natural images effectively incorporate the dynamic time points in DCE and the diffusion weighting strengths in DWI. High classification performance is achieved when high-dimensional images in three mpMRI sequences are utilized. In addition, promising performance is achieved using both standard and soft tissue CXR images combined via feature fusion for diagnosing COVID-19 on the initial CXR at patient presentation. Multiple instance learning fails to improve the detection of COVID-19 on CXR in this specific task but shows promise for potential use in related tasks. The method is also able to predict if COVID-19-positive patients would require intensive care within 24, 48, 72, and 96 hours after CXR acquisition.

The medical significance of this work is that it can potentially improve the current breast cancer CADx systems and enhance the diagnostic workup by providing accurate image assessment for the patients and alleviating the workload of interpreting mpMRI for the clinicians. This work is also clinically significant in light of the ongoing COVID-19 pandemic, as the findings can potentially assist in computer-aided COVID-19 early diagnosis and prognosis, contributing to optimizing patient care and reducing the burden on healthcare systems. Methods developed in this work can be potentially applied to other clinical tasks that can benefit from AI-assisted medical image analysis.

Keywords: breast cancer, multiparametric MRI, COVID-19, chest radiography, artificial

intelligence, computer-aided diagnosis, radiomics, deep learning, convolutional neural network.

CHAPTER 1

INTRODUCTION

Due to continuing technological advances in medical image acquisition, novel imaging modalities are being introduced in medical practices, such as volumetric and multi-energy computed tomography (CT), multi-parametric and dynamic magnetic resonance imaging (MRI), multi-dimensional ultrasound, multi-planar interventional imaging, and multi-modal positron emission tomography (PET)/CT and PET/MRI hybrid imaging technologies [20]. While these imaging technologies provide radiologists with more information than ever before for their clinical assessment, the analysis of large amounts of imaging data has led to new challenges. There is an increasing need for image interpretation expertise, and interpretation remains time-consuming, prone to human error, and sometimes unavailable. The desire to improve the efficacy and efficiency of clinical care continues to drive innovations, including artificial intelligence (AI) and its applications in medical imaging. AI offers the opportunity to optimize and streamline the clinical workflow and aid in many of the clinical decision-making tasks that involve image interpretations. AI's capacity to recognize complex patterns in images, even those that are not noticeable or detectable by human experts, transforms image interpretation into a more quantitative and objective process. AI also excels at processing the sheer amount of information in multimodal data, giving it the potential to integrate not only multiple radiographic imaging modalities, but also other forms of data such as genomics, pathology, and electronic health records to perform comprehensive analyses and predictions.

Fortunately, the adoption of digital picture archiving and communication systems (PACS) in radiology and their integration within the overall hospital information system have allowed large databases of medical images and associated relevant medical information (e.g., demographics, clinical findings, blood tests, pathology, genomics, proteomics) to be built up. Such databases have become increasingly accessible for research purposes, which, along with the recent advancements in machine learning and computing power, provides the driving force

for the ongoing digital transformation that continues to significantly impact radiology and medicine in general, offering new opportunities for data-driven medical research and development, such as detection and characterization of anomalies, the discovery of early biomarkers of disease onset and progression, optimal therapy selection and prediction of therapy outcome, and correlation of genotype and phenotype related findings.

The introduction of this dissertation starts with an overview of AI methods that assist in medical image analysis and then introduces two specific application domains explored in this work, namely, breast cancer diagnosis on multiparametric MRI (mpMRI) and COVID-19 diagnosis and prognosis on chest radiography (CXR). For each application, the clinical imaging modalities for the disease, along with the computerized methods that have been developed and deployed, are reviewed. The introduction further presents and technical background of the machine learning algorithms used in this research. Finally, it concludes with a statement of the scope of this work and the outline for this dissertation.

1.1 Artificial Intelligence in Medical Image Analysis

AI-assisted medical image analysis, also termed radiomics, extracts a large number of features from radiographic medical images using data-characterization algorithms. Computer-aided detection (CADe), diagnosis (CADx), and triaging (CADt) systems have been under development and deployment for clinical use since the mid-1980s to aid radiologists in making better clinical decisions [21]. The first CAD methods were developed for the analysis of chest radiographs and mammograms. Since then, successful automated image analysis was performed on various imaging modalities for various diseases, such as breast, lung, colon, and prostate cancers, osteoporosis, cerebrovascular disease, diabetic retinopathy, interstitial disease, and many more [1, 22–24]. CAD systems extract and analyze large volumes of quantitative information from image data, assisting radiologists in image interpretation as a concurrent, secondary, or autonomous reader at various steps of the clinical workflow.

They aim to effectively exploit the available imaging data and other relevant information, reduce human errors, intra- and inter-observer variability, and evaluation times, as well as democratize high-quality medical image assessment in resource-limited settings and enable personalized medicine.

1.1.1 Human-Engineered Radiomics

CAD systems can be categorized into two types, which we refer to as human-engineered and deep-learning-based radiomics. The former has existed since the start of CAD development. Human-engineered features, also known as hand-crafted features, are defined by mathematical expressions that quantify visually discernible characteristics, such as size, shape, texture, and morphology, collectively describing the phenotypes of the imaged lesion or tissue. These features can be automatically extracted from images using computer algorithms with analytical expressions encoded. Then, machine learning models, such as linear discriminant analysis, support vector machines, or multilayer perceptron, can then be trained on the extracted features for various clinical questions. The extraction of human-engineered features often involves a prior segmentation of the lesion from the parenchyma background. Note that the extraction and interpretation of features depend on the imaging modality and the clinical task required. For example, Fig. 1.1 presents a CADx pipeline that automatically segments breast lesions and extracts six categories of human-engineered radiomic features from dynamic contrast-enhanced (DCE)-MRI on a workstation [1, 25–30].

1.1.2 Deep-Learning-Based Radiomics

The machine learning field has been going through a period of explosive development in recent years, in which the innovation and application of more powerful solutions are driven by the increased accessibility of computing power and large datasets. A crucial part of this development is deep learning, a type of machine learning that enables end-to-end learning

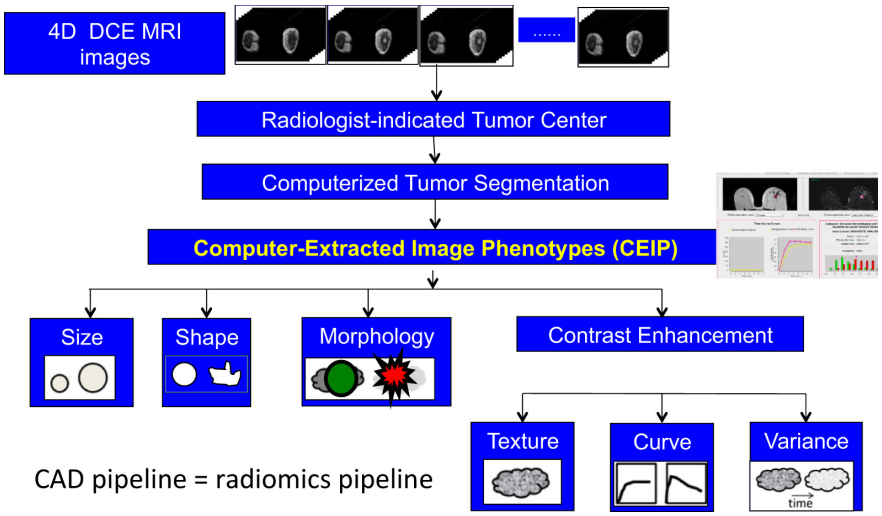


Figure 1.1: Schematic flowchart of a computerized tumor phenotyping system for breast cancers on DCE-MRI. The CAD radiomics pipeline includes computer segmentation of the tumor from the local parenchyma and computer-extraction of human-engineered radiomic features covering six phenotypic categories: (1) size (measuring tumor dimensions), (2) shape (quantifying the 3-D geometry), (3) morphology (characterizing tumor margin), (4) enhancement texture (describing the heterogeneity within the texture of the contrast uptake in the tumor on the first postcontrast MRIs), (5) kinetic curve assessment (describing the shape of the kinetic curve and assessing the physiologic process of the uptake and washout of the contrast agent in the tumor during the dynamic imaging series), and (6) enhancement-variance kinetics (characterizing the time course of the spatial variance of the enhancement within the tumor). CAD = computer-aided diagnosis; DCE-MRI = dynamic contrast-enhanced MRI. Reprinted from [1].

of very complex functions from raw data. Some of the greatest successes of deep learning have been in the field of computer vision, which considerably accelerated AI applications of medical imaging. Numerous types of deep learning algorithms, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), autoencoders, generative adversarial networks (GANs), and reinforcement learning, have been developed for medical imaging applications [23, 24, 31–33].

Deep learning methods have several advantages over the conventional CADx. Since they automatically learn useful features from the data for a given task during the training process, they eliminate the need for manual feature design and may be able to learn abstract representations that would not be described by human-engineered features. Medical images,

nevertheless, pose a set of unique challenges to deep-learning-based computer vision methods. For one, training of high-performing and robust deep learning models requires large amounts of well-annotated data, whereas medical imaging datasets are usually relatively small in size and can have incomplete or noisy labels. Further, the high-dimensionality and large size of medical images allow them to contain a wealth of clinically useful information, but the information cannot be optimally exploited by naive applications of the deep learning models developed for computer vision tasks for natural images. The lack of interpretability is another hurdle in building trustworthy deep-learning-based AI systems for healthcare purposes and adopting them for clinical use.

Due to data scarcity in the medical imaging domain, transfer learning is a commonly used technique when deep learning algorithms are employed, where the deep learning model is initialized with weights pretrained on millions of natural images (e.g., ImageNet, which contains over a million natural images in 1000 categories) or another related task [34–36]. The model initialized with pretrained weights can then either be used as a fixed feature extractor or be fine-tuned. For the former, the model's weights are frozen and applied directly to extract representations from medical images, and the extracted features can then be used to train a simpler machine learning model depending on the task. For fine-tuning, the weights for all or part of the model are updated during training based on the new data and the task of interest. Since earlier layers of a model are usually responsible for low-level features that are common to many types of images, such as shapes, gradients, and edges, earlier layers are sometimes frozen during fine-tuning. The choice and implementation of transfer learning techniques often have an impact on the model performance and run-time, and the decision needs to be made based on the specific scenario and/or through experimentation.

Human-engineered radiomics and deep learning methods for breast imaging analysis have both advantages and disadvantages regarding computation efficiency, amount of data required, preprocessing, interpretability, and prediction accuracy [37, 38]. They should be

chosen based on the specific tasks and scenarios, and they can be potentially combined to complement each other.

1.2 MRI for Breast Cancer Imaging

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death among women worldwide [39]. In the United States, breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer death among women, with over 281,000 estimated new cases and 43,000 estimated deaths in 2021 [40]. Due to its high prevalence, the advancement of clinical practice and basic research to predict the risk, detect and diagnose the disease, and implement the optimal therapy has a high potential impact. The progress against breast cancer so far is reflected in a substantial decrease in mortality. As of 2018, the female breast cancer death rate had dropped from its peak by 41% [40].

1.2.1 Breast MRI in Clinical Practice

Over the course of many decades, medical imaging modalities have been developed and used in routine clinical practice for these efforts in several capacities. First, screening techniques, including mammography and physical examination, are employed to detect abnormalities. Second, the abnormality is diagnosed through biopsy and further imaging such as ultrasound and MRI. Third, if found to be malignant, the lesion will be characterized by subtype and staged based on tumor size and extent of invasion. Finally, imaging such as MRI is used throughout treatment planning and monitoring.

Breast Cancer Screening

Mammography is the recommended method for breast cancer screening of women in the general population [41]. Screening with mammography is associated with a 20% – 40%

reduction in breast cancer deaths [2]. However, screening with mammography alone may be insufficient for women at high risk of breast cancer since its effectiveness is limited by its two-dimensional (2D) projection nature [41]. This concern is particularly important for women with dense breasts, as cancers can be missed at mammography in these women due to the camouflaging effect [42]. The need for more effective assessment strategies has led to the emergence of newer imaging techniques for supplemental screening, including digital breast tomosynthesis (DBT), MRI, and automated breast ultrasound [2, 43]. Breast MRI has been shown to detect additional cancers in women with negative screening mammography examinations regardless of their risk level for breast cancer [44]. Nevertheless, since the cost-effectiveness of screening MRI rises with increasing breast cancer risk, the American Cancer Society recommends screening breast MRI in women at high risk and some of the women at intermediate risk for breast cancer based on family history or genetic predisposition to supplement mammography or DBT, which has shown to identify earlier stage disease and is associated with improved survival rates [41, 43, 45]. Figure 1.2 shows an example where MRI detects a malignant lesion occult on mammography [2]. Ultrasound is an option for those high-risk women who cannot undergo MRI [41, 43]. There is insufficient evidence to support the use of other imaging modalities, such as thermography, breast-specific gamma imaging, PET, and optical imaging, for breast cancer screening [41].

High costs and limited availability of MRI units are the main factors that preclude the widespread use of screening MRI. The abbreviated MRI protocol, consisting of one pre- and one postcontrast T1-weighted acquisition, has shorter image acquisition and interpretation times [46]. Research has found similar diagnostic accuracy between abbreviated breast MRI and the full MRI protocol [47]. Also, ultrafast sequences may be used to obtain dynamic information without lengthening the protocol, maintaining a high diagnostic accuracy [48]. These emerging imaging technologies may reduce the costs of screening breast MRI and increase its availability.

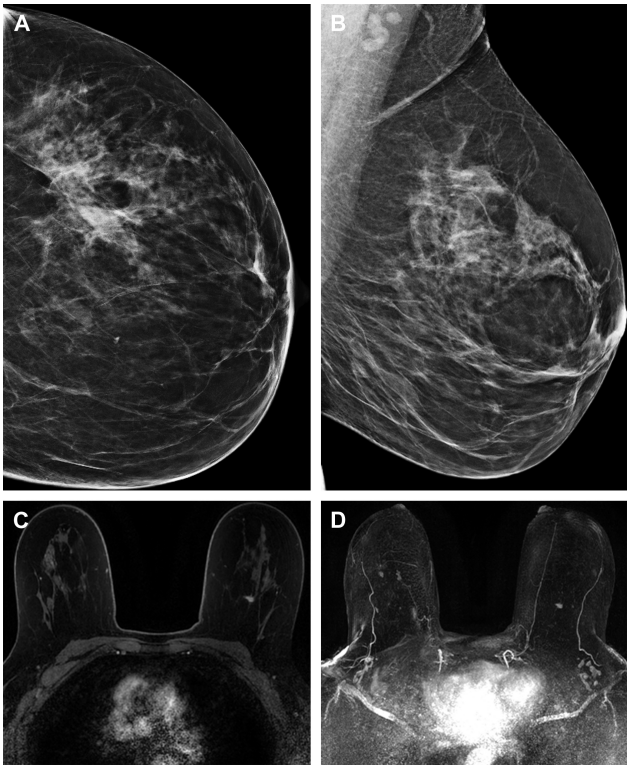


Figure 1.2: Screening breast MRI detects malignancies occult on other imaging modalities. (A) craniocaudal and (B) mediolateral oblique full-field digital mammography images of the left breast demonstrate no suspicious findings. (C) Early postcontrast T1-weighted fat subtracted axial and (D) maximum intensity projection images from screening breast MRI demonstrate a 7 mm enhancing mass with spiculated margins in the left breast at 12 o'clock 10 cm from the nipple. Pathology from an MRI–guided percutaneous breast biopsy yielded invasive ductal carcinoma (grade 1). Reprinted from [2].

Staging in Women with Known Breast Cancer

Although guidelines differ widely in their recommendations for the performance of preoperative breast MRI in women with a new diagnosis of breast cancer, in general, MRI is useful for determining the extent of the tumor, evaluation of the tumor's relation to the deep fascia, and screening of the contralateral breast. First, breast MRI provides high-quality preoperative staging. Most studies conclude that size estimations with MRI are more reliable than those with clinical examination, mammography, or ultrasound, and the benefit is particularly strong for invasive lobular carcinomas and DCIS components [49–53]. Moreover, breast MRI helps improve the management of detected lesions. With preoperative MRI, the detec-

tion of mammographic occult lesions in the affected breast is common, and the detection of additional disease that impacts treatment occurs in 20% of patients [54]. Besides incidental lesions, the use of a multiparametric MRI protocol is also valuable because it may allow the classification of lesions as certainly benign and obviate biopsy. Indeterminate lesions should be sampled with MRI-directed US-guided biopsy or MRI-guided biopsy [3]. In addition, although the benefit of using breast MRI findings in surgery is still under investigation, evidence clearly points to a reduction in the rate of re-excisions without increasing the rate of mastectomies for lobular cancers [55, 56]. Breast MRI also helps detect occult contralateral disease in 5.5%–9.3% of women with known unilateral breast cancer, with 37%–48% of these findings being malignant [54].

Evaluation of Women Treated with Neoadjuvant Chemotherapy

Monitoring the change in tumors during neoadjuvant chemotherapy (NAC) is important for the evaluation of treatment and preparation of downstream procedures. To evaluate residual tumor size, physical examination, mammography, ultrasound, and MRI have been used, among which MRI is the most accurate method, as it is difficult for other modalities to distinguish post-treatment fibrosis or post-biopsy change from residual tumor following NAC [57, 58]. Findings of breast cancers following NAC vary depending on tumor subtype, histologic type, and time points of MRI acquisitions, and thus a refined strategy for accurate interpretation is crucial. The purpose of the MRI examination should also be taken into account in the choice of MRI acquisition and interpretation strategies. From an oncologist's perspective, assessing response to a specific regimen and measuring changes in invasive tumor size is important, whereas residual DCIS might not be the primary concern. For a surgeon, to achieve a negative margin during BCS, tumor extent, including DCIS, should be measured [3].

1.2.2 MRI Physics

MRI is one of the wonders of modern science and medicine. MRI uses superconducting magnets with strong magnetic fields, typically 1.5 Tesla (T) or 3T, radiofrequency (RF) pulses, and magnetic gradients to generate images of internal structures, organs, and tissues of the patient. Hydrogen atoms are used to produce MR images because it is the most abundant nucleus in a human body, in water molecules and many other molecules, and thus it is able to produce the strongest signal. When placed in a strong magnetic field, the proton spins, i.e., their intrinsic magnetization that are randomly oriented in the absence of external magnetic fields, all align parallel to it. The RF pulse is applied perpendicular to the magnetic field, and when tuned to the Larmor frequency, i.e., the proton's precessional frequency, creates a phase coherence in the precession of the proton spins and tilts the magnetization away from the equilibrium alignment with the longitudinal magnetic field, so that a component of the magnetization lies in the transverse plane. The time-varying magnetic field from the rotating transverse magnetization will induce a tiny signal in the receiver coils, which are connected to sensitive amplifiers.

Spatial encoding of the MRI signal is accomplished through the use of magnetic field gradients, which are spatially variant small magnetic fields in addition to the main field, so that spins from protons in different locations precess at slightly different frequencies. The range of precessional frequencies in the signal detected by the receiver coil encode the locations in the body from which the signal originated.

The MRI signal is attenuated due to two simultaneous relaxation processes. The loss of coherence of the spin system attenuates the signal with a time constant called the transverse relaxation time, T2. Concurrently, the magnetization vector slowly relaxes towards its equilibrium orientation that is parallel to the external magnetic field, which occurs with a time constant called the spin-lattice relaxation time, T1. The contrast in MR images originates from the fact that different tissues usually have different T1 and T2 relaxation times, since

they are influenced by magnetic interactions of the nuclei with their local environment in the body. This is especially true for soft tissues, which explains the excellent soft-tissue contrast of MRI. The contrast in MR images can be changed by varying the weighting of the T1 and T2 relaxation times by changing the pulse repetition time (TR) and echo time (TE), respectively, in the image acquisition sequence.

1.2.3 The Multiparametric Breast MRI Protocol

While the basis for breast MRI is a T1-weighted dynamic contrast-enhanced (DCE) sequence, breast MRI has evolved from a primarily contrast-enhanced technique to a multiparametric technique, in which T2-weighted and diffusion-weighted imaging (DWI) are routinely performed. The various components of the multiparametric protocol are shown in Fig. 1.3 [3]. The T2-weighted and DWI sequences are acquired before contrast material administration since they do not rely on the contrast agent. Despite its high sensitivity for breast cancer detection and characterization, DCE-MRI has relatively moderate and variable specificity, which may lead to unnecessary secondary patient management and anxiety. T2-weighted and DWI sequences provide additional morphological and functional information that complement the DCE sequence, and the use of multiparametric MRI has shown to improve the specificity while preserving its sensitivity for the differentiation of benign and malignant lesions [59].

The T1-weighted DCE sequence involves intravenous injection of a gadolinium-based contrast agent. A native T1-weighted acquisition is obtained prior to contrast material administration, and multiple T1-weighted images are acquired afterward with appropriate time intervals in between. Contrast material is administered at a maximum dose of 0.1 mmol per kilogram of body weight, preferentially at a flow rate of 2 ml/sec, and flushed with saline [60]. Gadolinium administered in small doses affects the microenvironment by reducing the T1 relaxation time, therefore, increasing its signal in a T1-weighted MRI acquisition [61].

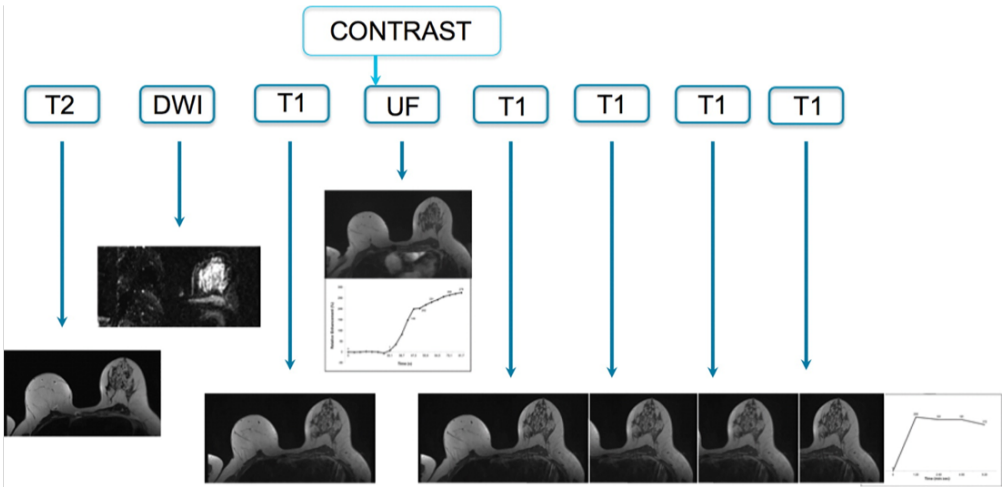


Figure 1.3: Components of the basic multiparametric breast MRI protocol. In general, the protocol is begun with the non–contrast-enhanced acquisitions (T2-weighted and diffusion-weighted imaging [DWI]). This is followed by a native T1-weighted acquisition and subsequently the contrast-enhanced series (ultrafast [UF] imaging and regular T1-weighted imaging). For screening purposes, this protocol may be abbreviated to contain only the T1-weighted acquisitions before and directly after contrast material administration, with or without the acquisition of ultrafast images (or only ultrafast images if they are of sufficiently high resolution). For lesion discrimination, adding T2-weighted imaging and DWI is beneficial. The information from ultrafast images is in essence similar to (although somewhat more discriminative than) the delayed phase dynamics, and these can therefore both be used. After neoadjuvant chemotherapy, the delayed phase is essential to document the presence of residual ductal carcinoma in situ. Reprinted from [3].

DCE-MRI allows for visualization of spatial and temporal variations of abnormalities, providing morphological details and functional information. By using modern MRI units and breast coils, high spatial resolution, 1 mm isotropic or lower, is obtainable without lengthening the acquisition time. 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, which enables easier visual and computerized discrimination of the lesion and surrounding tissue. Furthermore, the time-signal intensity curve, or kinetic curve, obtained at the multiple time points, carries highly useful information for clinical evaluation. As shown in Fig. 1.4, the kinetic curves allow for assessment of the initial phase, within approximately 2 minutes of contrast injection, and the late (or delayed) phase, after 2 minutes or after peak

enhancement [5]. In the initial phase, enhancement classifications of slow, medium, and fast are determined by signal intensity increase. In the delayed phase, enhancement curves can be classified by three basic curve types: persistent, plateau, and washout. According to a classification scheme based on the shape of the contrast-time intensity curves, breast masses with persistent contrast enhancement (Type I) are likely to be benign, whereas plateau delayed enhancement (Type II) is of intermediate suspicion for malignancy, and washout delayed enhancement (Type III) is the most indicative of malignancy. Although the most classic curve type for malignant breast lesions demonstrates rapid uptake followed by early washout, there is a significant overlap of kinetic curve types among benign and malignant lesions [5].

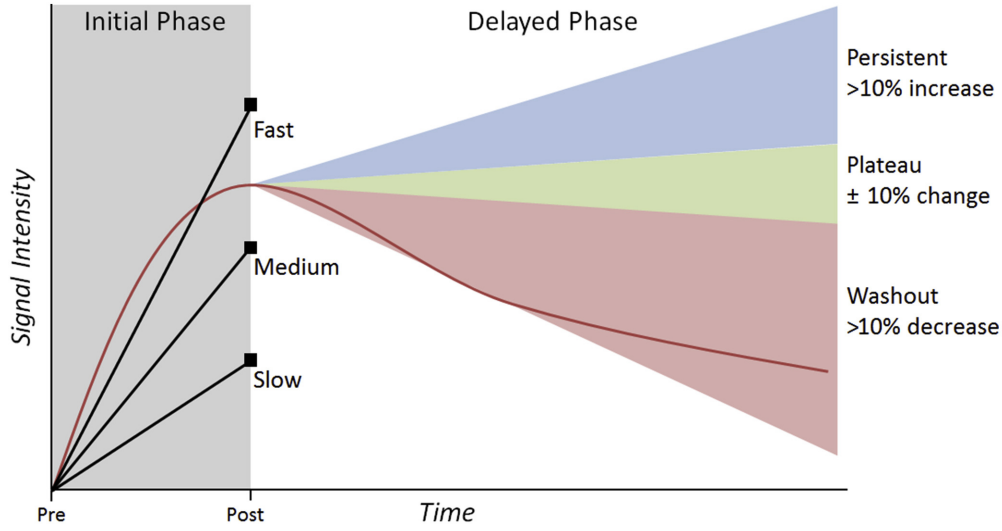


Figure 1.4: Semiquantitative breast DCE kinetics analysis approach, as defined in the ACR BI-RADS atlas [4]. The initial phase is classified based on the percent increase in signal intensity from precontrast levels, with increases of less than 50%, 50% to 100%, and greater than 100% classified as slow, medium, and fast, respectively. The delayed phase is classified by the curve type after initial peak enhancement as persistent (defined as a continuous increase in the enhancement of $> 10\%$ initial enhancement), plateau (constant signal intensity once the peak is reached $\pm 10\%$ initial enhancement), or washout (decreasing signal intensity after peak enhancement $> 10\%$ initial enhancement). Reprinted from [5].

Subtracting the precontrast image from postcontract images, thus forming subtraction images, is helpful for acquisitions with fat suppression because they help differentiate truly

enhancing structures from lesions with native high signal intensity at T1 [62]. Moreover, generating the maximum intensity projection (MIP) from these subtracted images aids in rapid lesion detection, but motion artifacts, chemical shift artifacts, and poor fat suppression may obscure lesions on MIP images.

T2-weighted imaging is included in the multiparametric MRI protocol to depict edema, hemorrhage, mucus, or cystic fluid, providing additional information that complements the T1-weighted DCE sequence [63, 64]. T2-weighted imaging with fat suppression enables easy visualization of cysts. T2-weighted imaging without fat suppression allows better depiction of lesion morphology. Most masses with high signal intensity at T2-weighted imaging are benign, such as apocrine metaplasia, cyst, myxoid fibroadenoma, fat necrosis, and lymph nodes [64]. Most cancers do not show high signal intensity relative to parenchyma at T2-weighted imaging because of their high cellularity and low water content. However, several rare types of breast cancer, including mucinous, medullary, papillary, and metaplastic carcinomas, can have high signal intensity on T2-weighted images [64]. Several studies have reported that T2-weighted imaging increases the specificity for differentiation of benign and malignant lesions [65]. For example, fibroadenomas, a type of benign lesion that can exhibit similar contrast agent enhancement to that of malignant lesions on T1-weighted DCE-MRI, are usually hyperintense on T2-weighted images, while malignant lesions are usually iso- or hypointense [63].

DWI quantifies the random movement of water molecules in tissue, which is influenced by tissue microstructure and cell density. This is achieved by applying motion-sensitizing gradients (b factors) to a primarily T2-weighted echo-planar imaging sequence [6, 66]. The DWI signal intensity decreases proportionally to the water diffusivity as follows:

$$S_D(b) = S_0\, e^{-b \times ADC}, \tag{1.1}$$

where $S_D(b)$ is the signal intensity with diffusion weighting b, S_0 is the signal intensity with-

out diffusion weighting. The apparent diffusion coefficient (ADC) is a quantitative measure of diffusivity derived from DWI, defined as the average area a water molecule occupies per unit time. Breast cancer has significantly lower ADCs than benign breast lesions or normal tissue (0.8–1.3 $\times$ 10^{-3} mm^2/sec versus 1.2–2.0 $\times$ 10^{-3} mm^2/sec), which is due to the relatively increased tumor cellularity that restricts diffusion, manifested by the bright signal on DWI and dark signal on a corresponding ADC map [6, 67]. An example is shown in Fig. 1.5 [6].

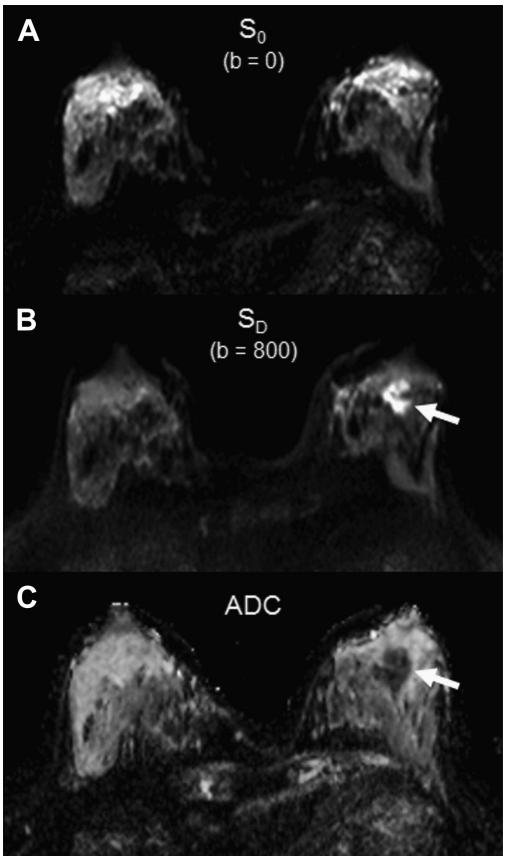


Figure 1.5: Example images obtained with DWI scan. Shown are corresponding slices from (A) S_0 with $b = 0$ s/mm^2, (B) S_D with $b = 800$ s/mm^2, (C) Apparent diffusion coefficient (ADC) map. An invasive tumor (arrow) exhibits reduced diffusivity on DWI, appearing hyperintense on S_D and hypointense on the ADC map. Reprinted from [6].

1.3 Chest Radiography for COVID-19 Imaging

The coronavirus disease 2019 (COVID-19) pandemic has emerged as an unprecedented health care crisis and has profoundly impacted global public health and the economy. The pathogen responsible for COVID-19 is severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is structurally related to the virus that causes severe acute respiratory syndrome (SARS). Since the outbreak in late 2019, scientific and clinical evidence is evolving on both acute and long-term effects of COVID-19, which can affect multiple organ systems. The principal mode of transmission of SARS-CoV-2 is through exposure to respiratory fluids carrying the infectious virus. Infected patients release respiratory fluids during exhalation (e.g., quiet breathing, speaking, singing, exercise, coughing, sneezing) in the form of droplets that carry the virus and transmit infection. Exposure occurs mainly in three ways: (1) inhalation of very fine respiratory droplets and aerosol particles, (2) deposition of respiratory droplets and particles on exposed mucous membranes in the mouth, nose, or eye by direct splashes and sprays, and (3) touching mucous membranes with hands that have been soiled either directly by virus-containing respiratory fluids or indirectly by touching surfaces with the virus on them [68].

The SARS-CoV-2 virus is highly contagious, and infection can cause severe and sometimes fatal diseases. Therefore, early detection and appropriate patient management are crucial when navigating the pandemic, both for the patient's well-being and for public health purposes. Early detection not only allows for prompt treatment at the earlier, more manageable stage of the disease, but also informs patient isolation based on disease mitigation and containment strategies. Accurate prognosis enables planning and optimization of medical resource allocation, as well as choosing the appropriate intervention and implementing necessary adjustments.

Laboratory confirmation of SARS-CoV-2 is performed with a virus-specific reverse transcription polymerase chain reaction (RT-PCR) test. Early on in the pandemic, early detec-

tion and containment of infection was hindered by the need to develop, mass produce, and widely disseminate the RT-PCR assay. While there have been successful efforts to increase the RT-PCR testing capacity, shortages of test kits and long processing times remain a problem in resource-limited settings during surges. Moreover, the RT-PCR test has moderate and variable sensitivity in clinical practice [69]. The value of an imaging test relates to the generation of results that are clinically actionable either for establishing a diagnosis or for guiding management, triage, or treatment. The value is diminished by associated costs, including the risk of radiation exposure to the patient, risk of COVID-19 transmission to uninfected healthcare workers and other patients, consumption of personal protective equipment (PPE), and need for cleaning and downtime of radiology rooms in resource-constrained environments. The appropriate use of imaging in each of these scenarios needs to be considered on the basis of the cost-benefit trade-off.

1.3.1 Chest Radiography and Computed Tomography

CXR and CT of the thorax are the primary imaging modalities that have been recommended as potential triage and diagnostic tools for COVID-19. CXR is the oldest and most commonly performed medical imaging examination. It involves exposing a part of the body to a low dose of ionizing radiation (average is around 0.02 mSv for a front view and 0.08 mSv for a lateral view) to produce 2D projection images of the heart, lungs, airways, blood vessels, and the bones of the spine and chest [70]. A standard CXR examination consists of an erect posteroanterior (PA) radiograph and a left lateral projection acquired during full inspiration with the patient facing the detector. A CXR exam can also be performed using a portable x-ray machine, acquiring a radiograph anteroposteriorly (AP) with the patient facing the x-ray beam. Portable radiography is recommended for patients too unstable or unable to travel to a radiology department; however, if a standard chest radiography exam is possible, it is preferred due to the superior diagnostic quality and acquisition of multiple projections

[71]. The type of image acquisition affects the image quality as well as the appearance of pathologic findings and thoracic structures due to photon beam divergence [72].

Dual-energy subtraction (DES) CXR is a robust and powerful tool used in many modern CXR exams to improve the ability to detect and accurately diagnose a wide variety of thoracic abnormalities on PA-lateral chest images [7, 73]. DES CXR takes advantage of the higher differential attenuation of bones as a function of photon energy compared to soft tissue, allowing for the ability to decompose two images taken at different x-ray energies into tissue-selective representations of the anatomy, namely soft-tissue and bone images. This helps in cases where the bony structure of the ribs and clavicle obscures the subtle soft-tissue abnormalities in the lung because of anatomical overlap caused by the projection process. An example is shown in Fig. 1.6 [7]. In portable CXR exams, postprocessing techniques can be applied to generate a synthetic soft tissue image, such as in ClearRead Xray Bone Suppress (Riverain Technologies) series. The algorithm can increase the visibility of soft tissue in the standard CXR image by suppressing the bone on the digital image without the need for two exposures and has also demonstrated its usefulness in helping radiologists identify missed nodules [74].

CT is also an x-ray-based imaging modality. CT scanners use a rotating x-ray tube and a row of detectors placed in the gantry to measure x-ray attenuation by different tissues inside the body. Cross-sectional images of the body are reconstructed from measurements of attenuation coefficients of x-ray beams in the volume of the object studied.

1.3.2 Use of Imaging in COVID-19

CXR is recommended for triaging at patient presentation and disease monitoring due to its fast speed, relatively low cost, wide availability, and portability [75, 76]. Characteristics such as bilateral lower lobe consolidations, ground glass densities, peripheral air space opacities, and diffuse air space disease on CXR have been related to COVID-19 [77, 78]. CXR is

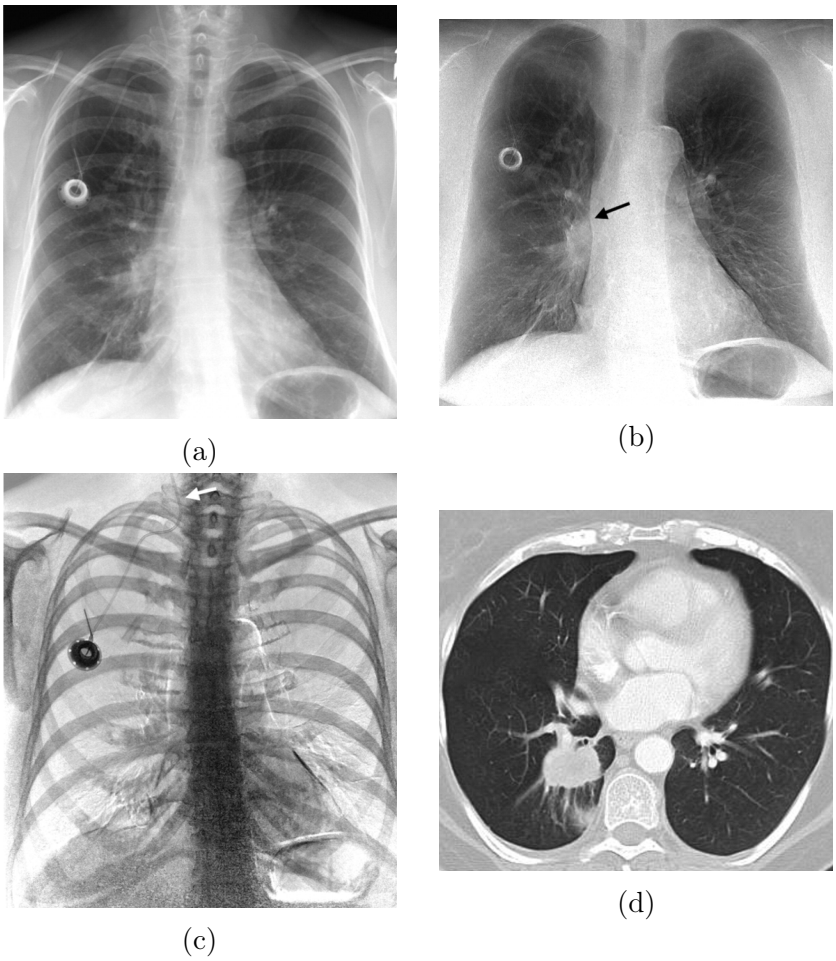


Figure 1.6: (a) Standard unsubtracted image obtained in a 54-year-old woman with a history of smoking. (b) Soft-tissue-selective image more clearly demonstrates increased opacity (arrow) in the right infrahilar region. (c) Bone-selective image shows a malpositioned left central venous catheter. Arrow indicates the tip of the catheter, located in the neck. Note the artifacts along the aortic arch (white streak), the border of the left side of the heart (black streak), and the left hemidiaphragm and stomach bubble (parallel white and black streaks) due to misregistration during the subtraction process. (d) Axial CT scan shows a right infrahilar mass that proved to be lung cancer. Reprinted from [7].

insensitive in mild or early COVID-19 infection [79]. However, the relative value of CXR or CT for detecting the presence of viral pneumonia depends on community norms and public health directives. When patients are encouraged to present early in the course of their disease, as was the case during the outbreak in Wuhan, China, CXR has little value. The greater sensitivity of CT for early pneumonic changes is more relevant in the setting of a public health approach that required isolation of all infected patients within an environment where the reliability of COVID-19 testing was limited and turnaround times were long [80]. Alternatively, during the surge in New York City, where patients were instructed to stay at home until they experienced advanced symptoms, CXR exams were often abnormal at the time of presentation. Furthermore, equipment portability with imaging performed within an infected patient's isolation room is another factor that may favor CXR in selected populations, effectively eliminating the risk of COVID-19 transmission along the transport route to a CT scanner and within the room housing a CT scanner, particularly in environments lacking PPE. In addition, CXR can be useful in hospitalized patients for assessing disease progression and alternative diagnoses [75].

CT is more sensitive for early parenchymal lung disease, disease progression, and alternative diagnoses, including acute heart failure from COVID-19 myocardial injury and, when performed with intravenous contrast material, pulmonary thromboembolism [81]. Leveraging these superior capabilities depends on the availability of CT capacity, particularly considering the potential reduction in CT scanner availability due to the additional time required to clean and disinfect equipment after imaging of patients suspected of having COVID-19. Although local practice patterns and resource availability do not articulate the relative merit of chest radiography versus CT, the choice of imaging modality is ultimately left to the judgment of clinical teams at the point of care, accounting for the differing attributes of CXR and CT, local resources, and expertise.

1.4 Machine Learning Algorithms Employed

As mentioned at the beginning of this section, the machine learning field has been advancing for decades and has seen explosive growth in recent years. The wave of development in machine learning algorithms has spun advancements in numerous application areas, including medical imaging. This subsection will provide an overview of the machine learning algorithms employed in this dissertation.

There are several types of machine learning algorithms in terms of the level of supervision. For example, supervised learning requires all training data to be labeled with truth, and the algorithm learns from the labeled training data to find a mapping that transforms the input data into a predefined output and predicts the correct label for the test data. In unsupervised learning, the algorithm is given unlabeled input data and tries to understand the internal structures of the data on its own. The most common example for unsupervised learning is clustering, where samples are automatically divided into groups based on their most distinct features. In recent developments of deep learning, the line between supervised and unsupervised has started to blur, yielding a few other types of algorithms.

Semi-supervised learning is an approach that uses partially labeled data sets. For example, the algorithm can first use labeled data to train itself, resulting in a partially trained model, which then predicts labels of the unlabeled data [82]. The combined labeled and "pseudo-labeled" data then train the algorithm in a supervised manner. This technique is useful when fully labeled training sets are infeasible, whereas the acquisition of unlabeled data is relatively inexpensive. Weakly supervised learning trains the algorithm on training data with limited, noisy, or imprecise supervision signals in a supervised learning setting, and the algorithm predicts additional information from the labels it was trained on [83]. This approach allows for the use of inexpensive weak labels to alleviate the burden of obtaining hand-labeled data sets, which can be impractical. Self-supervised learning obtains supervisory signals from the data itself, often leveraging the underlying structure in the data. The

general technique of self-supervised learning is to predict any unobserved part of the input from any observed part of the input [84]. For example, it can be used to predict past or future frames in a video from current ones. Self-supervised learning opens up a huge opportunity for better utilizing unlabelled data while learning in a supervised learning manner, which can potentially break the bottleneck for building more intelligent generalist models.

Due to the nature of the datasets and the tasks, the methods studied in this work are primarily under the supervised learning category, with the exception of one weakly supervised learning method.

1.4.1 Support Vector Machine

Support-vector machine (SVM) is a supervised learning algorithm used for classification and regression analysis. SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, such that the functional margin, i.e., the distance to the nearest training data point of any class, is maximized (Fig. 1.7) [85]. When the classes to discriminate are not linearly separable in the original finite-dimensional space, the space can be transformed into a much higher-dimensional feature space, making the separation by hyperplanes easier in that space. The mappings are defined in terms of a kernel function selected to suit the problem, which keeps the computational load reasonable [86]. The kernel trick efficiently allows much more complex, non-linear discrimination between sets that are not convex in the original space. Some common kernels include polynomial, Gaussian radial basis function, and hyperbolic tangent.

1.4.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of deep learning algorithm. CNNs assume a geometric relationship in the inputs, such as the rows and columns of the image [87]. They have been state-of-the-art methods for image and video classification, image segmentation,

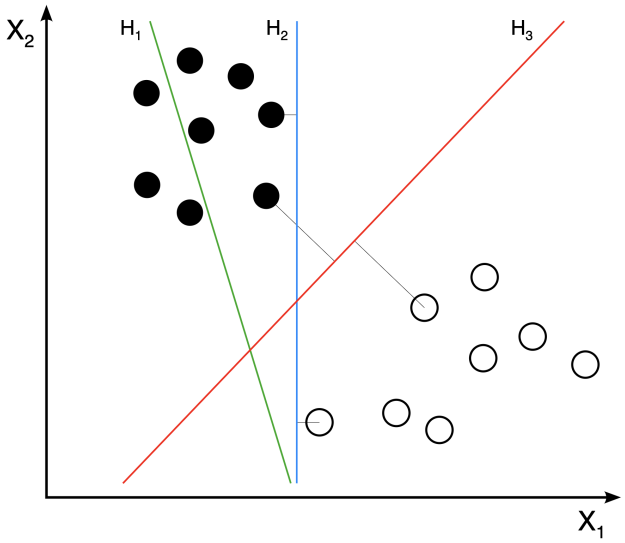
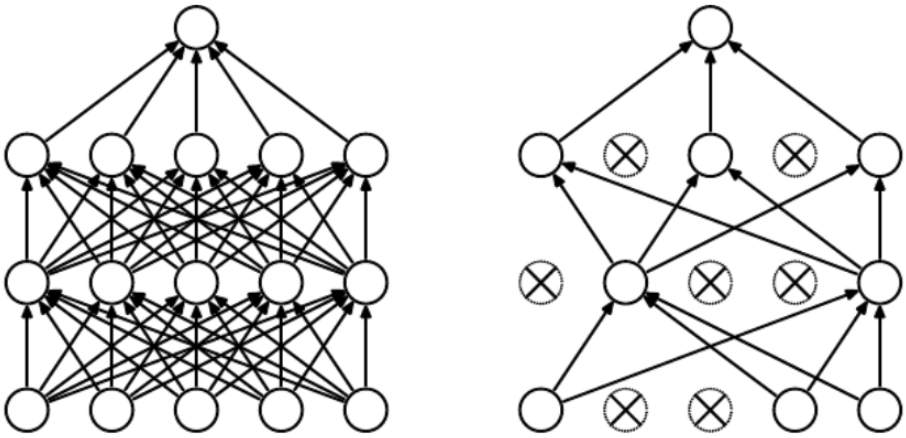


Figure 1.7: Example hyperplanes for discriminating the two classes (black and white circles). H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximal margin.

and image generation tasks [88]. One property of CNNs that makes them useful for many computer vision tasks is that the patterns they learn are translation invariant, meaning that once the network learns a pattern, it can detect the same pattern no matter where it is in the input image [89]. Each convolutional layer has filter elements (also called kernels). The filters are moved across the image to perform convolution operations. The step size for the movement of the filter element is called the stride. The receptive field is the region of the input space visible to a filter element. The output of a convolution layer is a tensor of feature maps, whose depth (or channel) dimension corresponds to the number of filters in the layer. Following each convolution operation, CNNs apply an activation function to each of the output units in the feature maps. The most commonly used activation is the rectified linear unit (ReLU), which has an output of zero for any negative value and keeps the value for any positive value. Pooling layers (e.g., max-pooling, average-pooling) are also commonly used in CNNs to downsample the feature maps and compensate for long computing times. A max-pooling layer, for example, takes the maximum value in the feature map within the filter size, rewarding the convolution function that extracts the most important features from the image.

Fully connected layers are usually used at the top of the model to complete the classification task. Fully connected layers have every unit in the previous layer connected to every unit in the next layer, and the final fully connected layer in the model is usually followed by a sigmoid (for binary classification or independent classes) or softmax (for mutually exclusive classes) activation operation to produce an output score for the probability of the input data belonging to each class.

Another critical component for training a CNN is regularization, which helps prevent overfitting on the training data, thus achieving balance in the bias-variance trade-off. One regularization approach is adding parameter norm penalties to the cost function. L1 and L2 regularizations are the most common in this category [90]. When the L1 norm (i.e., the sum of the absolute values of the vector) of the weights is used, the cost function penalizes the absolute value of the weights. When the L2 norm (i.e., the square root of the sum of the squared vector values) of the weights is added to the cost function, it is also known as weight decay. The difference in their effects is that L1 regularization shrinks the less important feature's coefficient to zero, thus removing some feature altogether, whereas L2 regularization forces the weights to decay towards but not exactly to zero. Another powerful regularization technique in deep learning is dropout [8]. At every iteration during training, some neurons are randomly selected to be removed along with all of their incoming and outgoing connections, as shown in Fig. 1.8 [8]. This means a different sub-model is trained in each iteration, resulting in multiple independent internal representations being learned by the model. As a result, the network becomes less sensitive to the specific weights of neurons, which, in turn, results in a network that is capable of better generalization and is less likely to overfit the training data. In addition, batch normalization is a layer often used in very deep neural networks to standardize the output of the previous layers on each mini-batch, which has the effect of stabilizing and accelerating the learning process and is also used as regularization to avoid overfitting [91].



(a) Standard Neural Net (b) After applying dropout.

Figure 1.8: Dropout neural net model. Left: A standard neural net with two hidden layers. Right: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped. Reprinted from [8].

All weights in a CNN are learned from the training data via the backpropagation algorithm [92]. While machine learning scientists and practitioners have a general understanding of the benefits and drawbacks of design choices when building a CNN model, there is not a formula to precisely determine all the design parameters such as the number and types of layers needed for a given problem; the process usually involves trial-and-error that requires experimentation of various model architectures to determine the optimal configuration for a given problem [89].

The CNN models employed in this work are mainly variations based on established state-of-the-art CNN architectures for image recognition tasks. The remainder of this section provides a brief overview of these architectures.

VGG

The VGG architecture was the winner of the localization portion and received second place in the classification portion of the 2014 ImageNet challenge [88, 93]. The primary innovation was the use of 3×3 convolutional filters in each layer, as well as increasing the depth compared

to prior architectures, such as AlexNet, the architecture that started the renaissance of deep learning [94]. The small size of the convolution filters allows VGG to have a large number of weight layers, which enables the model to learn more complex feature and hence improves performance.

There are two variants of the VGG architecture, namely VGG-16 and VGG-19, named for the number of layers in the models. Taking VGG-19 as an example, as illustrated in Fig. 1.9, the architecture consists of five convolutional blocks for a total of 16 convolutional layers, with a max-pooling layer between consecutive convolutional blocks, followed by three fully connected layers. All convolutional blocks consist of 3×3 filters, which is the smallest receptive field to capture the concepts of left/right, center, and up/down. The stride is 1 pixel. The max-pooling operation is performed over 2×2 pixel windows with a stride of 2. The first two fully connected layers both have 4096 channels, and the last fully connected layer's channel number depends on the number of output classes. The total number of weights is approximately 144 million [93].

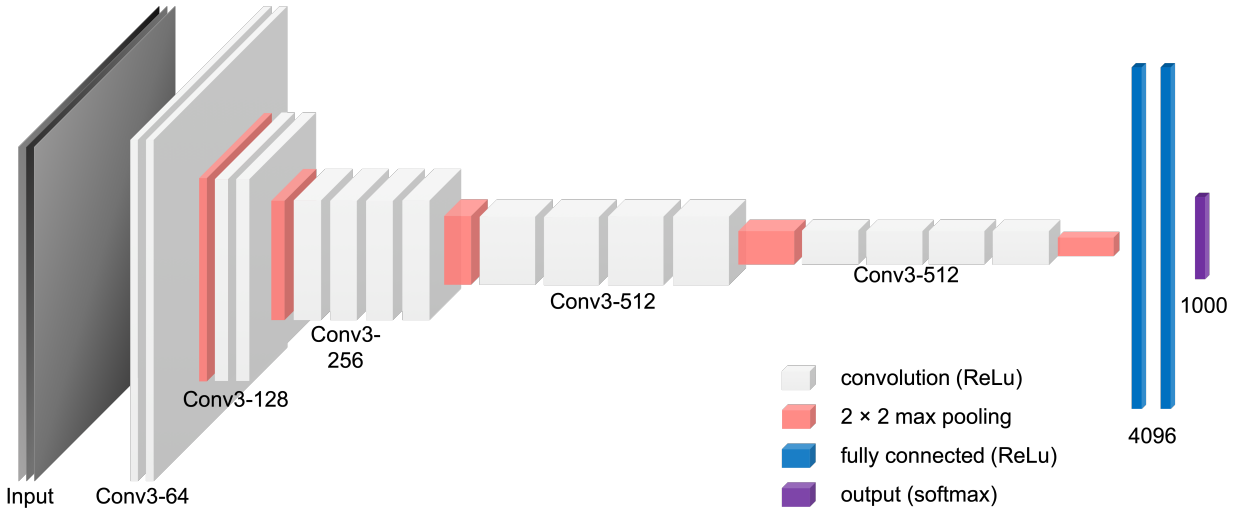


Figure 1.9: Illustration of the network architecture of VGG-19 model, in the case of a 1000-class classification, as in the ImageNet challenge.

VGG-19 has been successfully applied to medical imaging tasks [1], likely because the architecture has the capacity to learn complex features but is not too large relative to the

limited size of medical imaging datasets, achieving a balance in the bias-variance trade-off.

ResNet

ResNet won the 2015 ImageNet challenge with an accuracy of 95.5%, outperforming human classification performance (95%) on the ImageNet dataset [9, 88]. The innovation in ResNet was the addition of residual mapping, forming residual blocks as shown in Fig. 1.10. As CNN models grow deeper, they also become more difficult to optimize due to the problem of vanishing/exploding gradients, and model performances also saturate and then degrade beyond a certain depth, even on the training set [9]. Formally, if the desired underlying mapping is denoted as $\mathcal{H}(x)$, the stacked nonlinear layers in a residual block fit mapping $\mathcal{F}(x) := \mathcal{H}(x) - x$. The original mapping is recast into $\mathcal{F}(x) + x$. The authors of ResNet hypothesized that it would be easier to optimize the residual mapping than to optimize the original, unreferenced mapping. In the extreme case, if an identity mapping were optimal, the residual block would push the residual to zero rather than to fit an identity mapping by a stack of nonlinear layers, and the former is easier.

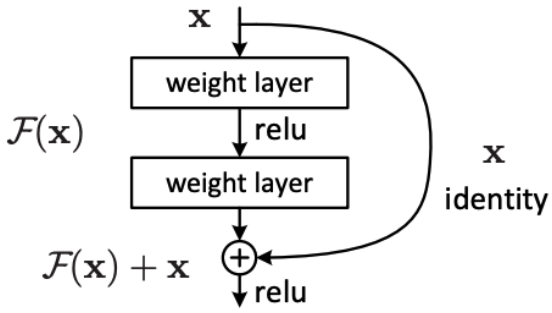


Figure 1.10: Illustration of a two-layer residual block, a building block of ResNet.

To update the weights during training of a CNN, the loss function is backpropagated through the network. When backpropagating through a network with many hidden layers, each layer has a small derivative, causing the gradient to decrease such that when it gets to the earlier layers, it is too small to update the weights in a meaningful way. This is referred

to as the vanishing gradient problem. Although this problem can be avoided by normalized initialization and intermediate normalization layers [91, 95, 96], residual connections are also a way to do so, because the short paths help preserve the gradient throughout the extent of very deep networks.

The authors of ResNet experimented with several ResNet models, whose detailed architectures are shown in Fig. 1.11 [9]. Thanks to residual connections, even the deepest variant, ResNet-152, presented has lower complexity than VGG-19 (11.3 billion FLOPs versus 19.6 billion FLOPs), despite being eight times deeper. Through experiments, they did observe improved convergence in very deep ResNet models compared with similarly structured plain networks that did not contain residual connections. Among the ResNet architectures, ResNet-50 has been chosen most frequently in medical imaging applications. Its complexity is usually appropriate for training with relatively small datasets in this domain.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 1.11: ResNet architectures for ImageNet classification. Building blocks are shown in brackets, with the numbers of blocks stacked. Downsampling is performed by conv3_1, conv4_1, and conv5_1 with a stride of 2. FLOPs = floating point operations per second. Reprinted from [9].

DenseNet

DenseNet is one of the state-of-the-art CNN architectures and has shown success in image classification tasks on ImageNet as well as medical images [10]. As illustrated in Fig. 1.12a, in a unit building block in DenseNet, named dense block, each layer has a direct connection with every other layer within the block, thus obtaining additional inputs from all preceding layers. Consequently, each layer is receiving collective knowledge from all preceding layers. The feature maps from previous layers and the current layer are concatenated in the channel dimension and passed on to all subsequent layers. A DenseNet model is constructed by stacking multiple dense blocks with transition layers in between, as shown in Fig. 1.12b.

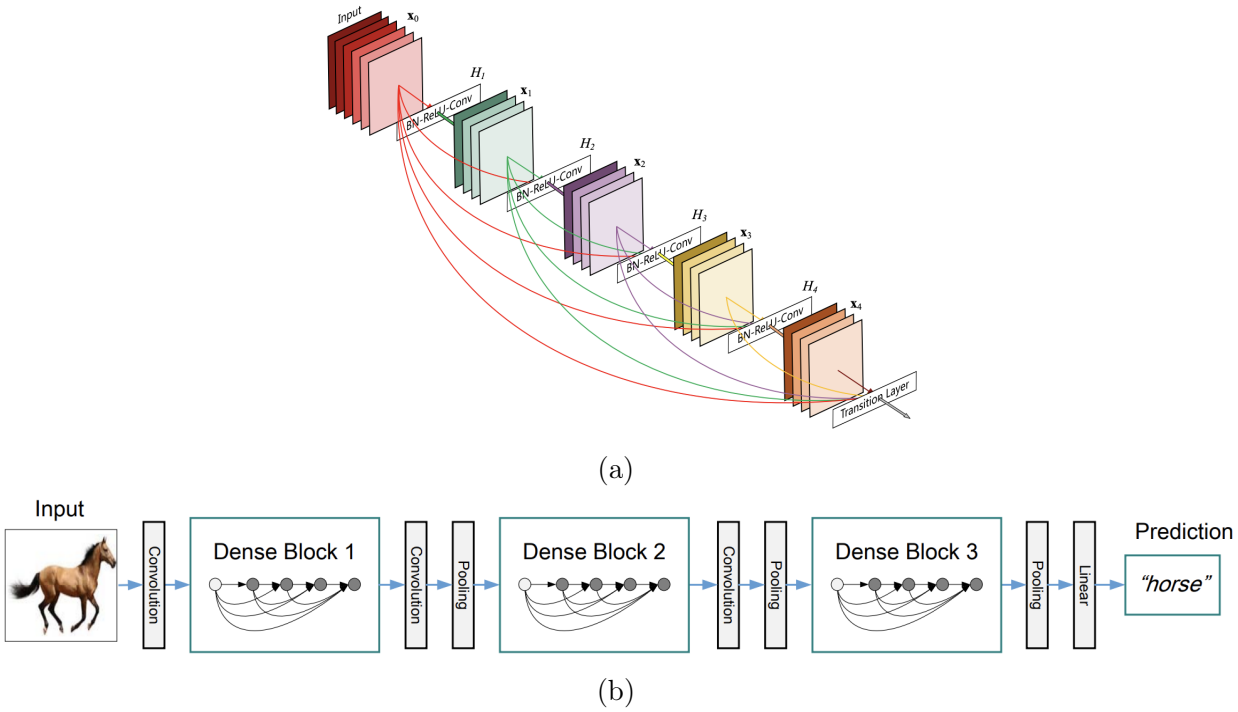


Figure 1.12: (a) Illustration of a five-layer dense block. (b) Illustration of a deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling. Each layer takes all preceding feature maps as input. Adapted from [10].

DenseNet has several advantages. As mentioned for ResNet, DenseNet also allows the error signal to be easily propagated to earlier layers more directly because of the skip connec-

tions. Moreover, DenseNet utilizes parameters more efficiently than alternative architectures such as ResNet. Since each layer receives feature maps from all preceding layers, the network can be thinner and compact, i.e., fewer channels can achieve the same classification performance. In addition, because features from various layers are concatenated, the features are more diversified and tend to have richer patterns than in ResNet, in which the element-wise addition is used. Using features of all complexity levels, including features with lower complexity levels from earlier layers, also tends to give more smooth decision boundaries. Thanks to these advantages, experimental results showed that, for example, DenseNet-201 with 20 million parameters yielded a similar validation error as the ResNet-101 with over 40 million parameters on ImageNet [10].

One drawback of DenseNet is that it is very demanding on the memory compared with ResNet, as the tensors from different layers are concatenated together. Among the DenseNet variants that have been investigated, DenseNet-121 has been used the most in medical imaging applications due to its relatively moderate complexity and memory requirement.

U-Net

U-Net is a segmentation network developed for image segmentation of biomedical images [11]. The network is based on the fully convolutional network and was modified to require fewer training images and to yield more precise segmentations [97]. The network consists of a contracting path and an expanding path, which gives it the symmetric u-shaped architecture, as illustrated in Fig. 1.13 [11]. The contracting path is a typical convolutional network that consists of repeated application of convolutions, each followed by a ReLU and a max-pooling operation. During the contraction, the spatial information is reduced while feature information is increased. The expanding path combines the feature and spatial information through a sequence of up-convolutions and concatenations with high-resolution features from the contracting path, allowing the network to propagate context information

to higher resolution layers. The output of U-Net gives per-pixel class predictions, hence serves as a segmentation algorithm.

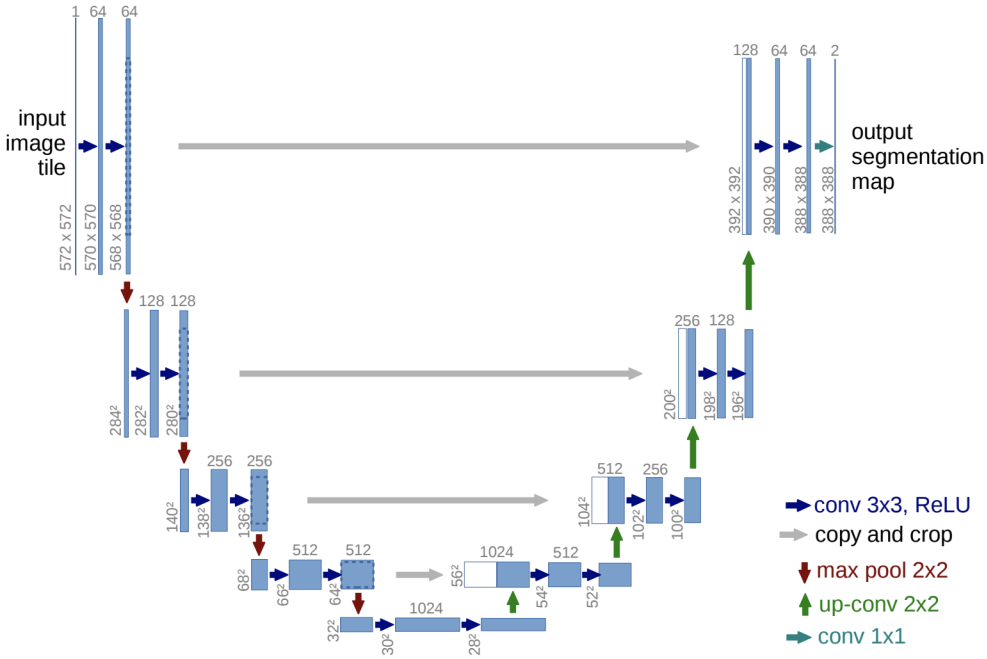


Figure 1.13: U-net architecture (example for 32×32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower-left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Reprinted from [11].

There have been many variations of U-Net adapted for specific segmentation tasks. One variation proposed by Clark et al. augments the U-Net with inception blocks and residual blocks [98]. The regular convolutional blocks in the architecture are replaced with inception blocks, shown in Fig. 1.14 [99]. Each inception block applies four convolution and pooling operations in parallel then concatenates the feature tensors at the end of the block. Merging of signals after parallel operations has been shown theoretically and experimentally to increase segmentation and classification accuracy [99]. Residual blocks (illustrated in Fig. 1.10) are added to the connections of the up- and down-sampling paths to enhance feature complexity, which is especially important in the first layer skip connection that has only undergone one convolution operation. This has been shown to help reduce areas of false

positive [98].

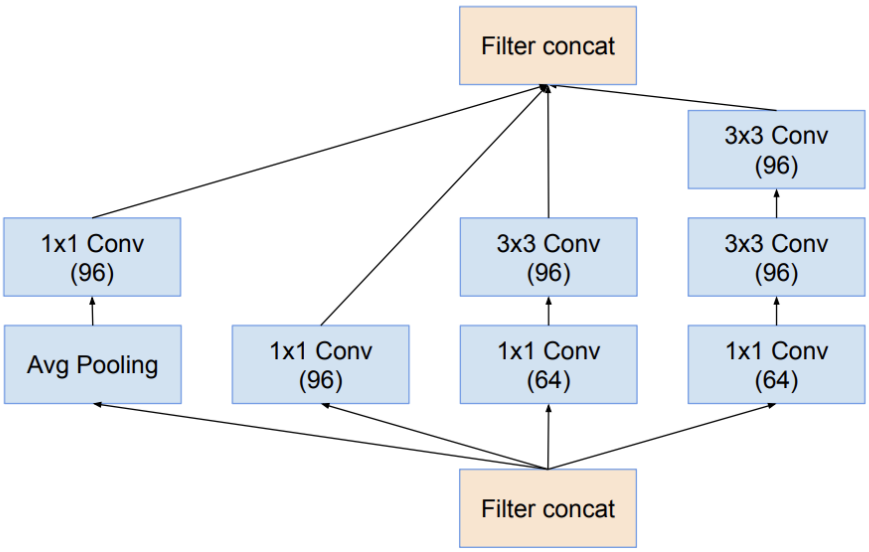


Figure 1.14: Illustration of the inception block used to augment the U-Net. An inception block is also the building block of an inception network architecture.

1.4.3 Multiple Instance Learning

Multiple instance learning (MIL) is a form of weakly supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag, but the individual instances are unlabeled. A bag is positively labeled if at least one instance in it is positive and is negatively labeled if all instances in it are negative. The goal of the MIL is to predict the labels of new, unseen bags. Several types of problems can be naturally formulated as MIL problems to leverage weakly labeled data. For example, in the drug activity prediction problem, the objective is to predict if a molecule can bind to the surface receptors of the target molecule [100]. A drug molecule can adopt a wide range of conformations, and it is labeled "active" if at least one of its conformations can bind to a binding site. Observing the effect of individual conformations is infeasible. By modeling a molecule as a bag and conformations that it takes as the instances in the bag, one can use

MIL to predict the drug activity for unseen molecules.

MIL has been increasingly used in many other application fields, such as image and video classification, document classification, and sound classification. Consequently, it has been used in diverse application fields such as computer vision and document classification [101]. In medical imaging, CADx algorithms can be trained with medical images for which only case-level diagnoses are available instead of costly local annotations provided by an expert. For example, digital pathology has seen rapid growth in recent years with the advent of digital microscopy, which makes it possible to convert glass slides into digital slides. The resulting large-scale whole-slide images (WSI) of tissue specimens contain billions of pixels and tens of thousands of cells. The size of 100 typical WSI studies is similar to the size of the entire ImageNet, a well-known visual database for computer vision research that contains over 14 million images [34]. Thoroughly reading a WSI is a laborious task that takes a highly trained pathologist several hours per slide. Deep learning, with its capabilities of processing huge amounts of data, holds a great promise to support pathologists in their daily routines; however, producing pixel-level annotations is labor-intensive and infeasible for large-scale datasets. MIL has shown great performance in addressing this issue by modeling WSIs as bags and smaller patches that contain cells as instances, and only using the slide-level labels that can be obtained much more easily.

Before the renascence of deep neural networks, the majority of machine learning systems consisted of two separated entities: a feature extractor and a classifier. Deep MIL, on the other hand, provides the possibility of training from end-to-end only using weakly labeled data. Figure 1.15 illustrates three types of deep MIL approaches: instance-based approach, embedding-based approach, attention-based approach [12]. In the instance-based approach, an instance score is obtained from the model for each instance in a bag, and then an MIL pooling layer is used to infer the bag label. In the embedded-based approach, MIL pooling occurs at the embedding (or feature) level to combine the instance embeddings to a single

bag embedding, which is then used to infer the bag label. The embedded-based approach generally yields better bag classification performance than the instance-based approach, but it is not desirable when interpretability is important as it cannot infer instance contribution [102]. The recently proposed attention-based approach combines the advantages of the previous two approaches. The instance embeddings are combined into a bag embedding by an attention mechanism consisting of two fully connected layers that are used to compute an attention weight for each instance. Let $H = \{\mathbf{h}_1, ..., \mathbf{h}_K\}$ be a bag of K embeddings, and then the attention-based MIL pooling is defined as follows:

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k, \tag{1.2}$$

where

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\right\}}, \tag{1.3}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$ are parameters [12]. Consequently, instances with a higher attention weight are contributing more to the bag embedding, and the attention scores provide interpretability. While the MIL pooling operators in previous approaches are predefined and nontrainable, the attention-based approach supports fully flexible and trainable MIL pooling alongside other components of a model.

1.4.4 Evaluation of Machine Learning Performance

There are two primary methods to evaluate the performance of a trained deep neural network: cross-validation and the use of an independent test set [89]. One round of cross-validation involves partitioning data into complementary subsets, using one subset for training and the other subset for testing. Multiple rounds of cross-validation are performed using different partitions, and the testing results are aggregated over the rounds to give an estimate of the

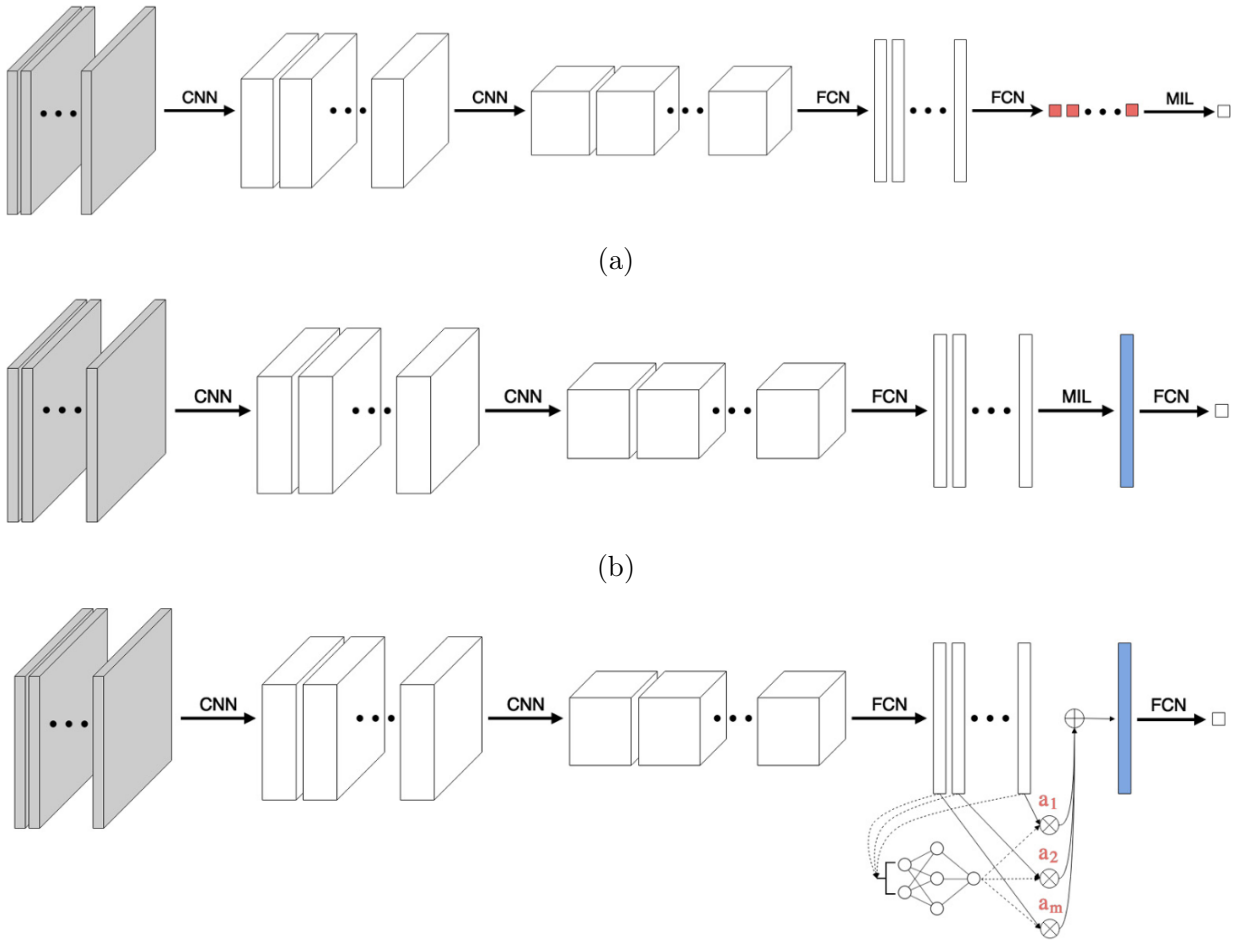


Figure 1.15: Deep MIL approaches: (a) the instance-based approach, (b) the embedding-based approach, (c) the proposed approach with the attention mechanism as the MIL pooling. Red color corresponds to instance scores; blue color depicts a bag vector representation. Reprinted from [12].

model's predictive performance. There are many types of cross-validation, such as leave-one-out and k-fold, that can be chosen depending on the specific scenario. When cross-validation is used simultaneously to select the best set of hyperparameters and to evaluate the generalization capacity, nested cross-validation is necessary. A $k \times l$-fold cross-validation, for example, contains an outer loop of k sets and an inner loop of l sets. The data set is split into k sets, and iteratively, each set is selected as the outer test set while the $k-1$ other sets are combined into the corresponding outer training set. Each outer training set is further divided into l sets, and iteratively, each set is selected as the inner test set, which we will refer to as validation set, and the $l-1$ other sets are combined into the corresponding inner training set. The inner loop is used to fit multiple models and/or using different hyperparameters and determine the best model type and/or hyperparameter set, while the outer loop is for providing an unbiased evaluation of the model. Alternatively, a single test set that is held out from the model development process can be used for independent testing of the trained model. Cross-validation can still be used during model development for hyperparameter tuning and/or model selection, but the test set needs to be sequestered until the evaluation step. It is also important to note that regardless of which evaluation method is used, the results are only meaningful if the human biases are controlled in the process of splitting the data into different subsets.

There are many metrics used to quantify the performance. For classification tasks, commonly used metrics include accuracy, the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, precision-recall curve, and more. AUC will be used as the primary metric in this work due to its robustness for unbalanced datasets and the conventions in the medical research community. Other metrics will also be calculated if they are appropriate and provide additional insights. Specific performance evaluation and statistical analysis for each method will be provided in each of the following chapters.

1.4.5 Explainability and Interpretability

In addition to the network performance evaluation, the ability of the algorithm to explain the result or provide interpretable output is of interest. Deep learning is often considered a "black box" due to its ability to learn features that are not explicitly programmed rules and the high complexity of learned features. "Black boxes" that yield high performances may be adequate for some applications; however, there is strong interest in explainable and interpretable AI algorithms in medical applications to benefit both the development of trustworthy algorithms and the clinical adoption of these technologies [103]. Explainability of the output can help developers identify abnormal behaviors or biases in the algorithm and allow them to make improvements by addressing those issues. Furthermore, interpretability of AI algorithms can not only help gain trust and confidence from the users in the medical community, but can also provide more valuable information to clinical practice than the prediction result alone.

1.5 Research Objectives and Scope

This dissertation studies AI-assisted medical image analysis in two applications: 1) breast cancer diagnosis on multiparametric MRI (mpMRI), and 2) COVID-19 diagnosis and prognosis on CXR. As mentioned in Section 1.2.1, mpMRI has assumed an important role in the routine clinical assessment of breast lesions, and therefore the ability to distinguish benign from malignant lesions on breast mpMRI is crucial. CADx methods, either using human-engineered features or deep learning, can potentially assist radiologists in image interpretation to reduce reading time and improve diagnostic performance. While mpMRI has shown benefits over DCE sequence alone in clinical studies, current CADx systems for breast lesion assessment on MRI are mainly focused on the DCE sequence. Thus, the first part of the work presented in this dissertation investigates CADx methods that leverage multiparametric information and high-dimensional information in mpMRI, with the goal of

improving the performance of breast lesion classification.

As COVID-19 emerged as a novel disease and has developed over the past 1.5 years, AI also holds promise to assist in the COVID-19 pandemic. While numerous studies have investigated the potential of using imaging data to help address critical questions and achieve optimal patient management, much is left to be explored due to the novelty and severe impact of the disease. The second part of this dissertation is dedicated to investigating computer-aided methods that can potentially assist in the early diagnosis and accurate prognosis of COVID-19 using CXR.

The outline for the remainder of the dissertation is as follows. Chapter 2 will propose and evaluate CADx methods for breast cancer diagnosis on mpMRI. Both human-engineered radiomics and deep-learning-based methods will be investigated, and classification performances will be compared to those based on single MRI sequences. Chapter 3 will discuss deep-learning-based methods that efficiently incorporate the high-dimensional information in MRI for breast cancer diagnosis. While high-dimensional medical images contain clinically valuable information, it is often underutilized in deep-learning-based image analysis due to computational constraints and data scarcity. The work presented in this chapter aims to address this bottleneck and further improve upon the performance reported in Chapter 2. Chapter 4 will focus on deep learning approaches to identify COVID-19 at patient presentation and predict future needs of intensive care using CXR. The role of various types of CXR images in the task will also be examined. Finally, Chapter 5 will summarize the findings and implications of this dissertation research and propose future research directions.

CHAPTER 2

MULTIPARAMETRIC MRI FOR BREAST CANCER DIAGNOSIS USING HUMAN-ENGINEERED RADIOMICS AND DEEP LEARNING

2.1 Introduction

As MRI is increasingly used for breast cancer screening and throughout the patient management process, computer-aided diagnosis (CADx)/radiomics systems continue to be developed to enable artificial intelligence (AI)-assisted image interpretation for radiologists and potentially enhance diagnostic performance [21, 30]. Section 1.2.1 discusses that clinical breast MRI has evolved from DCE-MRI to multiparametric MRI (mpMRI) to assess additional information and improve diagnostic performance and that T2-weighted (T2w) and diffusion-weighted MRI (DWI) are two commonly used sequences in mpMRI alongside the DCE sequence. Previous radiomics studies were primarily focused on using DCE-MRI [38, 104–106]. As MRI technology advances, radiomics methods for mpMRI have also started to be explored [37, 107, 108]. This chapter proposes and evaluates the performance of multiparametric radiomics methods that utilize multiple sequences in mpMRI and show that the complementary information provided in them can improve the diagnostic performance in the task of distinguishing between benign and malignant breast lesions. This work demonstrates strong potential in CADx systems to leverage multiparametric information from mpMRI to predict the probability of breast lesion malignancy. Both human-engineered radiomic features and deep-learning-based methods will be investigated.

In the human-engineered radiomics methodology, radiomic features are designed for and extracted from each MRI sequence, and classification is performed using support vector machines (SVMs). Information from different mpMRI sequences is integrated at two different levels of the classification framework, namely (i) at the feature level by concatenating ra-

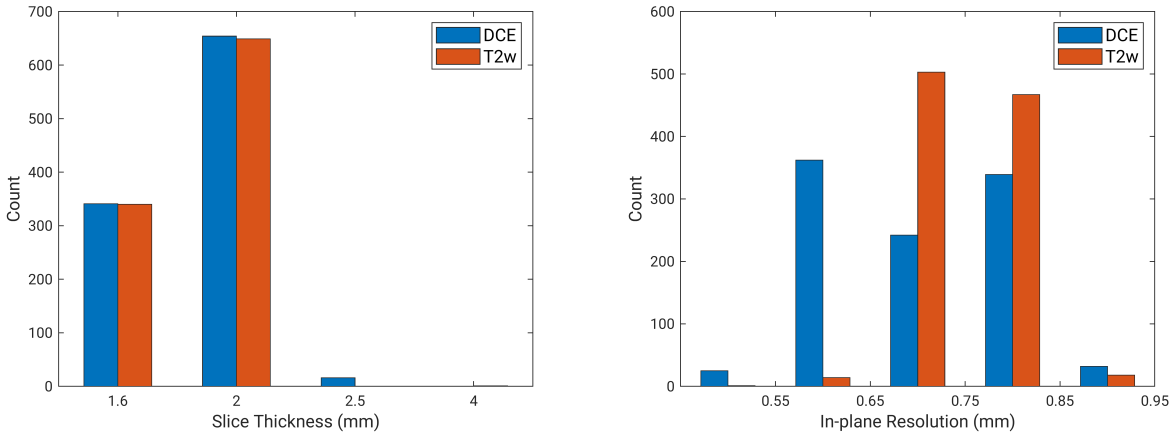
diomic features extracted from multiple sequences (feature fusion), and (ii) at the classifier output level by aggregating the outputs from the single-sequence SVMs (classifier fusion). The study also shows the effect of dataset size and demonstrates the value of handling variability in mpMRI protocols in CADx systems as in clinical settings. In the deep-learning-based methodology, deep transfer learning is used to extract and pool multi-level features using a pretrained CNN and perform classification using SVM. Multiparametric information integration is performed at three different levels of the classification framework, namely via (i) at the image level by inputting merged mpMRI images to the CNN (image fusion), (ii) feature fusion, and (iii) classifier fusion. As a supplementary investigation, a more novel convolutional neural network (CNN) architecture will be used to validated the findings. In addition, since the DWI sequence has not been sufficiently examined in prior CADx studies, CNN features will be compared with human-engineered radiomic features for the DWI sequence.

2.2 Dataset

The database was retrospectively collected from University of Chicago Medical Center under Health Insurance Portability and Accountability Act (HIPAA)-compliant Institutional Review Board (IRB) protocols. All clinical information and images in this study were de-identified before they were made available to the investigators, and hence consent from the participants was waived. The MRI exams in the database were consecutively acquired from 2007 to 2013 and imaged at a single institution. MRI studies that did not exhibit a visible lesion, lesions that did not have validation of the final diagnosis, or lesions that could not be clearly allocated to either the benign or malignant category were excluded. In total, the database used in the deep learning study consisted of 927 unique breast lesions from 616 women.

Images in the database were acquired using either 1.5 T (66%) or 3 T (34%) Philips

Achieva scanners. Each MR study contained a DCE-MRI sequence and a T2w MRI sequence acquired during the same exam, and exams for a subset of 397 lesions from 302 patients also contained a DWI sequence. The scanning sequences for DCE, T2w, and DWI were a T1-weighted spoiled gradient sequence with fat saturation, a T2-weighted fast spin echo sequence with flow compensation, and a diffusion-weighted fast spin echo sequence with fat saturation, respectively. In-plane resolution and slice thickness varied within each sequence across the dataset. The slice thickness was consistent across the DCE and T2w sequences in 96% of the exams, while the in-plane resolution was consistent across the two sequences in 46% of the exams (Fig. 2.1). The DWI sequence had coarser spatial resolutions than the other sequences. The DWI sequence was also obtained with various degrees (ranging from two to five) of diffusion weighting as measured by the b-value.



(a) (b)

Figure 2.1: Distribution of (a) slice thickness and (b) in-plane resolution of the dynamic contrast-enhanced (DCE) sequences and T2-weighted (T2w) sequences in the multiparametric MRI database [13].

The clinical characteristics of the dataset are detailed in Table 2.1. The ground truth for malignancy was obtained from pathology and radiology reports. For all lesions categorized at MRI as Breast Imaging Reporting and Data System (BI-RADS) category 4, 5, or 6, diagnosis validation was achieved by histopathologic analysis. For all lesions categorized at

MRI as BI-RADS category 2 or 3, benign diagnoses were validated by MRI follow-up of at least 24 months. There is no severe imbalance between the benign and malignant classes in terms of acquisition parameters and lesion characteristics, eliminating concerns for additional confounding variables.

Table 2.1: Clinical characteristics of the dataset [13]. The number of lesions is shown, along with the percentage of the total. Patient age is summarized on a patient basis, and lesion information (malignancy status and subtypes) is summarized on a lesion basis.

Benign/malignant prevalence	Benign: 199 (21.5) Malignant: 728 (78.5)
Age (years)[a]: mean ± sd[b]	55.0 ± 12.8 Unknown: 97
Benign lesion characteristics	
Lesion size (mm)[c]	Mean: 8.86 Median: 7.33 Range: 3.38–42.8
Lesion subtype	Fibroadenoma: 60 (30.2) Columnar change: 15 (7.5) Papilloma: 13 (6.5) Parenchyma tissue: 12 (6.0) Fibrotic tissue: 10 (5.0) Hyperplasia: 8 (4.0) Cystic change: 6 (3.0) Fat necrosis: 5 (2.5) Other: 27 (13.6) Unknown: 43 (21.6)

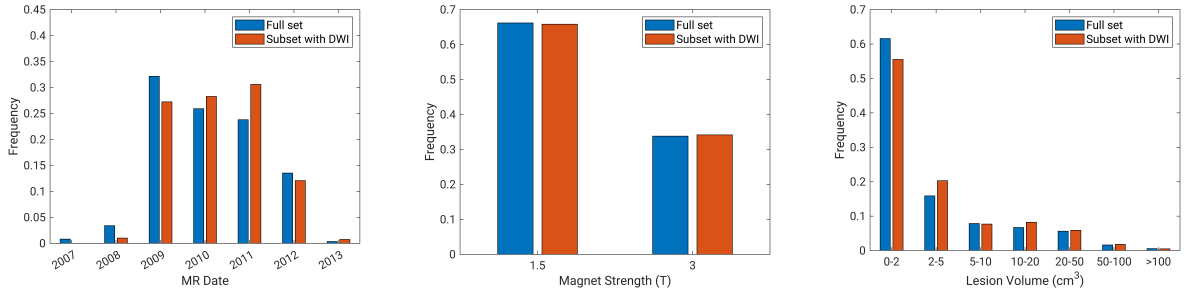
Table 2.1: Clinical characteristics of the dataset (continued)

Malignant lesion characteristics	
Lesion size (mm)[c]	Mean: 17.9
	Median: 14.9
	Range: 3.37–73.7
Lesion subtype	IDC[d]: 147 (20.2)
	DCIS[e]: 120 (16.5)
	IDC+DCIS: 359 (49.3)
	ILC[f]: 31 (4.3)
	ILC mixed: 26 (3.6)
	Other: 33 (4.5)
	Unknown: 12 (1.6)
Estrogen receptor status	Positive: 410 (56.3)
	Negative: 128 (17.6)
	Unknown: 190 (26.1)
Progesterone receptor status	Positive: 352 (48.4)
	Negative: 184 (25.3)
	Unknown: 192 (26.4)
HER-2[g] status	Positive: 87 (12.0)
	Negative: 404 (55.5)
	Equivocal: 5 (0.7)
	Unknown: 232 (31.9)

Table 2.1: Clinical characteristics of the dataset (continued)

[a] Numbers in parentheses are percentages. For some subjects, only the decade of age was available (e.g., "60s") as part of the patient information de-identification process. In these situations, the middle of the decade was used for the calculation of the mean subject age.
[b] sd = standard deviation
[c] Lesion size is measured by the effective diameter, i.e., the greatest dimension of a sphere with the same volume as the lesion.
[d] IDC = invasive ductal carcinoma
[e] DCIS = Ductal carcinoma in situ
[f] ILC = Invasive lobular carcinoma
[g] HER-2 = human epidermal growth factor receptor 2

For human-engineered radiomics, lesions whose DCE time intervals were unknown were also excluded, as it was a necessary parameter for calculating some of the features. Thus, a subset of 852 unique breast lesions from 612 women was included in this part of the study. Exams for a subset of 389 lesions from 299 patients contained a DWI sequence. For the full set of cases and the subset that contained DWI, the distributions of the acquisition date, magnetic strength, and lesion volume are shown in Fig. 2.2. The similar distributions in the full set and the subset with DWI eliminate concerns about additional confounding variables. The clinical characteristics of the dataset are detailed in Table 2.2.



(a) (b) (c)

Figure 2.2: Distribution of (a) MRI acquisition date, (b) magnet strength, and (c) lesion volume of the full dataset and the subset that contains diffusion-weighted imaging (DWI) sequence in the multiparametric MRI database.

Table 2.2: Clinical characteristics of the dataset [14]. Patient age is summarized on a patient basis, and lesion information (malignancy status and subtypes) is summarized on a lesion basis. The full set is a mixture of cases imaged using either two or three sequences, and the diffusion-weighted imaging sequence (DWI) subset contains cases imaged using three sequences.

	Full set (N = 852)	DWI subset (N = 389)
Benign/malignant prevalence	Benign: 195 (22.9)	Benign: 66 (17.0)
	Malignant: 657 (77.1)	Malignant: 323 (83.0)
Age (years)[a]: mean ± sd[b]	55.1 ± 12.8	56.4 ± 12.9
	Unknown: 96	Unknown: 12
Benign lesion characteristics		
Lesion subtype	Fibroadenoma: 60 (30.8)	Fibroadenoma: 18 (27.3)
	Columnar change: 15 (7.7)	Columnar change: 5 (7.6)
	Papilloma: 13 (6.7)	Papilloma: 6 (9.1)
	Parenchyma tissue: 11 (5.6)	Parenchyma tissue: 8 (12.1)
	Fibrotic tissue: 10 (5.1)	Fibrotic tissue: 5 (7.6)
	Hyperplasia: 8 (4.1)	Hyperplasia: 5 (7.6)
	Cystic change: 6 (3.1)	Cystic change: 3 (4.5)
	Fat necrosis: 4 (2.1)	Fat necrosis: 3 (4.5)
	Other: 26 (13.3)	Other: 12 (18.2)
	Unknown: 42 (21.5)	Unknown: 1 (1.5)
Malignant lesion characteristics		
Lesion subtype	IDC[d]: 133 (20.2)	IDC: 71 (22.0)
	DCIS[e]: 118 (18.0)	DCIS: 20 (6.2)
	IDC+DCIS: 316 (48.1)	IDC+DCIS: 197 (61.0)

Table 2.2: Clinical characteristics of the dataset (continued)

	ILC[f]: 27 (4.1)	ILC: 15 (4.6)
	ILC mixed: 24 (3.7)	5 (1.5)
	Other: 28 (4.3)	Other: 15 (4.6)
	Unknown: 12 (1.6)	
Estrogen receptor status	Positive: 408 (62.1)	Positive: 235 (72.8)
	Negative: 127 (19.3)	Negative: 83 (25.7)
	Unknown: 122 (18.6)	Unknown: 5 (1.5)
Progesterone receptor status	Positive: 350 (53.3)	Positive: 209 (64.7)
	Negative: 183 (27.9)	Negative: 108 (33.4)
	Unknown: 124 (18.9)	Unknown: 6 (1.9)
HER-2[g] status	Positive: 87 (13.2)	Positive: 54 (16.7)
	Negative: 401 (61.0)	Negative: 240 (74.3)
	Equivocal: 5 (0.8)	Equivocal: 2 (0.6)
	Unknown: 164 (25.0)	Unknown: 27 (8.4)

[a] Numbers in parentheses are percentages. For some subjects, only the decade of age was available (e.g., "60s") as part of the patient information de-identification process. In these situations, the middle of the decade was used for the calculation of the mean subject age.
[b] sd = standard deviation
[c] Lesion size is measured by the effective diameter, i.e., the greatest dimension of a sphere with the same volume as the lesion.
[d] IDC = invasive ductal carcinoma
[e] DCIS = Ductal carcinoma in situ
[f] ILC = Invasive lobular carcinoma
[g] HER-2 = human epidermal growth factor receptor 2

2.3 Methods

2.3.1 Human-Engineered CADx

Single-Sequence Methods

Figure 2.3 illustrates the human-engineered radiomic features extraction, machine learning classification, and evaluation process for both single-sequence and mpMRI approaches. Lesions were segmented separately from each sequence using a fuzzy C-means method requiring only the manual indication of a seed point [109]. Radiomic features were designed based on the biological phenotypes of lesions. Fifty radiomic features that characterize lesions in terms of their size, shape, morphology, enhancement texture, kinetics, and kinetics variance were extracted from DCE images [25–29, 110]. Likewise, three morphological features and 14 texture features, as well as the mean and the variance of the signal intensity, were extracted from T2w images [107]. In addition, six first-order radiomic features were extracted from the ADC maps of DWI images [16]. Morphological or texture features were not calculated from DWI due to its coarse resolution. Radiomic features related to contrast enhancement on the DCE sequence were calculated in 4D, and all other features were calculated in 3D across the entire lesion. A complete list of radiomic features and their descriptions is included in Appendix A.

SVM classifiers with Gaussian radial basis function kernel were trained on the extracted radiomic features to differentiate between benign and malignant lesions (Python Version 3.7, Python Software Foundation) [111]. SVM was chosen over other classification methods due to its relative robustness to correlated data, which is an attribute of the radiomic features. Each SVM classifier was trained and evaluated using nested five-fold cross-validation, where the inner cross-validation was used for model development, and the outer cross-validation was used for testing. Within each training fold in the outer cross-validation loop, two SVM hyperparameters, namely the scaling parameter γ and the regularization parameter C, were

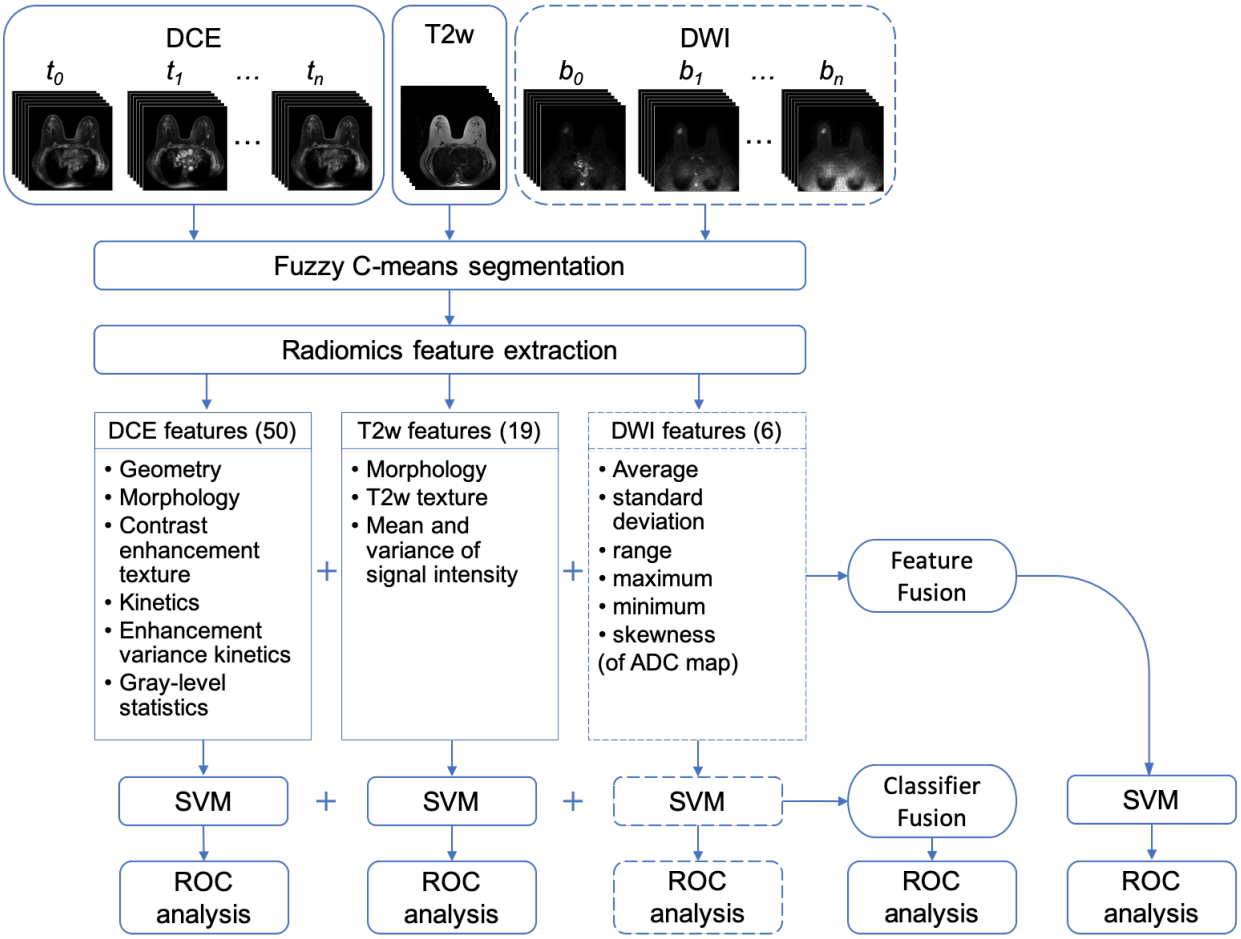


Figure 2.3: Lesion classification pipeline based on diagnostic images [14]. Radiomic features were extracted from dynamic contrast-enhanced (DCE), T2-weighted (T2w), and diffusion-weighted MRI (DWI) sequences. The mpMRI information was incorporated in two different ways: feature fusion, i.e., merging radiomic features extracted from all sequences to train a support vector machine (SVM) classifier, and classifier fusion, i.e., aggregating the probability of malignancy output from all single-sequence classifiers via soft voting. Parentheses contain the numbers of features extracted from each sequence. The dashed lines for DWI indicate that the DWI sequence was only included in the classification process when it was available, while the DCE and T2w sequences were available for all lesions and thus were always included. ADC = apparent diffusion coefficient, ROC = receiver operating characteristic.

optimized on a grid search with an internal five-fold cross-validation [112]. Predictions on the five test folds in the outer cross-validation loop were aggregated for classification performance evaluation. Splitting was performed by patient, keeping all lesions from a patient in the same fold to eliminate the bias due to using correlated lesions for training and testing. Class prevalence was held constant across all cross-validation folds. Each training set was standardized to zero mean and unit variance, and the corresponding test set was standardized using the statistics of the training set. To address the problem of class imbalance, a misclassification penalty for cases in each class was assigned to be inversely proportional to its prevalence in the training data.

Multiparametric Methods

We investigated integrating information from the three MRI sequences at two different levels of the classification framework, as illustrated in Fig. 2.3. The two mpMRI approaches are referred to as feature fusion and classifier fusion. For the feature fusion approach, radiomic features extracted from each sequence separately were concatenated to form an ensemble of features, which was then inputted to an SVM classifier. The classifier training process then followed the single-sequence methods. For the classifier fusion approach, probability of malignancy (PM) outputs from the single-sequence SVM classifiers were aggregated via soft voting. That is, the PM outputs were averaged across all single-sequence classifiers to yield prediction scores.

Protocol Variability

Missing data is a common challenge in multi-modality imaging studies. Conventional methods typically discard modality-incomplete subjects, which reduces the subjects that can be used to train a diagnosis model and hence may degrade the diagnostic performance. To mimic potential clinical situations where radiologists perform assessments based on MRI ex-

ams acquired using different imaging protocols that contain a variable number of sequences, the analyses were first performed on the entire dataset of 852 lesions, in which exams contained either two or three sequences. For the feature fusion approach, an SVM classifier was trained on features extracted from three sequences for the subset of lesions for which all three sequences were acquired during their MRI exams, and another SVM classifier was trained on features extracted only from DCE and T2w sequences for the remaining lesions for which DWI was not acquired. For the classifier fusion approach, output PMs from all applicable single-sequence SVM classifiers were aggregated via soft voting and subsequently inputted to ROC analysis and sensitivity/specificity calculations.

The same analyses were then performed on the subset of 389 lesions whose mpMRI protocol contained three sequences, discarding the modality-incomplete subset. The performances of mpMRI classifiers trained on this subset were compared with those trained on the full dataset to demonstrate the effect of the dataset size and the benefit of using all available data even when a subset contains missing sequences.

2.3.2 Deep-Learning-Based CADx

Single-Sequence Methods

Figure 2.4 schematically shows the machine learning classification and evaluation process for both single-sequence and mpMRI schemes. Lesions were segmented using a fuzzy C-means method requiring only the manual indication of a seed-point [109]. Lesion segmentations were not directly used as input to the CNN but enabled automatic region of interest (ROI) construction described below. To capture the 4D (volumetric and temporal) characteristics of the lesions from DCE sequences, maximum intensity projection (MIP) images of the second postcontrast subtraction DCE-MRI series were used as the input to a deep learning network [105]. The second post-contrast time point was chosen because the BI-RADS atlas defines the initial phase of enhancement as the first two minutes after contrast administration, which

has diagnostic utility for distinguishing benign and malignant breast lesions [4]. From the T2w sequence of each lesion, the slice that contained the largest lesion area according to the automatic lesion segmentation was selected as the representative center slice, which was used as the input to a deep learning network. The T2w center slice was rescaled using bicubic interpolation to match the in-plane resolution of its corresponding DCE sequence. To avoid confounding contributions from distant voxels, an ROI around each lesion was cropped from the image to use in the subsequent classification process. The ROI size was chosen based on the maximum dimension of each lesion and was held constant across sequences. A small part of the parenchyma, 3 pixels wide around the lesion, was included in each ROI. Appropriate shifts in the coordinates were applied to ensure that the DCE and T2w ROIs were cropped from the same location relative to the lesion. This study does not consider the DWI sequence due to its limited availability in the database, but a CNN-based method will be applied to DWI later in this chapter, and further investigations will be discussed in Chapter 3.

Through transfer learning, CNN features were extracted separately from the ROIs of the DCE subtraction MIPs and the ROIs of the T2w center slices using the publicly available VGG-19 model [93], pretrained on ImageNet [34]. Pretrained VGG-19 networks, which consist of three channels (red, green, and blue, or RGB), have previously been shown to be useful in transfer learning for breast lesion analyses [104, 105, 113]. The ROIs were grayscale for the single-sequence DCE and T2w image datasets and were duplicated across the three channels. Feature vectors were extracted at various network depths from the five max-pooling layers of the VGG-19. These features were then average-pooled along the spatial dimensions and normalized with Euclidian distance. The pooled features were then concatenated to form a CNN feature vector of 1472 features for a given lesion [104, 113].

SVM classifiers with Gaussian radial basis function kernel were trained on the CNN features to differentiate between benign and malignant lesions (Python Version 3.4.2, Python Software Foundation) [111]. SVM was chosen over other classification methods due to its

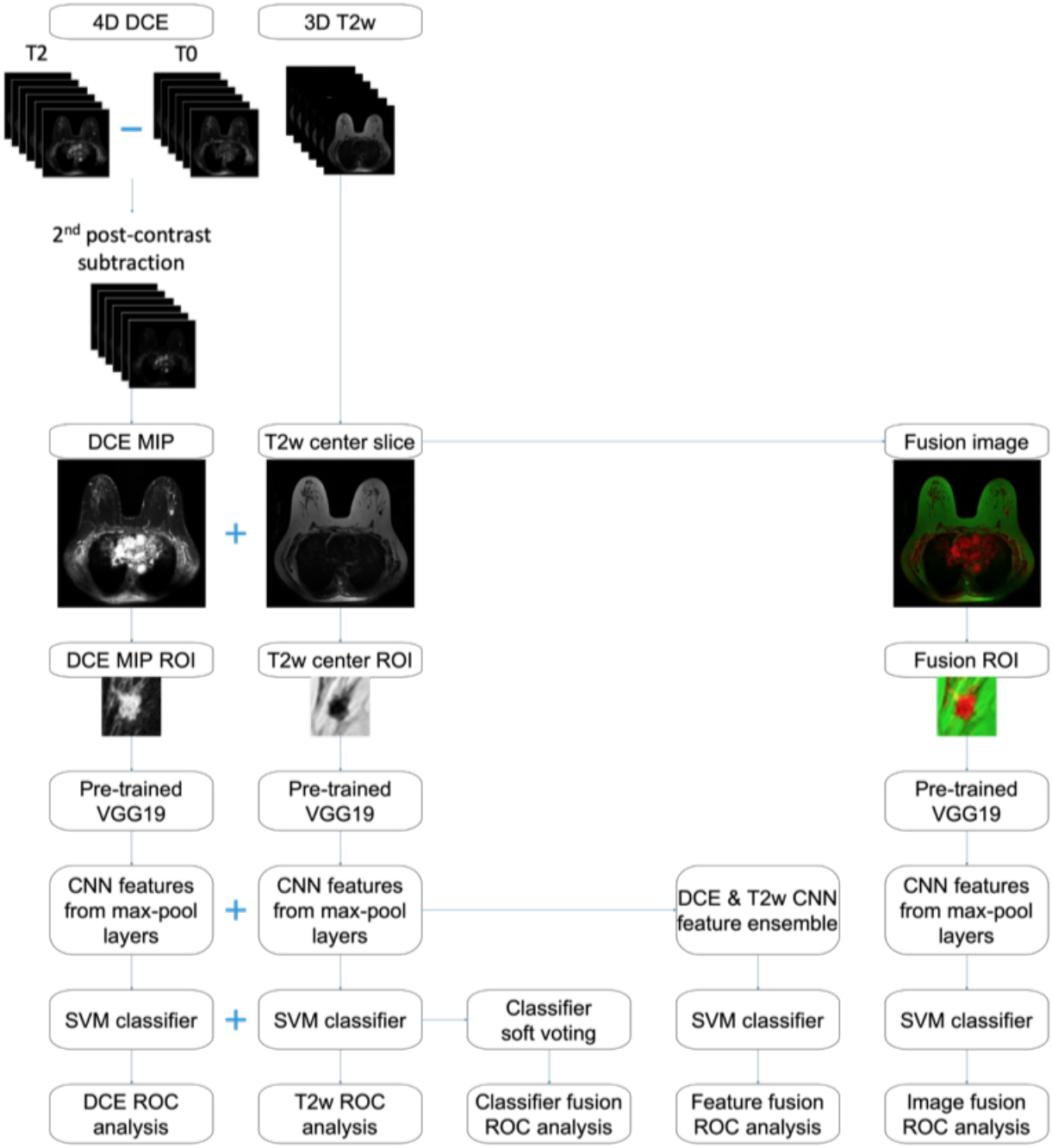


Figure 2.4: Lesion classification pipeline based on diagnostic images [13]. Information from dynamic contrast-enhanced (DCE) and T2-weighted (T2w) MRI sequences are incorporated in three ways: image fusion, i.e., fusing DCE and T2w images to create RGB composite image, feature fusion, i.e., merging convolutional neural network features extracted from DCE and T2w as the support vector machine (SVM) classifier input, and classifier fusion, i.e., aggregating the probability of malignancy output from the DCE and T2w classifiers via soft voting. MIP = maximum intensity projection. ROI = region of interest. ROC = receiver operating characteristic.

ability to handle sparse high-dimensional data, which is an attribute of the CNN features. Principal component analysis fit on the training set was applied to both training and test sets to reduce feature dimensionality [114]. The rest of the classifier training and evaluation process follows Section 2.3.1.

Multiparametric Methods

We explored integrating information from both the DCE and T2w MRI sequences at three different levels of the classification framework, as illustrated in Fig. 2.4. The three mpMRI schemes are referred to as image fusion, feature fusion, and classifier fusion.

For the input image fusion scheme, a three-channel RGB fusion image was constructed for each lesion by inputting the DCE MIP into the red channel, the T2w center slice into the green channel, and leaving the blue channel of the network blank. A composite ROI was cropped from the fusion image, which was then inputted into the pretrained VGG-19 network for feature extraction. Figure 2.5 includes an example to illustrate the process of ROI extraction from MRI images and creating RGB fusion ROIs. The classifier training process then followed the single-sequence methods to predict PMs. The feature fusion and classifier fusion approaches follow the descriptions in Section 2.3.1.

Inter-Sequence Image Registration

A preliminary study was performed to investigate whether image registration between DCE and T2w sequences would improve the performance of the proposed mpMRI classification schemes, especially the image fusion method. The T2w center slices were rescaled to match the in-plane resolution and then registered to the corresponding slice of the second post-contrast DCE image using a multi-modality rigid registration method that consists of translation and rotation [115, 116]. The same five classification mechanisms were evaluated after image registration.

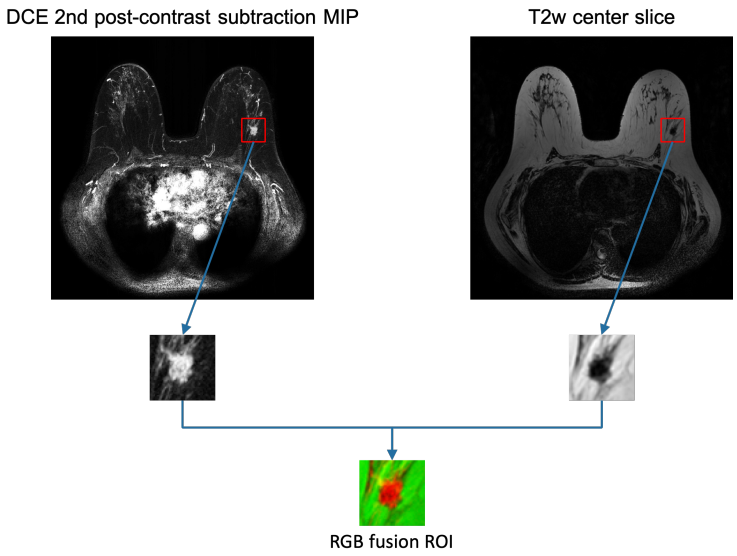


Figure 2.5: An example of the image fusion process [13]. A dynamic contrast-enhanced (DCE)-MRI transverse second post-contrast subtraction maximum intensity projection (MIP) and a T2-weighted (T2w)-MRI transverse center slice are shown with their corresponding regions of interest (ROIs) extracted. The RGB fusion ROI is created by inputting the DCE ROI into the red channel and the T2w ROI into the green channel.

2.3.3 Evaluation and Statistical Analysis

Classifier performances were evaluated using receiver operating characteristic (ROC) curve analysis, with the area under the ROC curve (AUC) serving as the figure of merit [117, 118]. The 95% confidence intervals (CIs) of the AUCs were calculated by bootstrapping the posterior PMs (2000 bootstrap samples) [119]. Other metrics, including sensitivity and specificity, positive predictive value (PPV), and negative predictive value (NPV), were calculated at the optimal operating point on the ROC curve that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$ and reported for each classifier [37].

The AUC values of the mpMRI approaches were compared with those of the single-sequence classifiers using the DeLong test [120, 121]. Bonferroni-Holm corrections were used to account for multiple comparisons [122], and a corrected $P < 0.05$ was considered to indicate a statistically significant difference in performance. Equivalence testing was performed to assess if image registration had any effect on the classification performances

of the CNN-based methods [123]. An equivalence margin of difference in AUC = 0.05 was chosen *prima facie*. To assess the performance reproducibility of the CNN-based methods, the highest performing classifier of the three mpMRI methods was trained and evaluated 100 times using different random seeds for the cross-validation split, and the mean and standard error of AUC was calculated from all the runs.

2.4 Further Investigations

2.4.1 ResNet Feature Extraction for mpMRI

A more novel CNN architecture, ResNet-50, was used to validated the findings using VGG-19 regarding the integration of multiparametric approaches as well as the effect of feature pooling across various levels within the network. The benefit of pooling features extracted from various depths of the network was examined on the set of DCE MIP images. Last-layer features (2048 features) were extracted from the last average pooling layer in the ResNet-50 architecture immediately before the final classification layer. In comparison, features were extracted from the 16 bottleneck layers of ResNet-50, i.e., the ReLU activations of the merging layer after each residual block (see Fig. 1.11 for detailed architecture). Similar to the feature pooling method previously applied to VGG-19, these features were then average-pooled along the spatial dimensions and normalized with Euclidian distance [104, 113]. The pooled features were then concatenated to form a feature vector for a given lesion, as illustrated in Fig. 2.6. The classification methods and the evaluation process follow Section 2.3.2 and Section 2.3.3.

2.4.2 Human-Engineered Radiomics versus Deep Learning for DWI

Since DWI is a relatively novel sequence in clinical practice and has not been sufficiently studied in CADx, we investigated the utility of human-engineered features and the CNN-

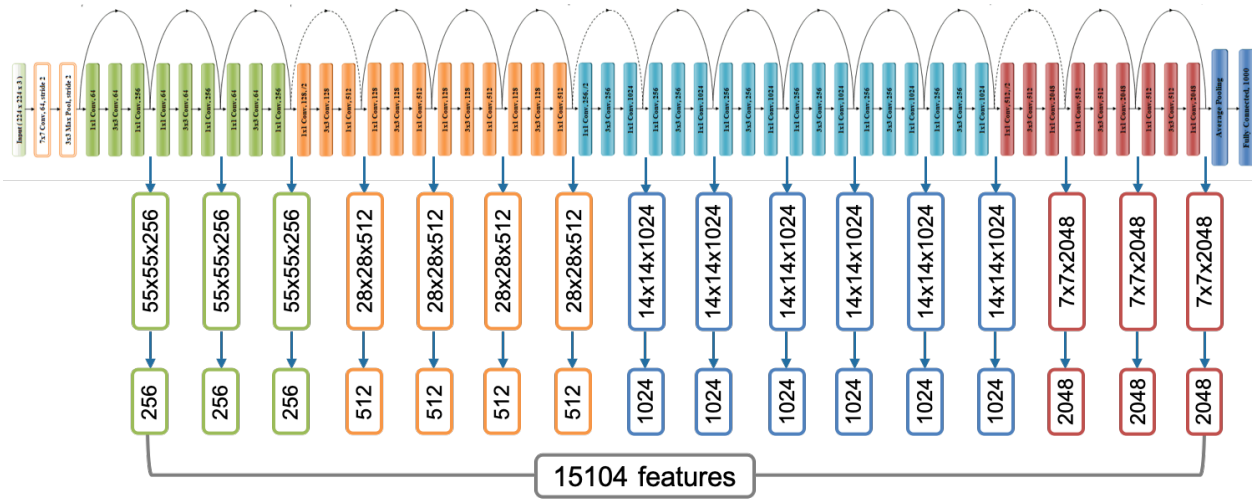


Figure 2.6: Feature pooling from various levels of ResNet-50.

based features extracted from DWI images. As described in Section 2.2, 397 unique breast lesions from 302 women in the database were imaged using protocols that included the DWI sequence. Among them, 69 lesions were benign and 328 were malignant. Following the method described in Section 2.3.1, six human-engineered radiomic features were extracted from the ADC map. For CNN feature extraction, a lesion ROI was constructed with DWI ROIs at two b-values, either 0 and 800 mm^2/s or 0 and 1000 mm^2/s, depending on availability. These two ROIs and a blank channel composed an RGB ROI for each lesion as input to the CNN. The two types of b-value pairs had similar prevalence across benign and malignant lesions (33%/67% for benign lesions and 31%/69% for malignant lesions). CNN features were extracted from the central slice ROIs using a VGG-19 model, following the method detailed in Section 2.3.2. The classification methods and the evaluation process follow Section 2.3. Furthermore, a fusion classifier was created by averaging the outputs from the ADC-based and CNN-based classifiers for each lesion and evaluated.

2.5 Results

2.5.1 Multiparametric MRI CADx

Human-Engineered CADx

Figures 2.7 and 2.8 show the comparison between the PMs predicted by the single-sequence classifiers using DCE and T2w features. Although the majority of benign and malignant classes are separated from each other, there exists notable disagreement between the two single-sequence classifiers, suggesting that a fusion technique for features extracted from various mpMRI sequences may improve the predictive performance. Figure 2.7 also shows example lesions upon which these two classifiers agree or disagree, with their lesion types noted in the caption. For example, the benign papilloma lesion on the lower right was inaccurately predicted to have a high PM score using DCE features, but more accurately assigned with a low PM score when using T2w features, providing an example where combining features from mpMRI sequences would be beneficial.

Figure 2.9 and Table 2.3 present the classification performances of the five classification models trained on the full dataset of 852 lesions imaged using either two- or three-sequence mpMRI protocols. Table 2.4 summarizes the p-values and the 95% CIs for the comparisons between the multiparametric and single-sequence classifiers' AUCs. Both mpMRI classification approaches significantly outperformed all single-sequence classifiers.

When only including the subset imaged using the three-sequence protocol and discarding the subset in which DWI was missing, the feature fusion and classifier fusion mpMRI approaches yielded AUCs [95% CIs] of 0.80 [0.73, 0.85] and 0.80 [0.74, 0.86], respectively, both significantly lower than their corresponding classifiers' performances when the full set was used (95% CI of $\Delta\text{AUC} = [0.01, 0.14]$ for both approaches). The results demonstrated that with the proposed method for handling exams acquired using different imaging protocols that contained inconsistent sequences, it would be beneficial to utilize the full dataset

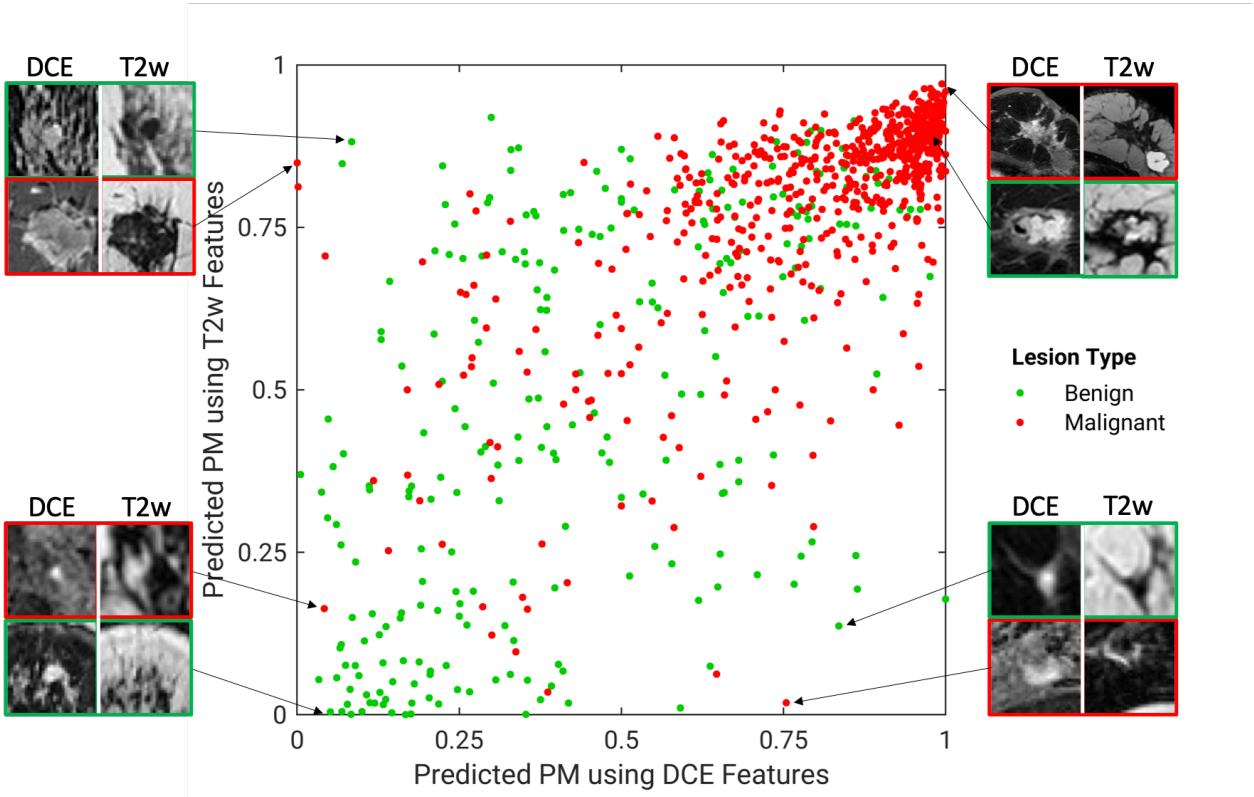


Figure 2.7: Diagonal classifier agreement plot between the T2-weighted (T2w) and dynamic contrast-enhanced (DCE) single-sequence classifiers trained on human-engineered radiomic features [14]. The x-axis and y-axis denote the probability of malignancy (PM) scores predicted by the classifiers using DCE and T2w features, respectively. Each point represents a lesion for which predictions were made. Points along or near the diagonal from bottom left to top right correspond to high classifier agreement; points far from the diagonal correspond to low agreement. Examples of lesions on which the two classifiers were in extreme agreement/disagreement are also included. Disagreement: lower right benign: papilloma; lower right malignant: IDC/DCIS, HER-2 enriched; upper left benign: fibroadenoma; upper left malignant: IDC/DCIS, luminal A. Agreement (both incorrect): upper right benign: hyalinized stromal fibrosis; lower left malignant: ductal carcinoma in situ. Agreement (both correct): upper right malignant: IDC/DCIS, triple negative, very large; lower left benign: fibroadenoma.

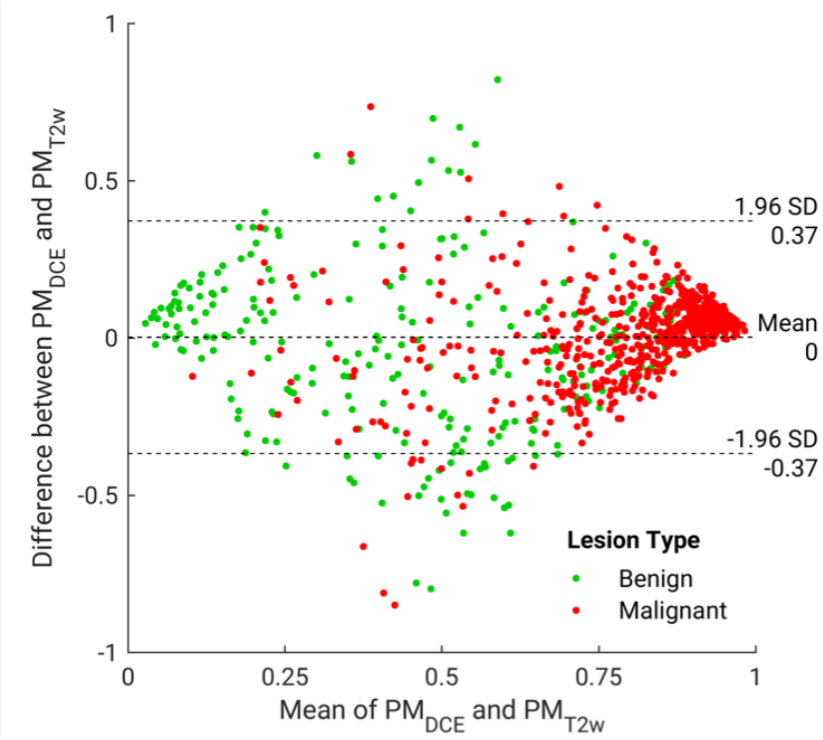


Figure 2.8: Bland-Altman plot illustrating classifier agreement between the single-sequence classifiers trained on human-engineered dynamic contrast-enhanced (DCE) features and T2-weighted (T2w) features [14]. The y-axis shows the difference between the support vector machine output scores of the two classifiers; the x-axis shows the mean of two classifiers' outputs.

Table 2.3: Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) along with the 95% confidence interval (CI) of AUC for each classifier based on human-engineered radiomic features, trained on the full set [14]. Sensitivity and specificity presented are for the optimal operating point determined using a metric for cut-off value that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$. Because all lesions were referred for biopsy, the sensitivity and specificity of the data set were not calculated for clinical assessment.

Classifier	DCE	T2w	DWI	Feature fusion	Classifier fusion
AUC	0.84	0.83	0.69	0.87	0.87
[95%CI]	[0.82, 0.87]	[0.80, 0.86]	[0.62, 0.75]	[0.84, 0.89]	[0.84, 0.89]
Sensitivity (%)	75.7	76.3	61.4	79.1	79.0
Specificity (%)	76.3	74.5	62.9	77.2	78.4

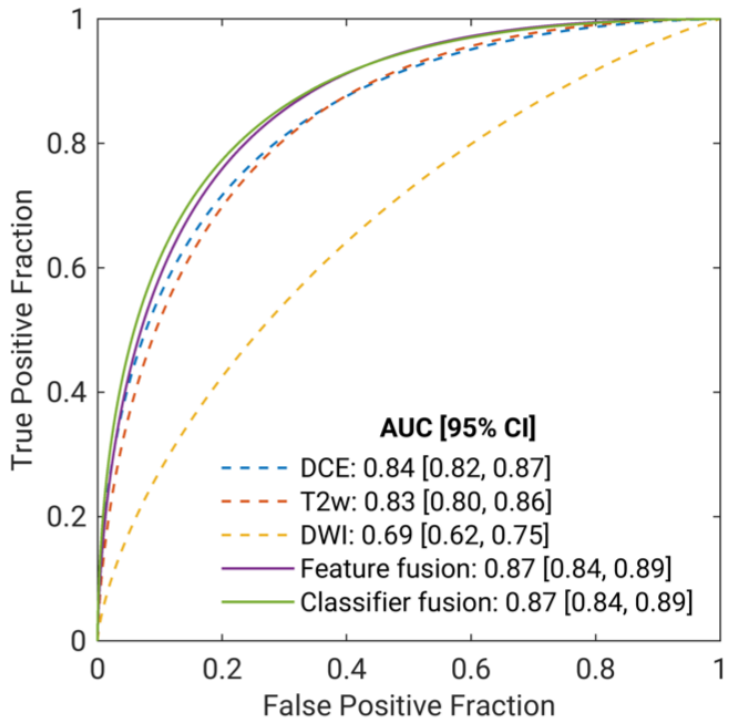


Figure 2.9: Fitted binormal receiver operating characteristic (ROC) curves for single-sequence (dashed line) and multiparametric MRI (mpMRI) classifiers (solid line) based on human-engineered radiomic features, trained on the full set [14]. The three single-sequence classifiers were trained separately on (i) dynamic contrast-enhanced (DCE), (ii) T2-weighted (T2w), and (iii) diffusion-weighted imaging (DWI) features. The mpMRI models (iv) were trained on the ensemble of features extracted from all available sequences, and (v) aggregated the probabilities of malignancy from the single-sequence classifiers via soft voting. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier.

Table 2.4: Performance comparison for the five classification methods based on human-engineered radiomic features, when classifiers were trained on the full set [14]. The classifier names are shown in the first column (single-sequence) and first row (multiparametric). P-value and 95% confidence interval (CI) of the difference in the area under the receiver operating characteristic curves (AUCs) are presented for each multiparametric classifier compared with each single-sequence classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections.

Classifier	Compared with feature fusion	Compared with classifier fusion
DCE	$P = .001^*$ 95% CI ΔAUC $= [0.01, 0.03]$	$P < .001^*$ 95% CI ΔAUC $= [0.01, 0.04]$
T2w	$P = .004^*$ 95% CI ΔAUC $= [0.01, 0.06]$	$P < .001^*$ 95% CI ΔAUC $= [0.02, 0.05]$
DWI	$P < .001^*$ 95% CI ΔAUC $= [0.11, 0.25]$	$P < .001^*$ 95% CI ΔAUC $= [0.11, 0.26]$

despite its incompleteness.

Deep-Learning-Based CADx

Figures 2.10 and 2.11 illustrate the comparison between the PMs predicted by the single-sequence classifiers using DCE and T2w. Figure 2.10 also shows example lesions on which these two classifiers agree or disagree. While the majority of benign and malignant lesions are separated from the other class, there appears to be moderate disagreement between the two classifiers, suggesting that a fusion technique could likely improve the predictive performance.

Figure 2.12 presents the ROC curves for the five classification schemes without image registration, and Table 2.5 summarizes the classification performances as measured by AUC, sensitivity, specificity, PPV, and NPV. Note that the mpMRI classifiers achieved improvements in terms of all these metrics for classification performance. Table 2.6 shows the p-values and the 95% CIs for the comparisons of AUCs between the mpMRI and single-sequence classifiers. Among the three mpMRI classification schemes, while all of them yielded statistically significantly higher AUCs than using T2w alone, only the feature fusion method significantly

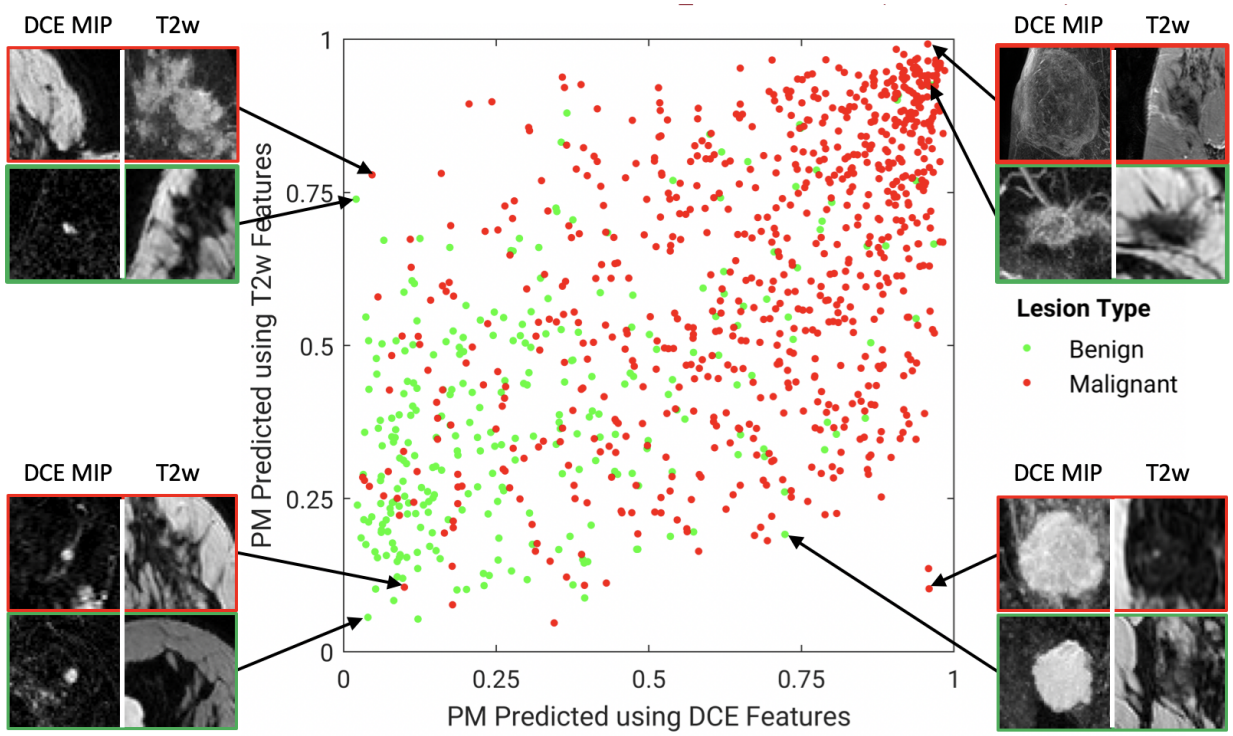


Figure 2.10: A diagonal classifier agreement plot between the T2-weighted (T2w) and dynamic contrast-enhanced (DCE) single-sequence classifiers trained on features extracted using VGG-19 [13]. The x-axis and y-axis denote the probability of malignancy (PM) scores predicted by the DCE classifier and the T2w classifier, respectively. Each point represents a lesion for which predictions were made. Points along or near the diagonal from bottom left to top right indicate high classifier agreement; points far from the diagonal indicate low agreement. Examples of lesions on which the two classifiers were in extreme agreement/disagreement are also included. Disagreement: lower right benign: fibroadenoma; lower right malignant: IDC/DCIS; upper left benign: unknown; upper left malignant: IDC/DCIS. Agreement (both incorrect): upper right benign: fat necrosis; lower left malignant: IDC.

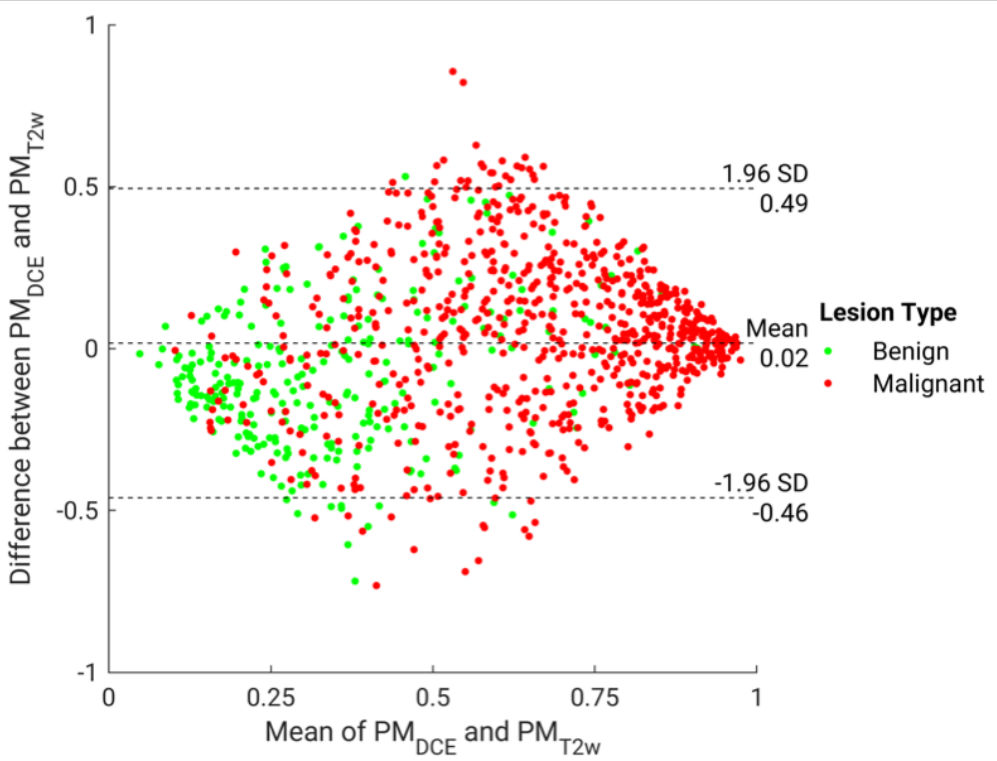


Figure 2.11: Bland-Altman plot illustrating classifier agreement between the dynamic contrast-enhanced (DCE) maximum intensity projection and T2-weighted (T2w)-based single-sequence classifiers trained on features extracted using VGG-19 [13]. The y-axis shows the difference between the support vector machine output scores (predicted posterior probabilities of malignancy) of the two classifiers; the x-axis shows the mean of two classifiers' outputs, which is also the probability of malignancy scores calculated in the classifier fusion method.

outperformed using DCE alone in terms of AUC, and the other two methods, image fusion and classifier fusion, failed to demonstrate a statistically significant difference in AUCs compared with using DCE alone.

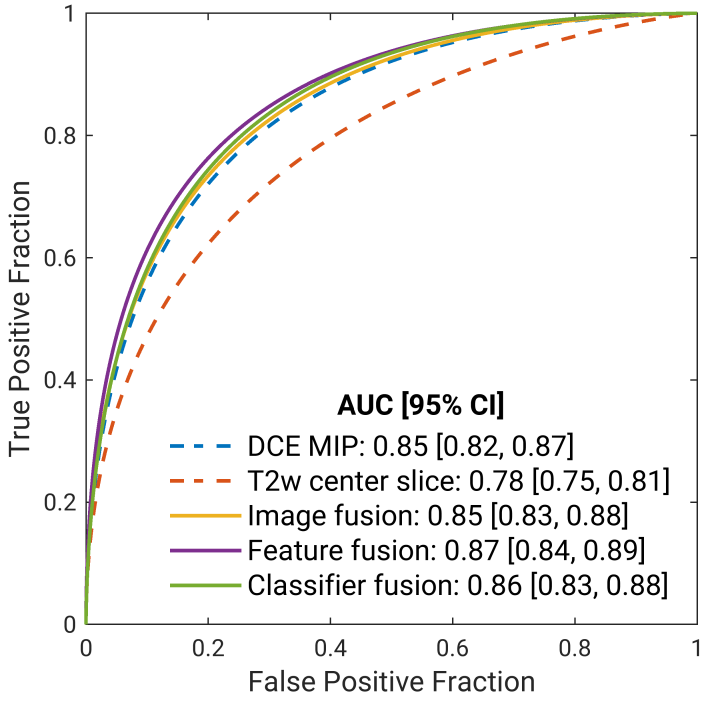


Figure 2.12: Fitted binormal receiver operating characteristic (ROC) curves for two single-sequence and three mpMRI classifiers trained on features extracted using VGG-19 [13]. The classifiers used (i) convolutional neural network (CNN) features extracted from dynamic contrast-enhanced (DCE) subtraction maximum intensity projections (MIPs), (ii) CNN features extracted from T2-weighted (T2w) center slices, (iii) CNN features extracted from DCE and T2w fusion images, (iv) ensemble of features extracted from DCE and T2w images, and (v) probability of malignancy outputs from the DCE MIP and T2w classifiers aggregated via soft voting. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier scheme. T2w images were rescaled to match the in-plane resolution of their corresponding DCE sequences, but image registration was not performed.

In assessing performance reproducibility, the mean and standard error of AUC for the feature fusion classifier was 0.864 ± 0.003, indicating that the classification performance was very stable regardless of the random seed chosen.

Performing inter-sequence rigid image registration did not have a significant effect on

Table 2.5: Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic curve (AUC) along with the 95% confidence interval (CI) of AUC for each classifier trained on features extracted using VGG-19 [13]. Sensitivity, specificity, PPV, and NPV presented are for the optimal operating point determined using a metric for a cut-off value that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$. Because all lesions were referred for biopsy, the sensitivity and specificity of the data set were not calculated for clinical assessment.

Classifier	DCE	T2w	Image fusion	Feature fusion	Classifier fusion
AUC	0.85	0.78	0.85	0.87	0.86
[95%CI]	[0.82, 0.88]	[0.75, 0.81]	[0.82, 0.88]	[0.84, 0.89]	[0.83, 0.88]
Sensitivity (%)	75.9	69.8	76.5	77.9	77.6
Specificity (%)	76.5	72.7	77.1	78.5	77.1
PPV (%)	89.7	87.3	90.0	90.7	90.1
NPV (%)	54.2	47.3	55.0	56.9	56.2

Table 2.6: Performance comparison for the five classification methods based on features extracted using VGG-19 [13]. The classifier names are shown in the first row (single-sequence) and first column (multiparametric). P-value and 95% confidence interval (CI) of the difference in area under the receiver operating characteristic curves (AUCs) are presented for each multiparametric classifier compared with each single-sequence classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections.

Classifier	Compared with DCE MIP	Compared with T2w center slice
Image fusion	$P = .73$ 95% CI ΔAUC $= [-0.01, 0.02]$	$P < .001^*$ 95% CI ΔAUC $= [0.05, 0.09]$
Feature fusion	$P < .001^*$ 95% CI ΔAUC $= [0.01, 0.03]$	$P < .001^*$ 95% CI ΔAUC $= [0.06, 0.11]$
Classifier fusion	$P = .14$ 95% CI ΔAUC $= [-0.00, 0.02]$	$P < .001^*$ 95% CI ΔAUC $= [0.06, 0.09]$

the classification performances of any classification scheme. Namely, the four classifiers affected by the registration (i.e., use information from T2w images) yielded AUC values of $\text{AUC}_{\text{T2w}} = 0.79 \pm 0.02$ (95% CI: [0.76, 0.82]), $\text{AUC}_{\text{ImageFusion}} = 0.84 \pm 0.01$ (95% CI: [0.81, 0.87]), $\text{AUC}_{\text{FeatureFusion}} = 0.87 \pm 0.01$ (95% CI: [0.84, 0.89]), and $\text{AUC}_{\text{ClassifierFusion}} = 0.86 \pm 0.01$ (95% CI: [0.83, 0.88]). Just as when T2w was not registered to DCE, while all three mpMRI classification schemes significantly outperformed using T2w alone, only feature fusion significantly outperformed using DCE alone. According to the 95% CIs of the difference in AUCs (ΔAUCs) between performing inter-sequence image registration or not, image registration between T2w and DCE failed to show a statistically significant effect on the performance of any classifiers examined. In addition, equivalence testing demonstrated that whether image registration was performed or not yielded equivalent performance with an equivalence margin of $\Delta\text{AUC} = 0.05$, chosen prima facie. Thus, all findings held regardless of whether image registration was employed or not, indicating that registration did not lead to a change in the performance of the mpMRI schemes.

2.5.2 ResNet Feature Extraction for mpMRI

The classifier using last-layer features and pooled features extracted from DCE MIPs yielded AUC values [95% CI] of 0.84 [0.82, 0.87] and 0.86 [0.83, 0.89], respectively. Pooling features extracted from multiple layers of ResNet statistically significantly improved the classification performance compared with using only the last-layer features ($P = .002$, 95% CI of ΔAUC: [0.01, 0.03]). This finding showed the advantage of utilizing mid- and low-level features learned by the network during training, even with the presence of skip connections in ResNet.

Figure 2.13 and Table 2.7 present the classification performances of the five classifiers when using pooled features. Table 2.8 summarizes the p-values and the 95% CIs for the comparisons of performances of the multiparametric and single-sequence classifiers. Among the multiparametric methods, the feature fusion method statistically significantly outper-

formed the classifier that used DCE alone, and all three methods statistically significantly outperformed using T2w alone.

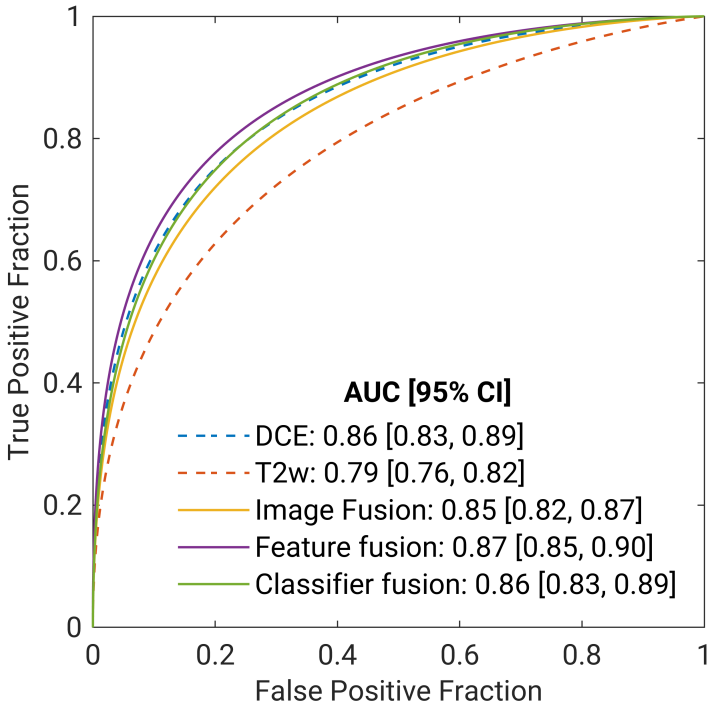


Figure 2.13: Fitted binormal receiver operating characteristic (ROC) curves for two single-sequence and three mpMRI classifiers trained on features extracted using ResNet-50 [15]. The classifiers used (i) convolutional neural network (CNN) features extracted from dynamic contrast-enhanced (DCE) subtraction maximum intensity projections (MIPs), (ii) CNN features extracted from T2-weighted (T2w) center slices, (iii) CNN features extracted from DCE and T2w fusion images, (iv) ensemble of features extracted from DCE and T2w images, and (v) probability of malignancy outputs from the DCE MIP and T2w classifiers aggregated via soft voting. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier scheme. T2w images were rescaled to match the in-plane resolution of their corresponding DCE sequences, but image registration was not performed.

2.5.3 Human-Engineered Radiomics versus Deep Learning for DWI

Figure 2.14 demonstrates the performances of the ADC-based and the CNN-based classifiers individually and the fusion classifier. They achieved AUC values [95% CI] of 0.68 [0.61, 0.75],

Table 2.7: Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) along with the 95% confidence interval (CI) of AUC for each classifier trained on features extracted using ResNet-50 [15]. Sensitivity and specificity presented are for the optimal operating point determined using a metric for a cut-off value that minimizes $m = (1 - sensitivity)^2 + (1 - specificity)^2$. Because all lesions were referred for biopsy, the sensitivity and specificity of the data set were not calculated for clinical assessment.

Classifier	DCE	T2w	Image fusion	Feature fusion	Classifier fusion
AUC	0.86	0.79	0.85	0.87	0.86
[95%CI]	[0.83, 0.89]	[0.76, 0.82]	[0.82, 0.87]	[0.85, 0.90]	[0.83, 0.89]
Sensitivity (%)	76.6	69.4	75.2	77.8	77.1
Specificity (%)	78.4	73.4	76.9	79.8	77.7

Table 2.8: Performance comparison for the five classification methods based on features extracted using ResNet-50 [15]. The classifier names are shown in the first row (single-sequence) and first column (multiparametric). P-value and 95% confidence interval (CI) of the difference in the area under the receiver operating characteristic curves (AUCs) are presented for each multiparametric classifier compared with each single-sequence classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections.

Classifier	Compared with DCE MIP	Compared with T2w center slice
Image fusion	$P = .05$ 95% CI ΔAUC $= [-0.03, 0.00]$	$P < .001^*$ 95% CI ΔAUC $= [0.03, 0.08]$
Feature fusion	$P = .004^*$ 95% CI ΔAUC $= [0.01, 0.02]$	$P < .001^*$ 95% CI ΔAUC $= [0.06, 0.10]$
Classifier fusion	$P = .96$ 95% CI ΔAUC $= [-0.01, 0.01]$	$P < .001^*$ 95% CI ΔAUC $= [0.05, 0.09]$

0.74 [0.68, 0.80], and 0.76 [0.69, 0.82], respectively. The fusion classifier performed significantly better than the ADC-based classifier ($P = 0.01$), whereas the CNN-based classifier failed to show a statistically significant difference from the other two.

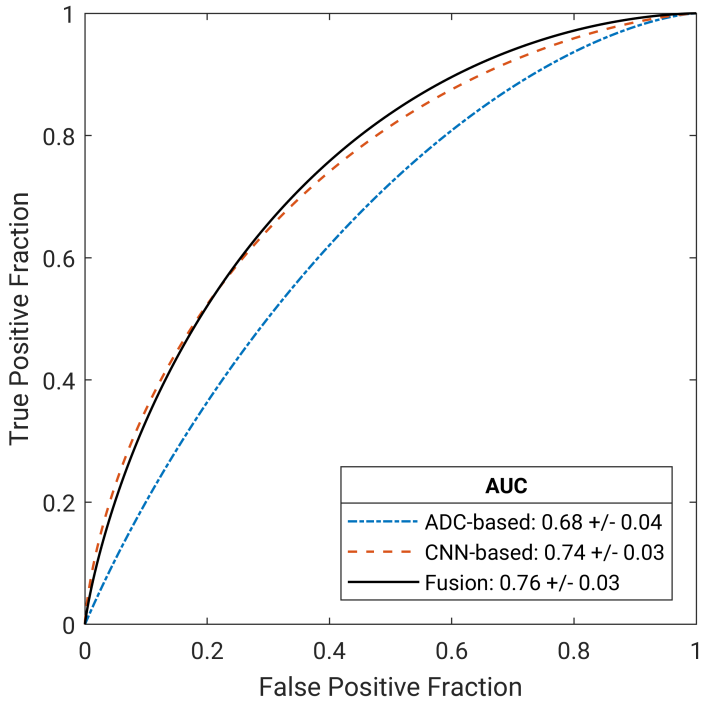


Figure 2.14: Fitted binormal receiver operating characteristic (ROC) curves for the ADC-based classifier, the CNN-based classifier, and the fusion classifier for DWI [16].

2.6 Discussion and Conclusions

The work presented in this chapter investigated radiomics methods that leverage the complementary information provided by the DCE, T2w, and DWI sequences in mpMRI and demonstrated the potential to improve performance over single-sequence radiomics methods in the task of distinguishing between benign and malignant breast lesions [13–16]. The study was performed on both human-engineered radiomic features and features extracted by pre-trained CNN models. Three mpMRI fusion approaches were proposed and evaluated: image fusion, i.e., fusing images from multiple MRI sequences into an RGB image to form the

input to the CNN (for CNN-based methods only); feature fusion, i.e., concatenating features extracted from mpMRI sequences to form the classifier input; and classifier fusion, i.e., aggregating the probability of malignancy output scores from single-sequence classifiers via soft voting. When human-engineered features were used, both feature fusion and classifier fusion methods achieved significantly higher classification performance using any sequence alone. When CNN features were used, the feature fusion method significantly outperformed using the DCE sequence alone, and all fusion methods significantly outperformed using the T2w sequence alone. The findings in this work can potentially improve the currently available breast cancer CADx systems based on DCE-MRI.

Using nested cross-validation as the evaluation technique allowed for more efficient use of the limited data by reporting an overall score across five test sets; however, methods should ideally be evaluated on an independent held-out test set to demonstrate the algorithm's generalizability and robustness. We will perform such evaluation in later chapters when larger datasets are available.

Moreover, only six first-order human-engineered radiomic features were extracted from ADC maps. Other radiomic features were not calculated because the coarse resolution of DWI limited the utility of high-order features, such as texture features. Also, feature selection was not included in the human-engineered radiomics study, because our preliminary investigation of several feature selection and dimension reduction methods, including stepwise feature selection, recursive feature selection, principal component analysis, and t-distributed stochastic neighbor embedding, showed that none of these methods resulted in improved classification performance. It is worth noting that our approach was to extract radiomic features that are clinically or physiologically relevant to the diagnosis of breast cancer, rather than extracting as many features as possible and then selecting a subset based on statistical importance. A total of 75 features were extracted from three modalities, which was a reasonable size for SVM classifiers without feature selection, especially given the fairly large

size of the database.

In addition, MRI exams used in this study were collected over the span of six years, during which imaging technology advanced and some acquisition parameters did not remain constant. We ensured that no severe imbalance that would potentially bias the results was present. For example, the field strengths distribution was similar between the benign and malignant class: among the 195 benign lesions, 141 (72%) of them were imaged with 1.5 T scanners and 54 were imaged with 3T (28%) scanners; among the 657 malignant lesions, 422 (64%) were imaged with 1.5 T scanners and 235 were imaged with 3T (36%) scanners. The spatial resolution and the temporal resolution in the DCE sequence were also variable within the dataset. The use of such a retrospectively collected, heterogeneous dataset positively contributed to the algorithm robustness and generalizability. Future work will continue investigating the harmonization of differences in acquisition parameters to improve performance.

Common alternative approaches for handling missing modalities in multiparametric imaging studies include image imputation and feature imputation. Image imputation methods are task-specific, and while developing a satisfactory image imputation method for diagnosing breast cancer on mpMRI is an interesting topic for future investigation, it is beyond the scope of this study. As for feature imputation, a comparative experiment was performed in which the missing DWI radiomic features were imputed using a regression-based multivariate iterative feature imputation method, and the classification results were compared with those from our original approach. The performance for all classifiers that utilized DWI features, namely the DWI single-parametric classifier, the feature fusion mpMRI classifier, and the classifier fusion mpMRI classifier, slightly decreased. Their AUC values [95% CIs] were 0.66 [0.62, 0.70], 0.85 [0.82, 0.88], and 0.86 [0.84, 0.89], indicating that the imputed DWI features did not benefit the classification performance. Besides classification performance, the advantages of our original approach also include its computational efficiency as it eliminates

the imputation step, and its close analogy to the clinical diagnostic process, i.e., radiologists basing their assessment on either two or three sequences available in the mpMRI exam for a particular case.

In this study, the CNN models were used as feature extracted with fixed pretrained weights. Preliminary investigation suggested that given the task in question and the characteristics of the dataset, feature extraction was the most appropriate use of the CNN models in this particular study. In a different scenario, such as in later chapters in this dissertation, fine-tuning a part or all of the CNN model may be beneficial. Moreover, the pretrained CNN network requires 2D input, which limited the inclusion of the high-dimensional information contained in breast MRI exams. The 4D information in DCE-MRI was captured by using second post-contrast MIP images in this study, and Chapter 3 will investigate approaches to more effectively leverage high-dimensional information in medical images in deep transfer learning frameworks.

When performing inter-sequence image registration, multi-modality rigid registration that consists of translation and rotation was performed. Scaling and shifting were performed based on acquisition parameters provided in the exams' DICOM information. Shearing or deformable registration was not employed because it was not desirable for the quantitative image analysis in this study to alter the geometry of and the texture within the lesions. More in-depth registration optimizations can be explored in future studies. Image registration can be computationally expensive and time-consuming. Given that all classifier performances were equivalent with or without image registration, image registration might not be a necessary step in this proposed method of distinguishing between benign and malignant breast lesions using mpMRI.

The margin in equivalence testing is ideally a predetermined clinically meaningful limit. However, due to complexities and impracticalities in applying the statistical principles of equivalence testing to diagnostic performance studies, there is currently no widely used

standard procedure to establish this margin [123]. Nonetheless, a rather conservative margin of 5% for ΔAUC was used in this study to demonstrate equivalence between classifier pairs.

CHAPTER 3

HIGH-DIMENSIONAL DEEP LEARNING IMAGE ANALYSIS FOR BREAST CANCER DIAGNOSIS ON MULTIPARAMETRIC MRI

3.1 Introduction

Deep learning methods have demonstrated success in computer-aided medical imaging analysis, as mentioned in Chapter 1, where transfer learning techniques are usually employed to circumvent the need for massive datasets [31, 36]. Standard transfer learning techniques for convolutional neural networks (CNNs) have achieved promising results for breast MRI analysis [13, 15, 38, 104, 113]. However, CNNs pretrained on two-dimensional (2D), natural images in ImageNet require 2D inputs, which has resulted in an underutilization of the high-dimensional information in MRI that critically contributes to lesion classification in clinical practice.

High dimensionality and data scarcity are unique challenges in deep learning applications for medical imaging. In order to exploit the rich, clinically valuable information inherent in medical images without sacrificing computational efficiency or model performance, it is important to devise methods to use transfer learning in creative ways so that volumetric, temporal, and other aspects of high-dimensional images can be incorporated even when networks pretrained on 2D images are used. To take advantage of the four-dimensional (4D) (volumetric and temporal) information inherent in DCE-MRI without sacrificing the efficiency provided by transfer learning, a previously proposed method, which was shown to outperform methods using only 2D or 3D information, used the second post-contrast subtraction maximum intensity projection (MIP) images to classify breast lesions as benign or malignant [105]. This chapter proposes a new transfer learning method that makes use of both volumetric and temporal information in DCE-MRI more effectively than MIP images.

Instead of collapsing the volumetric information at the image level to form MIP images, the dimension reduction occurs at the feature level by taking the maximum of CNN features along the axial dimension for a given lesion within the CNN, which will be referred to as "feature MIP." Additionally, the use of subtraction images was replaced by inputting images acquired at three of the dynamic time points to the three channels of the CNN. The feature MIP method was then validated on an external dataset with the temporal information from four dynamic time points incorporated in the form of three post-contrast subtraction images inputted to the three channels of the CNN.

Furthermore, the study further extended to multiparametric MRI (mpMRI). First, two CNN transfer learning approaches to utilizing the diffusion weighting information in the diffusion-weighted image (DWI) sequence were investigated. Then, the feature MIP method was applied to three sequences in mpMRI, namely, DCE, T2-weighted (T2w), and DWI sequences, and the feature fusion method discussed in Chapter 2 was applied to integrate the multiparametric information.

3.2 Datasets

3.2.1 UChicago Medicine DCE-MRI Dataset

The first dataset involved in this study, specifically for investigation on leveraging 4D information in DCE-MRI, was derived from the same retrospective database detailed in Section 2.2. Due to the addition of recent cases and not requiring all MRI exams to be multiparametric, the DCE-MRI dataset used in this study included more cases than that in Section 2.2. In total, the dataset consisted of 1161 unique breast lesions from 855 women who had undergone breast MR exams. Of all lesions, 270 were benign (23%) and 891 were malignant (77%). Patient age ranged from 23 to 89 years old, with a mean and standard deviation of 55 ± 13 years. Images in the database were acquired over the span of 12 years, from

2005 to 2017, using either 1.5 T or 3 T Philips Achieva scanners with a T1-weighted spoiled gradient sequence. Image in-plane resolution ranged from 0.55 mm to 1.37 mm, and image slice thickness varied 1.6 mm to 2.5 mm. The other aspects of the dataset, including Institutional Review Board protocol, image acquisition, and ground truth validation, follow the description in Section 2.2.

3.2.2 Tianjin Medical University MRI Dataset

Another breast MRI dataset was collected from the Tianjin Medical University Cancer Institute and Hospital, which allowed for independent, external validation of the methods developed on the UChicago Medicine database. The dataset used in this study was retrospectively collected and de-identified prior to analysis, and thus the study was deemed exempt by the institutional review board-approved protocol. Initially, the 4704 patients presenting for breast MRI examinations between 2015 and 2017 were consecutively collected. Exclusion criteria included patients with previous surgical excision, systemic hormone therapy or chemotherapy, exams that did not exhibit a visible lesion, and lesions without final pathology results. A total of 1990 unique lesions from 1979 patients imaged by DCE-MRI were ultimately included in this study. There were 1494 (75%) malignant lesions from 1483 patients with cancer, including eight bilateral and three bifocal cancers, and 496 (25%) benign lesions from 496 benign patients. A subset of 1827 unique lesions from 1825 women was imaged with mpMRI protocols that contained DCE, T2w, and DWI sequences, among which 1372 (75%) were malignant lesions from 1370 patients, including one bilateral and one bifocal cancer patient, and 455 (25%) were benign from 455 patients. Ground truth for each lesion was based on histopathology from surgical specimens. The flowchart in Fig. 3.1 illustrates patient enrollment and inclusion.

MR images were acquired with 3 T GE scanners using a dedicated eight-channel phased-array breast coil (Discovery 750, GE Medical Systems, Milwaukee, WI). The MRI protocol

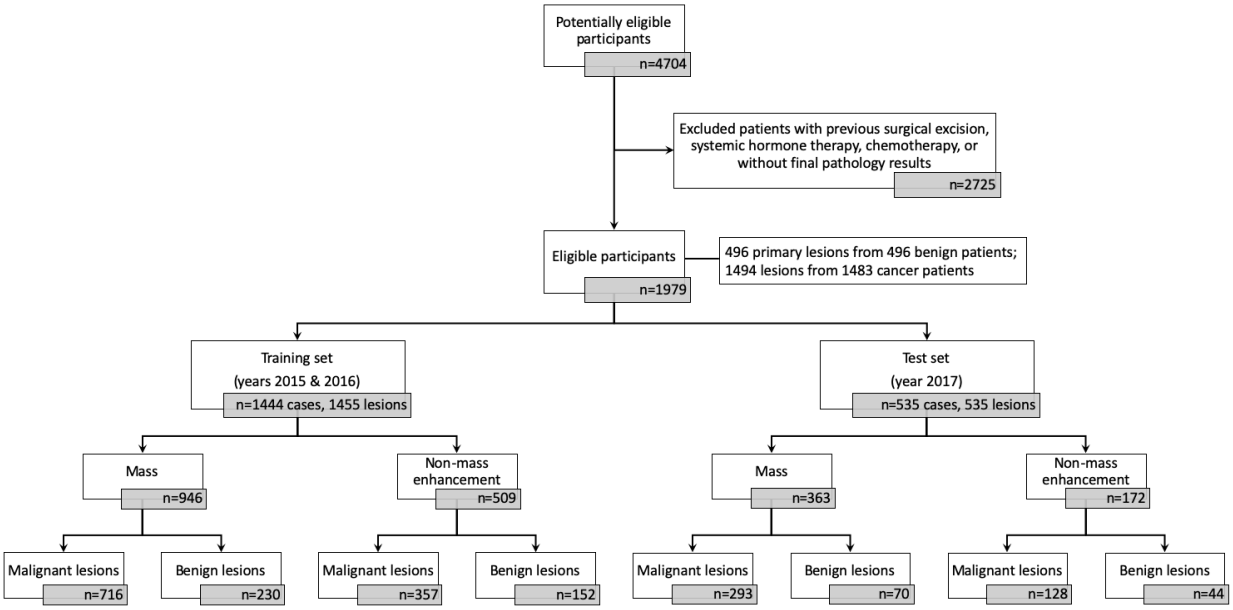


Figure 3.1: Flowchart of study participants enrollment [17].

consisted of a sagittal T1-weighted DCE gradient echo sequence using the volume imaging for breast assessment (VIBRANT) bilateral breast imaging technique, a sagittal T2w fast spin echo sequence with flow compression and fat saturation, and an axial diffusion-weighted echo planar imaging sequence. The average spatial resolution is 0.5 mm $\times$ 0.5 mm $\times$ 1.8 mm for DCE, 0.5 mm $\times$ 0.5 mm $\times$ 4.0 mm for T2w, and 1.2 mm $\times$ 1.2 mm $\times$ 4.5 mm for DWI. The temporal resolution for dynamic acquisition ranged from 51 s to 129 s. The contrast agent, gadolinium-diethylenetriamine pentaacetic acid (0.1 mmol/kg body weight, flow rate 2.0 ml/s), was injected after the serial mask images were obtained, followed by flushing with the same total dose of saline solution. All DWI sequences were acquired using at least three common b-values (0, 500, and 1000 s/mm^2), with up to nine additional b-values in some exams.

To minimize the bias in case selection for the computerized image analysis and to mimic a development-then-clinical-use scenario, the dataset was divided into a development set

(training and validation) and an independent test set based solely on the date of the MRI examinations. The development dataset included 1455 lesions (1323 imaged with mpMRI) from years 2015 and 2016, and the test set included 535 lesions (504 imaged with mpMRI) from the year 2017. No patients were in both the development set and the test set, and there was one lesion per patient in the test set. The clinical characteristics of the study population are listed in Table 3.1. Lesion characteristics were similar in the training and validation data compared with the test data for lesion size for benign ($P = .29$) and malignant ($P = .09$) lesions. Similar distributions were noted in other subcategories as well.

Table 3.1: Clinicopathological characteristics of the lesions from patients in the Tianjin breast MRI dataset [17].

	Training and Validation		Test	
	Malignant	Benign	Malignant	Benign
Total	1073	382	421	114
Age[a,h], years	47.6 (19-77)	42.2 (16-76)	49.3 (25-75)	41.9 (19-65)
Size[b], mm	19.1 ± 8.6	14.7 ± 10.7	18.5 ± 7.6	12.9 ± 6.8
Lesion type				
Mass	716 (75.7%)	230 (24.3%)	293 (80.7%)	70 (19.3%)
Non-mass	357 (70.%)	152 (29.9%)	128 (74.4%)	44 (25.6%)
MRI BI-RADS category[c,h]				
0	0 (0%)	2 (0.5%)	0 (0%)	0 (0%)
1	0 (0%)	1 (0.3%)	0 (0%)	2 (1.8%)
2	0 (0%)	4 (1.0%)	0 (0%)	0 (0%)
3	4 (0.3%)	202 (52.9%)	0 (0%)	50 (43.8%)
4	351 (33.1%)	170 (44.5%)	113 (26.8%)	60 (52.6%)
5	529 (49.8%)	3 (0.8%)	221 (52.5%)	2 (1.8%)

Table 3.1: Clinical characteristics of the dataset (continued)

6	178 (16.8%)	0 (0%)	87 (20.7%)	0 (0%)
Histology				
IDC[d]	914 (85.2%)		366 (86.9%)	
ILC[e]	22 (2.1%)		4 (1.0%)	
DCIS[f]	76 (7.1%)		18 (4.3%)	
Other malignant	61 (5.6%)		33 (7.8%)	
Fibroadenoma		165 (43.2%)		46 (40.4%)
Papilloma		66 (17.3%)		28 (24.6%)
Inflammation		19 (5.0%)		10 (8.8%)
Other benign		132 (34.5%)		30 (26.3%)
Estrogen receptor[i]				
$< 1\%$	192 (18.0%)		77 (18.3%)	
$>= 1\%$	876 (82.0%)		344 (81.7%)	
Progesterone receptor[i]				
$< 1\%$	222 (20.8%)		104 (24.7%)	
$>= 1\%$	846 (79.2%)		317 (75.3%)	
HER-2[g,i]				
0 or 1+	632 (59.2%)		243 (57.7%)	
2+ or 3+	436 (40.8%)		178 (42.3%)	
Ki-67[i]				
$< 14\%$	180 (16.9%)		60 (14.3%)	
$>= 14\%$	887 (83.1%)		361 (85.7%)	

Table 3.1: Clinical characteristics of the dataset (continued)

[a] Patient age is shown as mean (range).
[b] Lesion size is measured by the effective diameter, i.e., the greatest dimension of a sphere with the same volume as the lesion, and shown as mean ± standard deviation.
[c] BI-RADS = Breast Imaging Reporting and Data System
[d] IDC = invasive ductal carcinoma
[e] ILC = Invasive lobular carcinoma
[f] DCIS = Ductal carcinoma in situ
[g] HER-2 = human epidermal growth factor receptor 2
[h] Age and BI-RADS are reported by patient, and the other information is reported by lesion.
[i] There were five lesions with unknown estrogen receptor, progesterone receptor, and HER-2 status, and six lesions with unknown Ki-67 status.

3.3 Methods

3.3.1 Volumetric and Temporal Information in DCE-MRI

Input Construction

As illustrated in Fig. 3.2, the subtraction images were created by subtracting the pre-contrast (t_0) images from their corresponding second post-contrast (t_2) images in order to emphasize the contrast enhancement pattern within the lesion and suppress constant background. To generate MIP images, the 3D volume of subtracted images for each lesion was then collapsed into a 2D image by selecting the voxel with the maximum intensity along the axial dimension, i.e., perpendicular to the transverse slices.

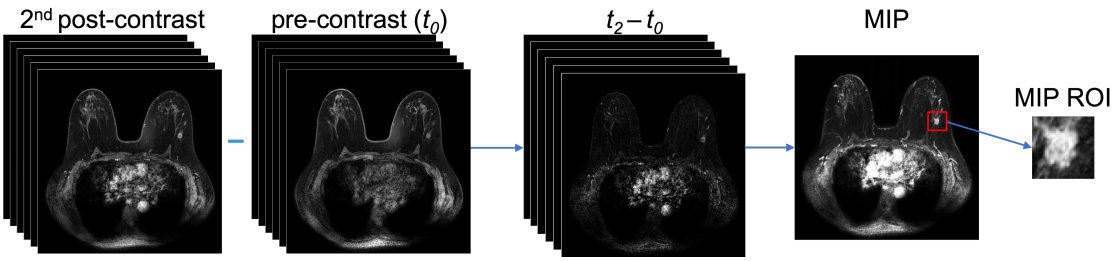


Figure 3.2: Illustration of the processes to construct the second post-contrast subtraction images, the subtraction maximum intensity projection (MIP) images, and region of interest (ROI) [18].

To avoid confounding contributions from distant voxels, a region of interest (ROI) around each lesion was automatically cropped from the image to use in the subsequent classification process. The ROI size was chosen based on the maximum dimension of each lesion, and a small part of the parenchyma around the lesion was included. The minimum ROI size was set to 32×32 pixels as required by the model architecture. The cropping process is illustrated in Figs. 3.2 and 3.3. ROIs were not rescaled.

Additionally, an ROI that contained red, green, and blue (RGB) channels was created for each slice of each lesion. As illustrated in Fig. 3.3, the pre-contrast (t_0), first post-contrast (t_1), and second post-contrast (t_2) DCE time points were input to the red, green, and blue channels, respectively.

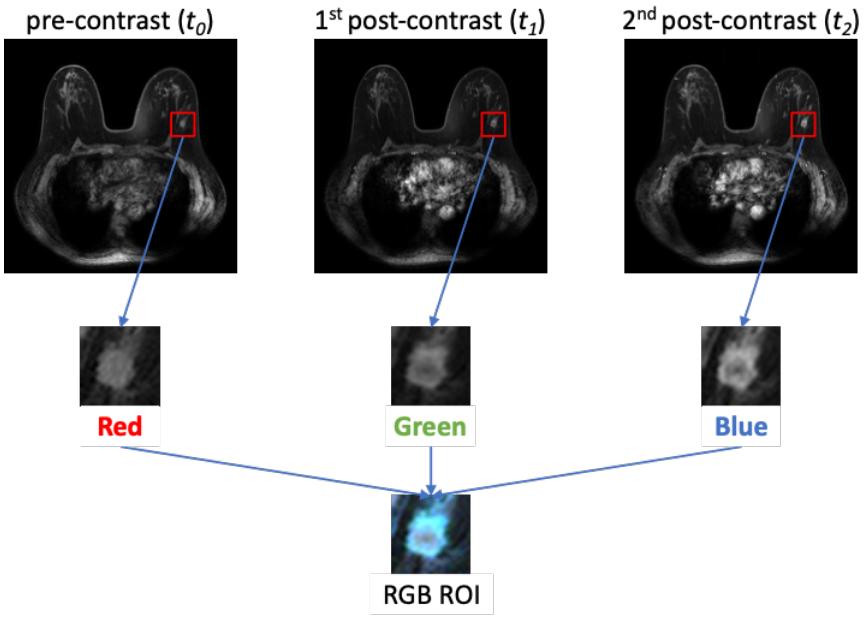


Figure 3.3: Illustration of the RGB region of interest (ROI) construction process.

Classification and Evaluation

Figure 3.4 schematically shows the transfer learning classification and evaluation process for the three methods following the ROI construction. For each lesion, CNN features were

extracted separately from the subtraction MIP ROI, the subtraction ROIs of all slices, and the RGB ROIs of all slices using a VGG-19 model pretrained on ImageNet [34, 93]. The RGB ROI volumes comply with the desired three-channel input format of VGG-19, while the subtraction ROI volumes and MIP ROIs were grayscale and were duplicated across the three channels. Feature vectors were extracted at various network depths from the five max pooling layers of the VGG-19. These features were then average-pooled along the spatial dimensions and normalized with Euclidian distance. The pooled features were then concatenated to form a CNN feature vector for a given lesion [104, 113].

For the method using subtraction or RGB ROIs of all slices of a lesion, the 2D feature vectors extracted by VGGNet from each slice were further concatenated to form a 3D feature vector, which was subsequently collapsed into a 2D feature vector by selecting the maximum feature value along the axial dimension (i.e., taking the MIP of the feature vector along the direction in which slices were stacked). This method will be referred to as "feature MIP." Max pooling was chosen over average pooling along the axial dimension because it was desirable to select the most prominent occurrence of each feature among all transverse slices of a lesion. Average pooling would have smoothed out the feature map and obscured the predictive features.

Three linear support vector machine (SVM) classifiers were trained on the CNN features extracted from subtraction MIPs, subtraction volumes, and RGB volumes separately to differentiate between benign and malignant lesions (Python Version 3.7.3, Python Software Foundation), following the method described in Section 2.3.2. Training and evaluation were performed using nested five-fold cross-validation, with the area under the receiver operating characteristic (ROC) curve (AUC) serving as the figure of merit, as detailed in 2.3.3 [117, 118].

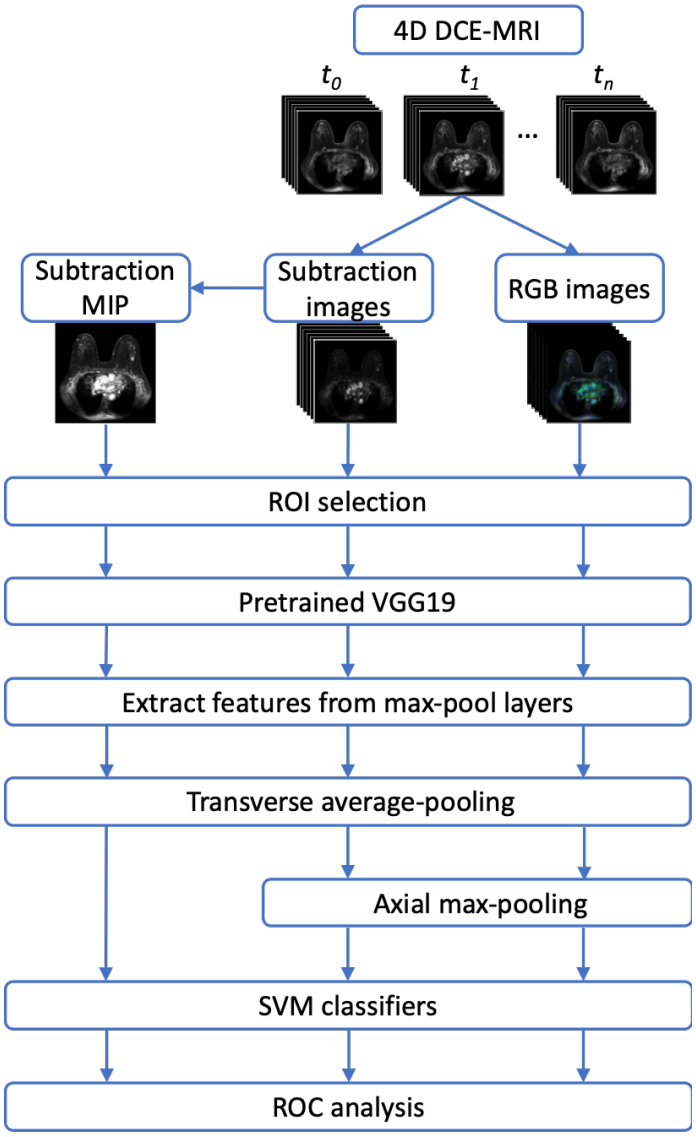


Figure 3.4: Lesion classification pipelines based on diagnostic images. Three-dimensional volumetric lesion information from dynamic contrast-enhanced (DCE)-MRI is collapsed into 2D by maximum intensity projection (MIP) at the image level (left) or at the feature level (middle and right) along the axial dimension. Temporal information is incorporated via either subtraction images (left and middle) or inputting different time points in a DCE sequence into the RGB channels of the CNN input (right).

3.3.2 Independent Validation of Feature MIP on DCE-MRI

The procedure for creating the input for the CNN architecture from the 4D DCE-MRI sequence is illustrated in Fig. 3.5. Subtraction images were created by subtracting the pre-contrast (t_0) images from their corresponding first, second, and third post-contrast images (t_1, t_2, and t_3, respectively) to emphasize the contrast enhancement pattern within the lesion and suppress constant background.

To avoid confounding contributions from distant voxels, an ROI around each lesion was automatically cropped from all of its subtraction images with a seed-point manually indicated by a breast radiologist (Dr. Yu Ji) with five years of experience in breast DCE-MRIs. The ROI cropping process follows the description in Section 3.3.1. For the feature MIP method, the 3D ROIs from the three subtraction image volumes (i.e., the first, second, and third post-contrast subtraction 3D ROIs) were input into the network through the RGB channels, respectively, forming a 3D RGB ROI for each lesion. The pixel intensity in each ROI was normalized over the 3D ROI volume. For the image MIP method, the 3D RGB ROI volume for each lesion was subsequently collapsed into a 2D MIP ROI by selecting the voxel with the maximum intensity along the axial dimension (i.e., perpendicular to the transverse slices).

Figure 3.5 also shows a schematic of the transfer learning classification and evaluation process for the two methods. For each lesion, CNN features were extracted from the inputted MIP RGB ROIs and the 3D RGB ROIs separately using a VGG-19 model pretrained on ImageNet [34, 93]. Cases from years 2015-2016 (1455 lesions) were split into 80% for training and 20% for validation, holding the class prevalence constant across the two sets and under the constraint that lesions from the same patient were kept together in the same set to eliminate the impact of bias from data leakage. Cases from the year 2017 (535 lesions) were held out for testing. The feature extraction, classification, and evaluation process for both image MIP and feature MIP methods followed the descriptions in Section 3.3.1, but on independent training, validation, and test sets instead of nested cross-validation.

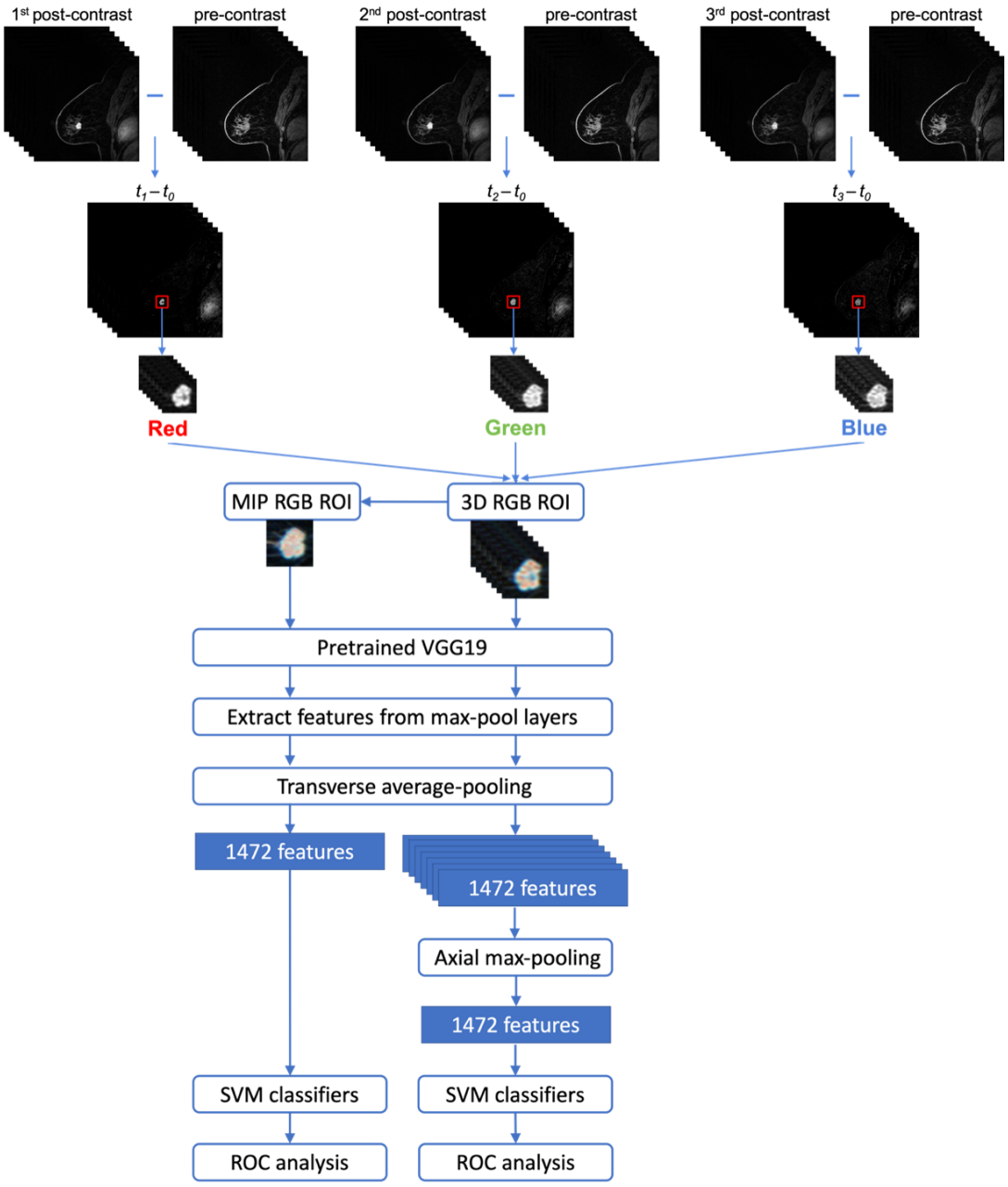


Figure 3.5: Lesion classification pipelines for image maximum intensity projection (MIP) and feature MIP [17]. The top portion illustrates the construction of the region of interest (ROI) that incorporates volumetric and temporal information from the four-dimensional dynamic contrast-enhanced MRI sequence. The same ROI was cropped from the first, second, and third post-contrast subtraction images and combined in the red, green, and blue (RGB) channels to form a three-dimensional (3D) RGB ROI. For image MIP (left branch of the bottom portion), the MIP RGB ROI was generated from the 3D RGB ROI, collapsing volumetric lesion information at the image level. For feature MIP (right branch of the bottom portion), volumetric lesion information was integrated at the feature level by max pooling feature extracted from all slices. SVM = support vector machine.

The optimal operating point reported for each classifier was determined, using the ROC curve of the training data, by finding the sensitivity and specificity pair that maximizes the function $sensitivity - m(1 - specificity)$, where m is the slope of the ROC curve at the optimal operating point given by

$$m = \frac{Prob_{Norm}}{Prob_{Dis}} \times \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}, \tag{3.1}$$

with $Prob_{Norm}$ and $Prob_{Dis}$ being the probability that a case from the population studied is negative and positive, respectively, and C_{FP}, C_{TN}, C_{FN}, C_{TP} being the cost of a false-positive, true-negative, false-negative, and true-positive result, respectively [124, 125]. An equal cost was assumed for false positive and false negative predictions and no cost for the correct predictions. The predicted posterior probabilities of malignancy (PMs) of the test set were converted to match the cancer prevalence in the training set [126], and the sensitivity and specificity of the test set were reported using the optimal thresholds pre-determined on the training data. The two classifiers' sensitivities and specificities were each compared at the optimal point using the McNemar test [127, 128]. $P < 0.05$ was considered to indicate a statistically significant difference in each performance metric. Statistical analyses were performed in MATLAB (MATLAB R2019b, The MathWorks Inc., Natick, Massachusetts).

3.3.3 Deep Learning for DWI

The role of deep learning in the diagnosis of breast cancer on DWI was investigated on the mpMRI subset in the database from Tianjing, which is detailed in Section 3.2.2. A DWI sequence acquired using at least three b-values (0, 500, and 1000 s/mm^2) was included in each exam. Lesions were automatically segmented from the DWI images using a fuzzy C-means method with a manually indicated seed-point. The slice with the largest segmented area was chosen to represent each lesion, and a square ROI was cropped around the lesion in the same

manner as described in Section 3.3.1. Two types of CNN input were investigated: a) the apparent diffusion coefficient (ADC) image derived pixel by pixel from the three b-values; b) an RGB image in which the red, green, and blue channels of the CNN contained DWI images at the three b-values, respectively. The CNN, a VGG-19 pretrained on ImageNet, was fine-tuned on 1323 lesions images in years 2015 and 2016 to distinguish between benign and malignant lesions and then independently tested on 504 lesions images in the year 2017 on the task of distinguishing between benign and malignant lesions. Classification performance was evaluated using ROC analysis, and the AUC served as the figure of merit. Statistical tests on superiority and noninferiority were performed to compare the two methods.

3.3.4 Feature MIP on Multiparametric MRI

As illustrated in Fig. 3.6, the feature MIP method detailed in Section 3.3.1 was applied to three sequences in mpMRI, and the feature fusion method investigated in Chapter 2 were employed to integrate multiparametric information. The input images for the DCE sequence and the DWI sequence were RGB images of the lesion ROI, following the construction processes described in Section 3.3.2 and 3.3.3, respectively. The input image for the T2w sequence was a grayscale volume of the lesion ROI. The modified VGG-19 model, as described in 2.3.2, was connected with a multilayer perceptron (MLP), as shown in Fig. 3.7, which was fine-tuned on each MRI sequence to differentiate benign and malignant breast lesions. Feature MIP was applied to the lesion volume in each of the three MRI sequences, and the features vectors from these sequences were subsequently concatenated and input into an MLP to produce a probability of malignancy output score. The evaluation for the single-sequence and mpMRI classifiers followed Section 3.3.2. Gradient-weighted class activation mapping (Grad-CAM) was generated to provide a visual explanation of the model's classification [129].

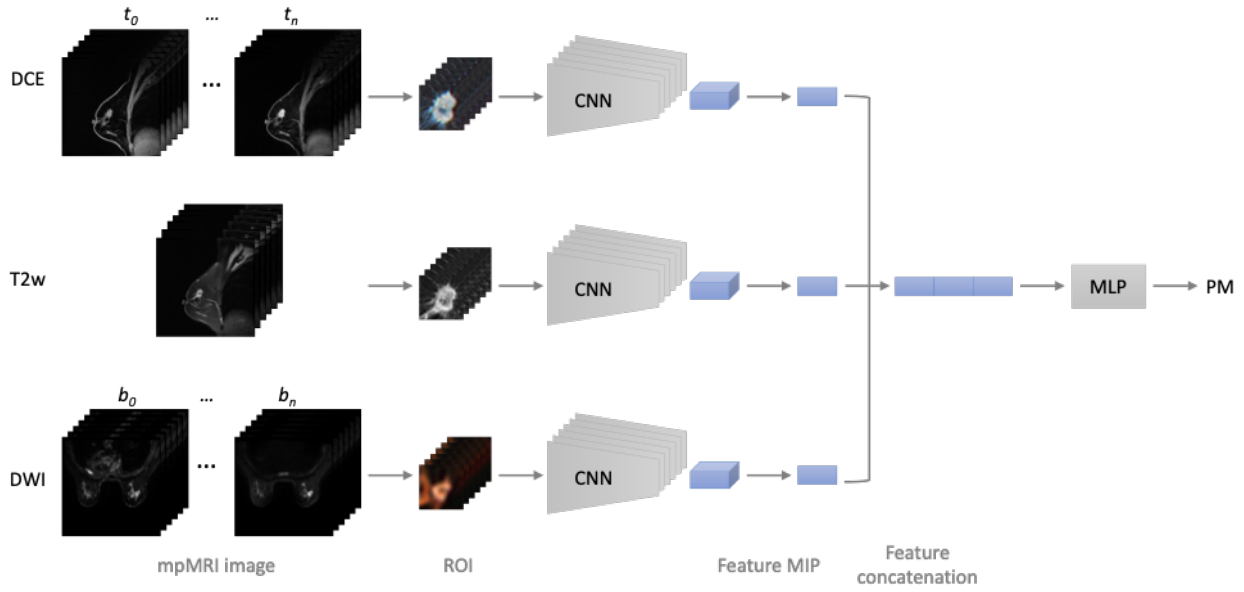


Figure 3.6: Lesion classification pipelines for multiparametric MRI. Feature MIP was applied to the lesion volume in three MRI sequences, and features from these sequences were concatenated and input to a multilayer perceptron (MLP).

3.4 Results

3.4.1 Volumetric and Temporal Information in DCE-MRI

Figure 3.8 presents the ROC curves of the three classification schemes in the task of distinguishing benign and malignant breast lesions. Table 3.2 summarizes the classification performances and compares the two newly proposed approaches with the method using MIP images. The 95% CIs of the difference in AUCs (ΔAUC) and the p-values demonstrate that both classifiers that collapsed 3D volumetric information by feature MIP significantly outperformed the previously proposed method of using image MIP. Meanwhile, the two feature MIP classifiers, which leverage temporal information differently, failed to show a significant difference in performance from each other. In addition, equivalence testing demonstrated that these two feature MIP classifiers yielded equivalent performance with an equivalence margin of $\Delta\text{AUC} = 0.05$, chosen *prima facie*.

The results suggest that in the task of distinguishing benign and malignant breast lesions

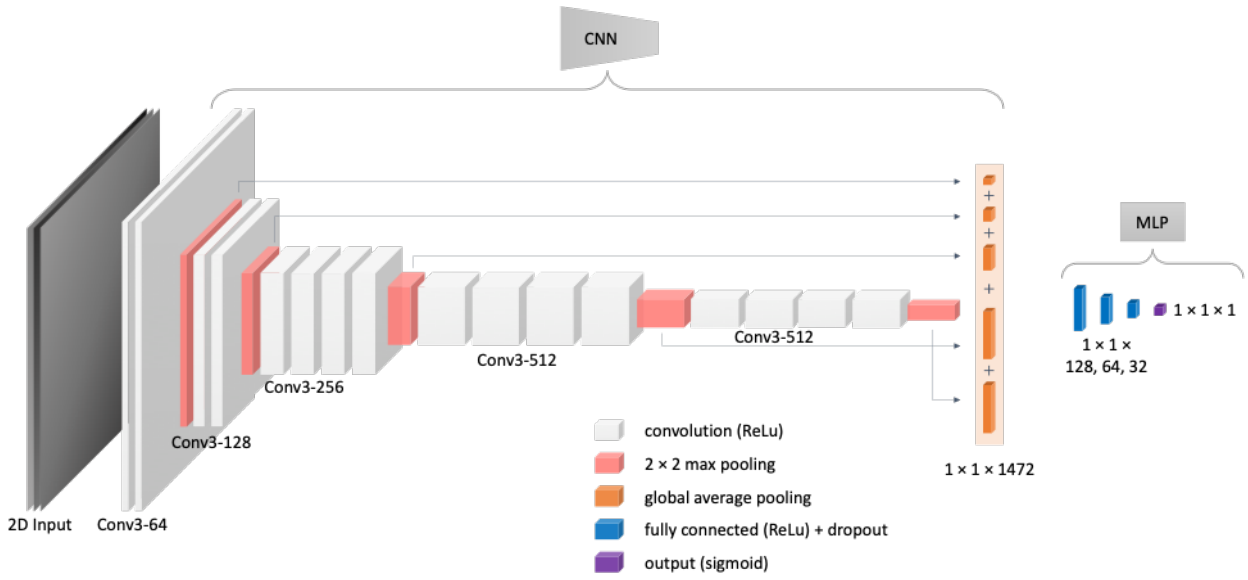


Figure 3.7: Architecture of the modified VGG-19 and the multilayer perceptron (MLP) of which the overall model was composed.

using deep transfer learning, 3D volumetric information in DCE-MRI may have superior predictive power when collapsed along the axial dimension by maximum intensity projection at the feature level rather than at the image level. The temporal information in DCE-MRI, on the other hand, contributes equivalently to the classification task when incorporated in subtraction images or in the three channels of RGB images.

Table 3.2: Classification performances and comparisons between the three classification schemes using the DeLong test. The p-value and 95% confidence interval (CI) of the difference in the areas under the receiver operating characteristic curves (AUCs) were computed with respect to the classifier using maximum intensity projection (MIP) images. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections.

	AUC ± SE	P-value	95% CI of ΔAUC
Subtraction image MIP	0.86 ± 0.01	—	—
Subtraction volume CNN feature MIP	0.89 ± 0.01	$< .001^*$	$[0.01, 0.04]$
RGB volume CNN feature MIP	0.89 ± 0.01	$< .001^*$	$[0.02, 0.04]$

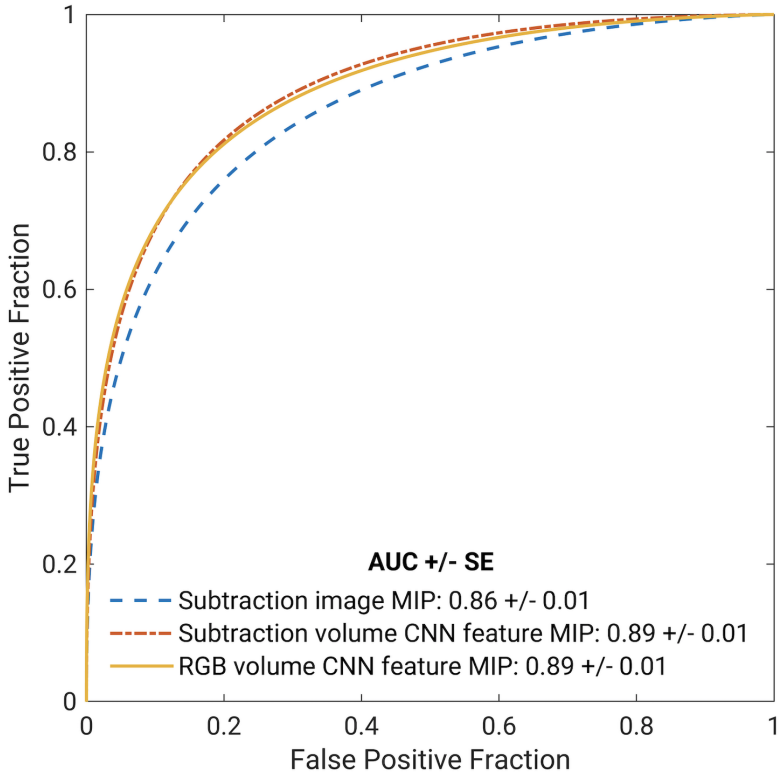


Figure 3.8: Fitted binormal receiver operating characteristic (ROC) curves for two classifiers that utilize the volumetric and temporal information from dynamic contrast-enhanced (DCE)-MRI. The legend gives the area under the ROC curve (AUC) with standard error (SE) for each classifier scheme.

Table 3.3: Performance metrics comparison between image maximum intensity projection (MIP) and feature MIP models [17]. The area under the receiver operating characteristic curve (AUC), along with the standard error and the 95% CI, as well as the sensitivity and specificity (in percentage and ratio of cases) for each method. The 95% CI and p-value for the difference (Δ) between the two methods are also presented for each metric. The AUCs were compared using the DeLong test, and the sensitivities and specificities were compared using the McNemar test.

Classifier	Image MIP	Feature MIP	95% CI of Δ	P-value
AUC	0.91 ± 0.02 [0.87, 0.94]	0.93 ± 0.01 [0.91, 0.96]	[0.003, 0.051]	.03
Sensitivity	90% (379/421)	94% (395/421)	[0.014, 0.062]	.002
Specificity	73% (83/114)	72% (82/114)	[-.094, 0.076]	$> .99$

3.4.2 Independent Validation of Feature MIP on DCE-MRI

Figure 3.9 presents the ROC curves of the image MIP (AUC: 0.91, 95% CI: [0.87, 0.94]) and the feature MIP (AUC: 0.93, 95% CI: [0.91, 0.96]) approaches, and Table 3.3 summarizes the classifiers' performance metrics in the task of distinguishing benign and malignant breast lesions. A DeLong test comparing the feature MIP method with the image MIP method demonstrated that the feature MIP method achieved a higher classification performance (ΔAUC 95% CI: [0.003, 0.051], $P = .03$). These results suggest that collapsing 3D volumetric information by taking the maximum intensity projection at the feature level retained higher predictive power than collapsing at the image level. McNemar tests showed that, at the operating point determined using the training set, the sensitivity of the feature MIP method on the test set was significantly higher than that of the image MIP method, and the specificities failed to demonstrate a significant difference.

Figures 3.10 and 3.11 illustrate the comparison between the PMs predicted using the image MIP method and feature MIP methods. Although the majority of benign and malignant lesions were separated from the other class by both image MIP and feature MIP, these two methods exhibit moderate disagreement between these figures. Overall, the feature MIP method assigned malignant cases with higher PMs and benign cases with lower PMs as compared with image MIP, indicating that the feature MIP classifier has higher discriminatory

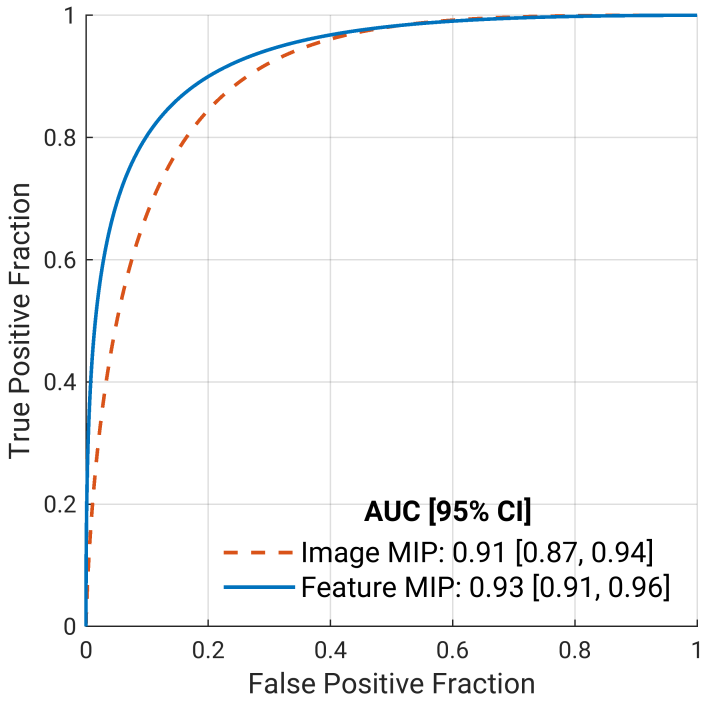


Figure 3.9: Fitted binormal receiver operating characteristic (ROC) curves for two classifiers that use the four-dimensional volumetric and temporal information from dynamic contrast-enhanced MRI [17]. The legend gives the area under the ROC curve (AUC) with the 95% CI for each classifier.

power than image MIP in distinguishing between benign and malignant lesions. Figure 3.10 also shows several example lesions on which one method generated more accurate predictions than the other or on which the two methods agreed. For lesions on which feature MIP predicted more accurately than image MIP, the MIP images either failed to retain important features of the lesions or captured misleading features that do not accurately represent the lesions volumes in the projection process.

3.4.3 Deep Learning for DWI

As shown in Fig. 3.12, classification of benign and malignant lesions using the ADC map input and the RGB input yielded AUC values [95% CI] of 0.81 [0.76, 0.85] and 0.83 [0.79, 0.87], respectively. The AUCs failed to demonstrate a statistically significant difference ($P = .27$, 95% CI of $\Delta\text{AUC} = [-0.07, 0.02]$). The RGB input was noninferior to the ADC map input within a margin of $\Delta\text{AUC} = 0.05$. Figure 3.12 also shows examples of ADC input ROIs and RGB input ROIs of both malignant and benign lesions. Given the classification performances, the RGB input will be used in the mpMRI part of this study, because it does not require the computationally intensive calculation of the ADC map and is not subject to variations in the image post-processing algorithm.

3.4.4 Feature MIP on Multiparametric MRI

Figure 3.13 and Table 3.4 present the classification performance of each classifier that utilized the high-dimensional information in their corresponding MRI sequence and of the fusion mpMRI classifier. Table 3.4 also summarizes the comparison results between the single-sequence classifiers and mpMRI classifier. The mpMRI classifier yielded a high AUC [95% CI] of 0.94 [0.92, 0.96] and significantly outperformed all single-parametric classifiers. Figure 3.14 shows examples of the input ROI and Grad-CAM visualization for a benign lesion and a malignant lesion. One slice is shown for each lesion volume. In both cases, the deep

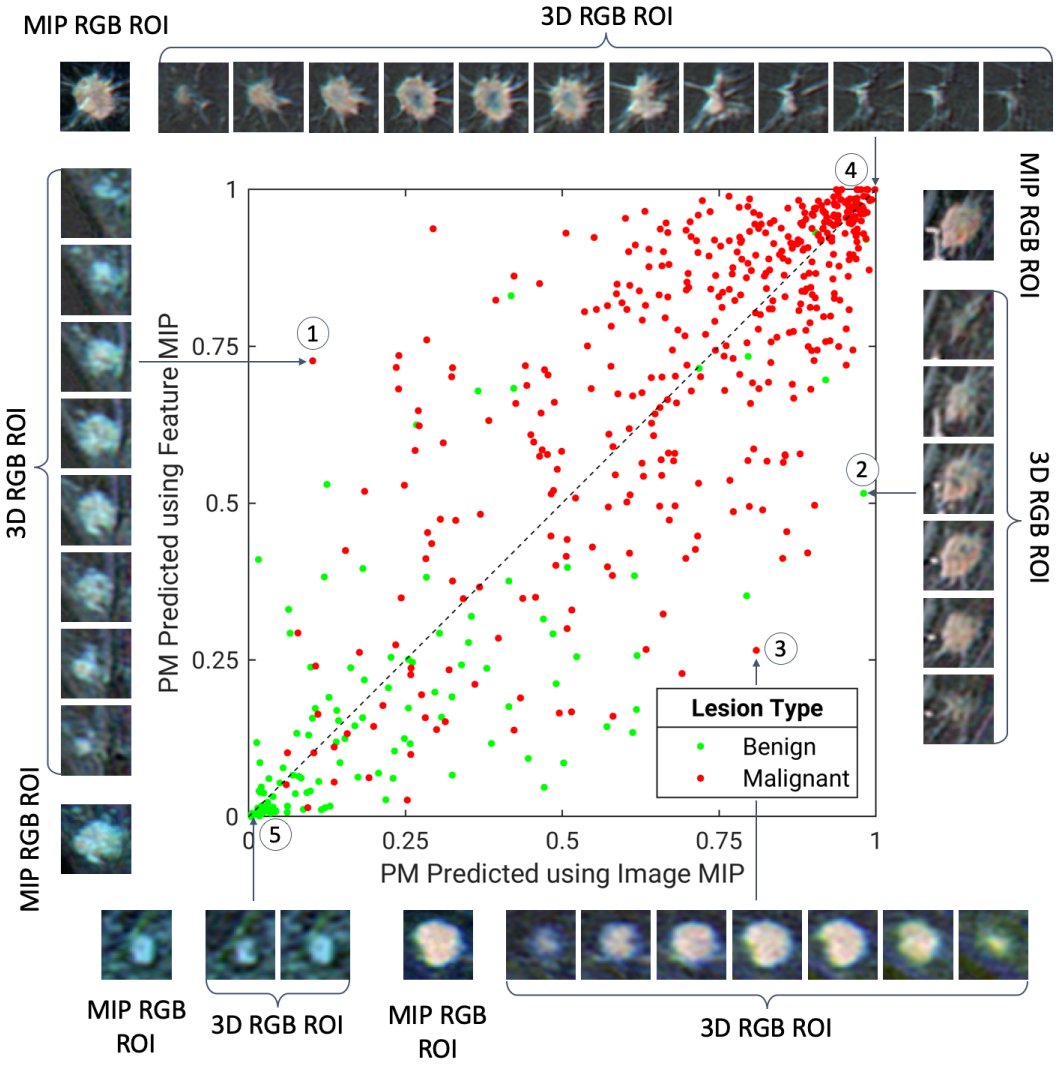


Figure 3.10: A diagonal classifier agreement plot between the image maximum intensity projection (MIP) and feature MIP methods [17]. The x-axis and y-axis denote the probability of malignancy (PM) scores predicted by the image MIP classifier and feature MIP classifier, respectively. Each point represents a lesion for which predictions were made. Points along or near the diagonal from bottom left to top right indicate high classifier agreement; points far from the diagonal indicate low agreement. The insets are the MIP regions of interest (ROIs) and three-dimensional (3D) ROIs, which served as convolutional neural network (CNN) inputs for the image MIP and feature MIP methods, respectively, of extreme examples on which using feature MIP resulted in more accurate predictions than using image MIP (lesion 1-2), on which using image MIP resulted in more accurate predictions than using feature MIP (lesion 3), and on which the two methods both predicted accurately (lesion 4-5). Lesion 1: invasive micropapillary carcinoma; lesion 2: fibromatosis; lesion 3: invasive ductal carcinoma, grade II; lesion 4: invasive ductal carcinoma, grade II; lesion 5: non-mass enhancement, fibroadenoma.

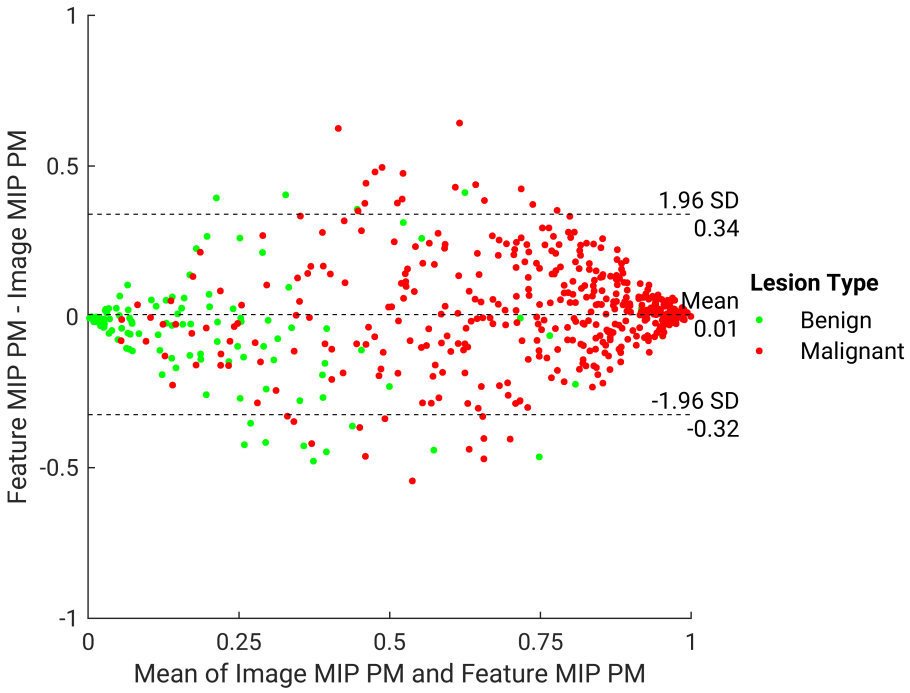


Figure 3.11: Bland-Altman plot for the image maximum intensity projection (MIP) and feature MIP classifiers [17]. The x-axis and y-axis show the mean and difference between the support vector machine output scores (i.e., predicted posterior probabilities of malignancy [PMs]) of the two classifiers, respectively.

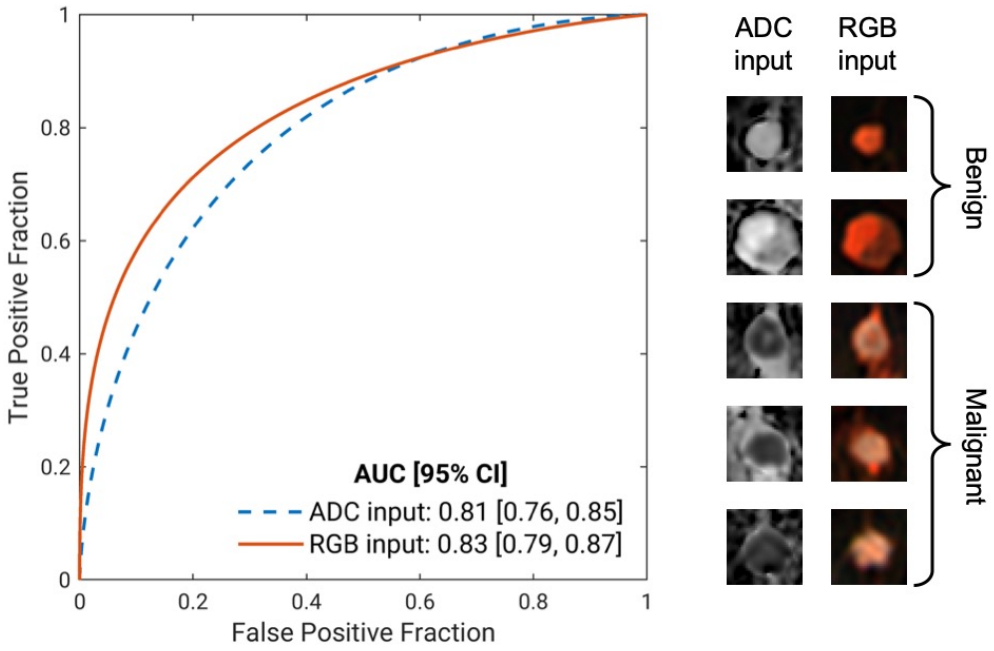


Figure 3.12: Fitted binormal receiver operating characteristic (ROC) curves for two diffusion-weighted imaging classifiers. Examples of the two classifiers' inputs are shown on the right. The legend gives the area under the ROC curve (AUC) with the 95% CI for each classifier.

learning algorithm identified complementary information on the three sequences in mpMRI, and consequently, the classification performance of the mpMRI fusion classifier was improved compared with using any single sequence alone.

Table 3.4: Area under the receiver operating characteristic curve (AUC) for single-sequence classifiers of three MRI sequences and a fusion mpMRI classifier. P-value and 95% confidence interval (CI) of the difference in AUCs are presented for the comparison between each single-sequence classifier and the multiparametric classifier using the DeLong test. Asterisks denote significance after accounting for multiple comparisons using Bonferroni-Holm corrections.

Classifier	DCE	T2w	DWI	mpMRI
AUC [95% CI]	0.92 [0.90, 0.95]	0.84 [0.80, 0.88]	0.86 [0.82, 0.90]	0.94 [0.92, 0.96]
95% CI of ΔAUC	[0.001, 0.036]	[0.065, 0.137]	[0.050, 0.116]	—
P-value	$P = .04^*$	$P < .001^*$	$P < .001^*$	—

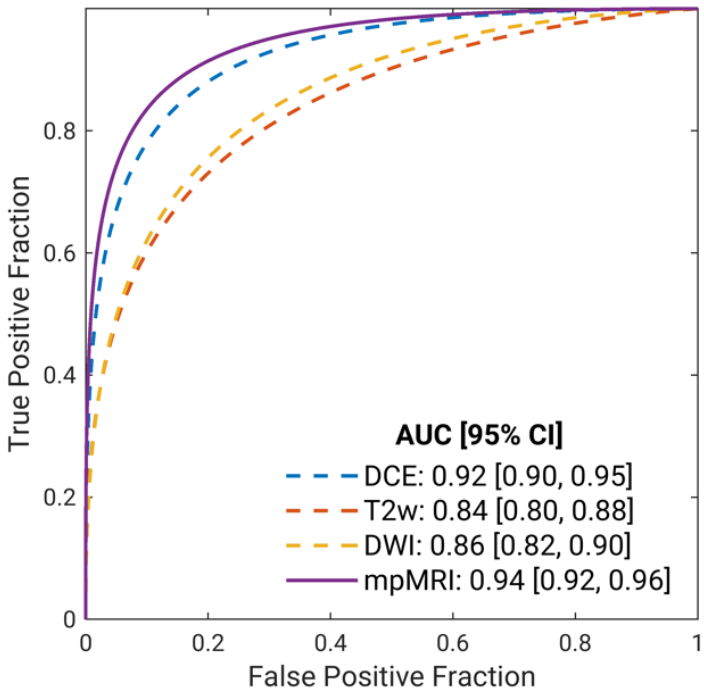
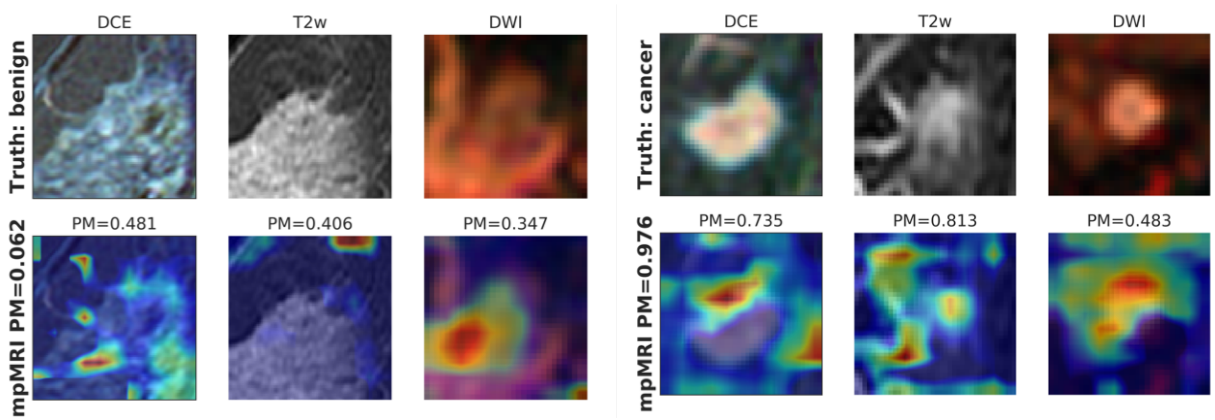


Figure 3.13: Fitted binormal receiver operating characteristic (ROC) curves for single-sequence classifiers of three MRI sequences and a fusion mpMRI classifier, all of which utilize the high-dimensional information in the images. The legend gives the area under the ROC curve (AUC) with the 95% CI for each classifier.

3.5 Discussion and Conclusions

The work presented in this chapter proposed an approach, referred to as feature MIP, to effectively incorporate the high-dimensional information inherent in MRI exams when using deep transfer learning in the task of distinguishing between benign and malignant breast lesions [17, 18]. The feature MIP method globally max pools the features extracted from a lesion volume along the lesion's axial dimension within a CNN. For 4D sequences, namely DCE and DWI, the RGB channels of CNNs pretrained on natural images are utilized to incorporate the images acquired at different time points in DCE and different diffusion weighting strengths in DWI. Compared with a previous method of using image MIP on DCE-MRI, the feature MIP method demonstrated significantly higher performance in the



(a) (b)

Figure 3.14: Example input ROI (top row) and their Grad-CAM heatmap overlays (bottom row) of (a) a benign breast lesion and (b) a malignant breast lesion. The probability of malignancy (PM) predicted by each single-sequence classifiers is shown above its corresponding heatmap overlay, and the PM predicted by the mpMRI classifier is shown on the left.

breast lesion classification task. When applying the feature MIP method along with the feature fusion method from Chapter 2, the high-dimensional mpMRI classifier achieved high classification performance (AUC = 0.94, 95% CI: [0.92, 0.96]) and significantly outperformed using any single sequence alone. Methods examined for incorporating the temporal aspect of DCE-MRI and the diffusion weightings of DWI did not yield significant differences in classification performance.

It is worth noting that the advantage of feature MIP relative to image MIP for utilizing volumetric information in deep learning is relatable to the perception of human readers. Given the anatomical complexity in breast parenchyma, the anatomical clutter caused by projecting a 3D volume onto a 2D image is a limiting factor for human readers' assessment [130–132]. Therefore, although conventional clinical MIP images are a convenient way of reducing the dimensionality of DCE-MRI for image interpretation by either radiologists or AI algorithms, they do not yield optimal results due to the loss of information and the enhanced anatomical noise in projection images.

A prior study from our group was based on the same dataset and training, validation,

and test split as our study, but used a single representative slice for each lesion and input the pre-contrast, first post-contrast, and second post-contrast image ROIs into the RGB channels [38]. The study reported an AUC of 0.85 using the same modified VGG-19 feature extraction and support vector machine classification approach. The newly proposed feature MIP method in this study outperformed the above-mentioned method by 10% (ΔAUC 95% CI: [0.035, 0.120], $P < .001$).

Training 3D CNNs from scratch is another common approach for taking advantage of high-dimensional information provided by medical images. However, it is computationally expensive and is usually not suited for moderately sized medical datasets. A recent study by Dalmis et al. trained a 3D CNN from scratch on 4D ultrafast DCE-MRI data after reducing the dimensionality using MIPs and achieved an AUC [95% CI] of 0.81 [0.77, 0.85] [108]. Another study by Li et al. trained a 3D CNN on the volume of DCE-MRI and incorporated the temporal information in the classification by calculating the enhancement ratio; they reported an AUC of 0.84 [133]. Compared with training 3D CNNs from scratch, our methodology of using transfer learning on 4D medical imaging data involves training of much fewer free parameters and therefore is computationally more efficient and has demonstrated high performance on the moderately sized dataset used in this study.

Moreover, there exist several variations of transfer learning strategies as mentioned in Section 1.1.2, including using the CNN as a feature extractor with fixed weights, or optionally adding fully connected layers on top of the pretrained network and fine-tuning the network end-to-end. In this work, the specific transfer learning strategy employed in each section was chosen through preliminary experiments and comparisons to optimize for the performance in the task of distinguishing benign and malignant breast lesions on our dataset.

A limitation in the evaluation is that without sufficient knowledge about the specific clinical use case, the operating points at which sensitivity and specificity are reported may not be clinically optimal. A different threshold may be chosen if the relative cost of false-

positive and false-negative diagnoses is known. Another limitation, similar to Chapter 2, is the choice of equivalence margin, which can be revised in the future when guidelines for selecting clinically meaningful margins are available. In addition, as in Chapter 2, MRI exams used in this study were collected over several years and image acquisition parameters such as spatial resolution and DCE temporal resolution were also variable within the dataset. The use of such a retrospectively collected, heterogeneous dataset contributed to the algorithm robustness and generalizability. Future work will continue investigating the harmonization of differences in acquisition parameters to improve performance.

CHAPTER 4

ARTIFICIAL INTELLIGENCE FOR COVID-19 DIAGNOSIS AND PROGNOSIS ON CHEST RADIOGRAPHY

4.1 Introduction

The prolonged COVID-19 pandemic has profoundly impacted global public health and the economy. As mentioned in Section 1.3, since the SARS-CoV-2 virus is highly contagious and infection can cause severe and sometimes fatal disease, early diagnosis and appropriate patient management are crucial when navigating the pandemic, both for the patient's well-being and for public health purposes. Early diagnosis not only allows for prompt treatment at the earlier, more manageable stage of the disease but also informs patient isolation based on disease mitigation and containment strategies. Accurate prognosis enables planning and optimization of medical resource allocation as well as choosing the appropriate intervention and implementing necessary adjustments.

Early on in the pandemic, the containment of infection was hindered by a shortage of the reverse transcription polymerase chain reaction (RT-PCR) assay. While there have been successful efforts to increase the production capacity, shortages of test kits and long processing times remain a problem in resource-limited settings during surges. Moreover, the RT-PCR test has moderate and variable sensitivity in clinical practice [69]. Chest radiography (CXR) is recommended for triaging at patient presentation and disease monitoring due to its fast speed, relatively low cost, wide availability, and portability [75, 76]. Characteristics such as bilateral lower lobe consolidations, ground glass densities, peripheral air space opacities, and diffuse air space disease on CXR have been related to COVID-19 [77, 78]. Unfortunately, the non-specificity of these features and the shortage of radiological expertise due to the stress on healthcare resources during the pandemic make precise interpretation of such images challenging. Under such circumstances, deep learning can potentially assist in this task.

The work presented in this chapter investigates deep learning image analysis methods for automated COVID-19 diagnosis at patient presentation and for predicting COVID-19 patients' future needs of intensive care using CXR. The role of standard and soft-tissue CXR is also examined.

There have been numerous studies on AI applications for COVID-19 using CXR. However, due to difficulties in collecting sizeable datasets, many of these studies utilized publicly available datasets that consist of images extracted from publications [134–136], which by nature are not a representative selection of the patient population and may lead to biased results that cannot be recommended for clinical use [137, 138]. We therefore curated a large CXR database consecutively collected from our institution for this research. Moreover, large public CXR datasets established prior to the pandemic were usually utilized to enrich the training set [139–141]. These images, all COVID-19 negative, differed from the COVID-19 positive cases in newly acquired datasets in image acquisition protocol, scanners, and patient population. Consequently, pooling them together would have introduced confounding variables into the classification task and potentially yielded over-optimistic results, because the models might learn to use these irrelevant factors to distinguish COVID-19 positive and negative, rather than identify disease presentations. In order to leverage pre-pandemic CXR datasets without adding confounding variables in this study, we designed a three-phase learning curriculum to sequentially fine-tune the model during training instead of pooling the datasets. Furthermore, while most current diagnostic imaging AI research was developed on all COVID-19 cases available, including images acquired when the disease has progressed, our study tackled the challenge of COVID-19 early diagnosis at initial patient presentation, which is important for implementing isolation and treatment promptly. Finally, while most prior studies only considered standard CXR images, our work also investigated the role of soft-tissue images in automated COVID-19 diagnosis using deep learning as they exhibit diagnostic utility in radiologists' clinical assessment.

4.2 Datasets

A database has been retrospectively and consecutively curated under a HIPAA-compliant, IRB-approved protocol during the COVID-19 outbreak. From adult patients who underwent the RT-PCR test for SARS-CoV-2 virus at the University of Chicago Medical Center, their CXR exams after and up to a year prior to their initial RT-PCR tests were collected. Table 4.1 summarizes the database as of February 12, 2021.

Table 4.1: Dataset statistics by patient and total images.

	Adult	Pediatric
COVID-19+	3046	170
COVID-19-	16190	1334
Total	19236	2127
Total images	65288	9290

For the early diagnosis study, the first CXR exam after (with a limit of two days) each patient's initial RT-PCR test was selected for this study. Dual-energy subtraction (DES) exams and portable exams with a ClearRead bone suppression series (Riverain Technologies) were included; the former generate a soft-tissue image per exam using images obtained at two different beam energies, while the latter generate a synthetic soft-tissue image using post-processing algorithms. Exams with a missing standard image or soft-tissue image were excluded. Ultimately, the dataset for this study consisted of 9860 adult patients who had CXR exams within two days after their initial RT-PCR tests acquired between January 30, 2020 and February 3, 2021, 1523 (15.5%) of whom tested positive and 8337 (84.5%) of whom tested negative for COVID-19. The dataset was split at the patient level into 64% for training, 16% for validation, and 20% for testing using stratified sampling, keeping the class prevalence constant across the three subsets. Detailed statistics on the dataset are summarized in Table 4.2. Figure 4.1 shows the distribution of the patient visit status, i.e., settings in which the CXR exams were acquired among COVID-19 positive patients in the

dataset. Note that the COVID-19 positive patients who were hospitalized or in the intensive care unit (ICU) at the time of the CXR could have been receiving treatment for diseases other than COVID-19, and COVID-19 might not have been the primary reason for their hospital stay. Thus, we did not assume COVID-19 severity based on patient visit status.

Table 4.2: Dataset statistics and the prevalence of cases for initial CXR exams.

	COVID-19+	COVID-19-	Total
Training	974	5336	6310 (64%)
Validation	244	1334	1578 (16%)
Test	305	1667	1972 (20%)
Total	1523 (15%)	8337 (85%)	9860

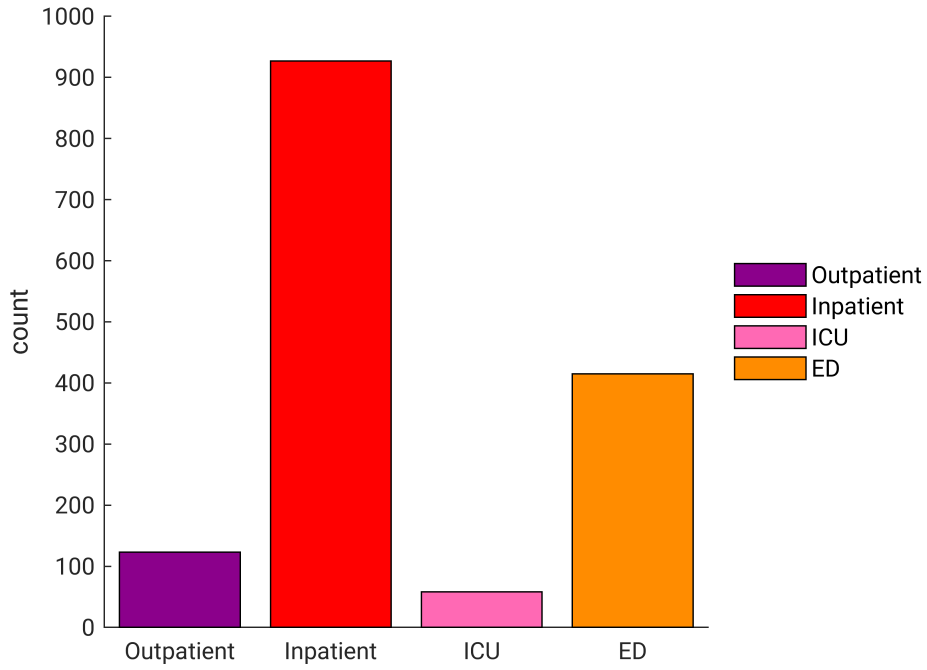


Figure 4.1: Distribution of the patient visit status in which chest radiography exams were acquired among COVID-19 positive patients in the dataset. ICU = intensive care unit, ED = emergency department.

For the prognostic study, images acquired after a positive RT-PCR were included, and images obtained after ICU admission or intubation were excluded. Ultimately, the dataset for

this study consisted of 1670 CXR exams of 1178 COVID-19 positive adult patients, acquired between March 20, 2020 and September 13, 2020. Intensive care was defined as intubation (invasive mechanical ventilation) and/or ICU admission. Since the medical resources at our institution were not overwhelmed by any measure throughout the pandemic, we assumed that all patients who needed intensive care received it without delay. The dataset was split at the patient level into 64% for training, 16% for validation, and 20% for testing using stratified sampling, holding the class prevalence for the least frequent outcome, i.e., intubation or ICU admission within 24 hours, constant across all subsets. Detailed statistics on the dataset are summarized in Table 4.3.

Table 4.3: Dataset statistics for patients who required intensive care within 24, 48, 72, and 96 hours after chest radiography exams. The numbers of patients and images (in parentheses) in each subset are listed.

	24 hours		48 hours		72 hours		96 hours	
	+	-	+	-	+	-	+	-
Training	135 (152)	601 (916)	144 (174)	592 (894)	147 (193)	589 (875)	148 (210)	588 (858)
Validation	34 (38)	162 (230)	40 (48)	156 (220)	41 (51)	155 (217)	42 (53)	154 (215)
Test	43 (47)	203 (287)	47 (56)	199 (278)	48 (57)	198 (277)	49 (58)	197 (276)
Total	212 (237)	966 (1433)	231 (278)	947 (1392)	236 (301)	942 (1369)	239 (321)	939 (1349)

4.3 Methods

4.3.1 Early Diagnosis

Inspired by curriculum learning proposed by Bengio et al. [142], a sequential transfer learning strategy was employed to train the model in three phases on gradually more specific and complex tasks, mimicking the human learning process. As illustrated in Fig. 4.2, instead

of presenting the model with a random mixture of CXR examples and directly training it to diagnose COVID-19, the curriculum was designed to fine-tune the model in a cascade approach in three phases: 1) First, the model was pretrained on natural images in ImageNet and fine-tuned on the National Institutes of Health (NIH) ChestX-ray14 dataset to diagnose a broad spectrum of 14 pathologies [34, 139]. 2) Then, the model was refined on the Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset, which has a high pneumonia prevalence, to detect opacities caused by pneumonia [140]. 3) Finally, the model was fine-tuned further on the training set of the COVID-19 dataset. The final model was then evaluated on the held-out independent test set to distinguish between CXRs of COVID-19 positive and COVID-19 negative patients. The DenseNet-121 architecture was chosen for the task because of its advantages mentioned in Section 1.4.2 and its success in diagnosing various diseases on CXR in previous publications [143, 144].

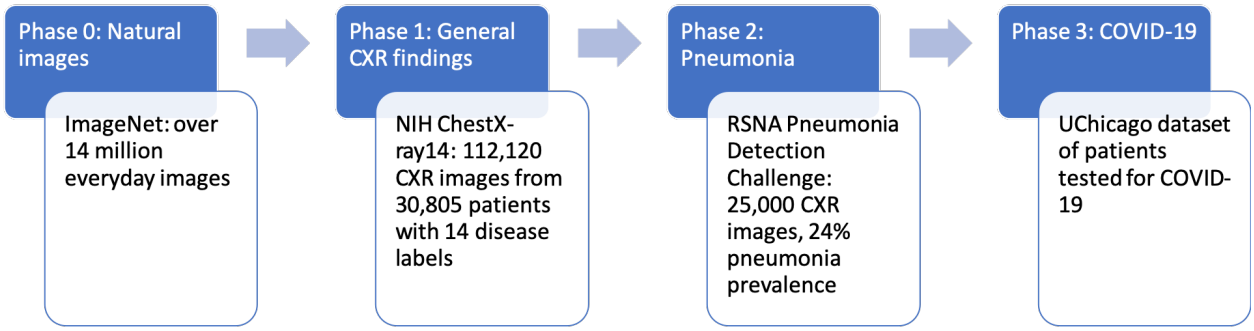


Figure 4.2: The sequential transfer learning curriculum for the diagnosis of COVID-19, and information on the dataset for each phase of training [19].

The phase 1 test set was specified by the original database curators of NIH ChestX-ray14, and the rest of the dataset was randomly divided at the patient level into approximately 80% for training and 20% for validation. The DenseNet-121 model was initialized with weights optimized for ImageNet (Phase 0), and the final classification layer was replaced with a 14-node fully connected layer with sigmoid activation. Images were downsampled

by a factor of four to 256×256 pixels, gray-scale normalized, and randomly augmented by horizontal flipping, shifting by up to 10% of the image size, and rotation of up to 8 degrees. The model was trained with a batch size of 64, weighted cross-entropy loss function, and Adam optimizer with an initial learning rate of 0.001. Step decay on learning rate and early stopping were employed. The misclassification penalty for cases in a class was assigned to be inversely proportional to its class prevalence to address the problem of class imbalance.

The phase 2 test set containing 1000 images was specified by the database curators of the RSNA Pneumonia Detection Challenge dataset, and the rest was split randomly at the patient level into 80% for training and 20% for validation, holding the class prevalence constant in the subsets. The DenseNet-121 model was initialized with the weights from phase 1, and the final classification layer was replaced with a one-node fully connected layer with sigmoid activation to differentiate whether CXR images contain evidence of pneumonia. Methods for image preprocessing and model training followed that in phase 1, except the initial learning rate was reduced to 0.0001.

In phase 3, as illustrated in Fig. 4.3, a U-Net-based model was used to segment the lung region on all images in our dataset in this phase in order to reduce the influence on the classification model from irrelevant regions in the images. Compared to the original U-Net proposed by Ronneberger et al., the architecture used in this study was augmented with inception blocks and residual blocks, as detailed in Clark et al. [11, 98]. We used the weights that had been pretrained on a pre-pandemic public CXR dataset for lung segmentation and fine-tuned on an external CXR dataset that included COVID-19 patients [135, 145, 146]. Open and close operations were performed in the post-processing steps to fill holes and reduce noise in the predicted masks. Then the smallest rectangular region that was able to enclose the predicted lung mask was cropped from each image. The masks were predicted using standard CXR images, and their corresponding soft-tissue CXR images were cropped using the same masks. The cropped images were resized to 256×256 pixels.

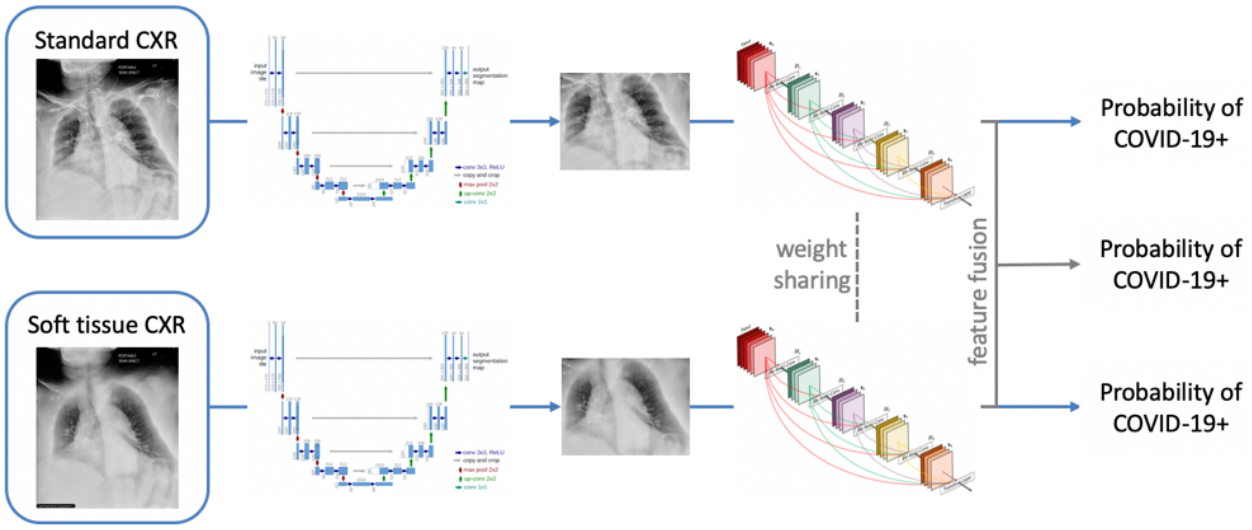


Figure 4.3: Illustration of Phase 3 in the sequential training process, fine-tuning the model on the pandemic-era CXR dataset to distinguish between COVID-19 positive and negative patients. The model architectures shown are for illustration purposes and are not the precise or complete architectures of the modified U-Net and the DenseNet models.

Image preprocessing other than the cropping step and model training followed the methods in phase 2, and the model was initialized with the weights from phase 2. The fine-tuning and evaluation in this phase were performed on the full standard CXR images, the cropped standard CXR images, and the cropped soft-tissue images in the dataset. The combined use of cropped standard and soft-tissue images was then investigated in a feature fusion manner as described in Section 2.3.1, which had demonstrated superior ability for leveraging multiparametric images [13, 15]. Specifically, the standard and soft-tissue images were input to two DenseNet-121 models trained with shared weights, whose activation maps prior to the final fully connected layer were concatenated, forming the ensemble of features extracted from the two types of input for classification. Training of the parallel models was split across four GPUs memory allocation.

4.3.2 Prognosis

The image processing and classifier training process for the prognosis task also followed the learning curriculum described in Section 4.3.1. The last phase of fine-tuning was adjusted for the task, making four independent predictions for the probabilities that a patient will need intensive care within the next 24, 48, 72, and 96 hours from the time of the CXR exam. The last layer in the model was changed to a four-node fully connected layer with sigmoid activation. Since the soft-tissue images in this database did not show significant additional benefit in previous studies, only standard CXR images were used.

4.3.3 Evaluation and Statistical Analysis

In each phase, the classification performance for each task and each label was evaluated using receiver operating characteristic (ROC) curve analysis with the area under the ROC curve (AUC) as the figure of merit. The 95% confidence intervals (CIs) of the nonparametric Wilcoxon–Mann–Whitney AUCs were calculated by bootstrapping (2000 bootstrap samples) [119]. The proper binormal model was used to plot the ROC curves [118]. All reported classification performance metrics pertain to the held-out independent test set in each phase ($N = 25596$ for phase 1, $N = 1000$ for phase 2, $N = 1972$ patients for phase 3 for diagnosis, and $N = 1178$ for prognosis). Multiple methods for diagnosis on the in-house COVID-19 dataset in phase 3 were compared in terms of AUC using the DeLong test and equivalence test [120, 123]. Bonferroni-Holm corrections were used to account for multiple comparisons [122]. A corrected $P < 0.05$ was considered to indicate a statistically significant difference in performance, and an equivalence margin of $\Delta\text{AUC} = 0.05$ was chosen *prima facie*. Additional evaluation metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score, were calculated and reported at four sensitivity levels for the different methods in phase 3. Bland-Altman analysis was used as an adjunct method to compare the estimated COVID-19 probabilities for individual patients for

the different classification approaches. Gradient-weighted class activation mapping (Grad-CAM) was generated to provide a visual explanation of the model's classification [129]. The test set evaluation was also performed on the DES exam subset and the portable exam subset separately for the early diagnosis task.

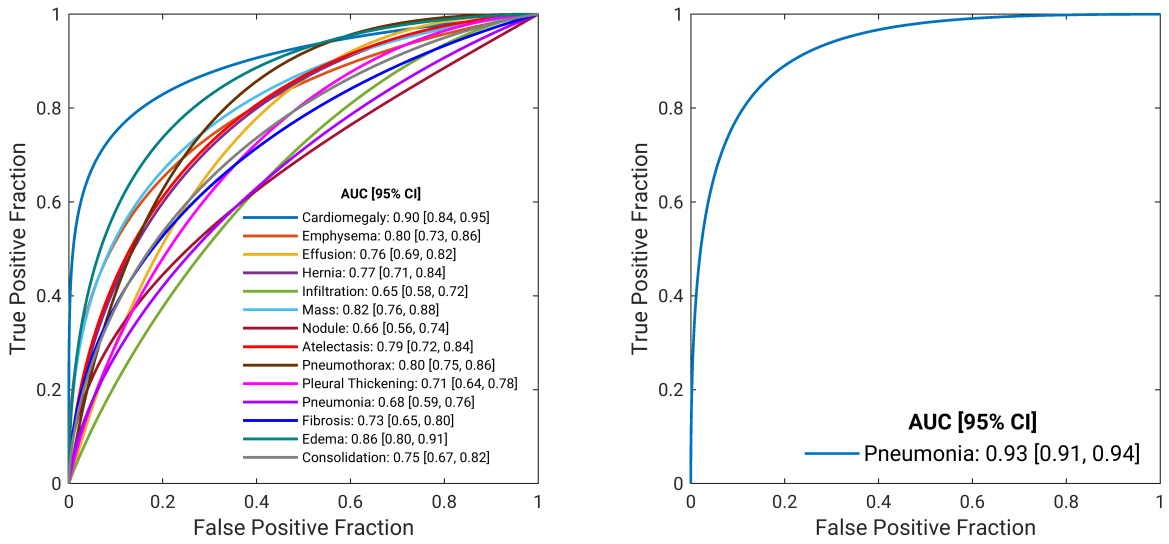
4.4 Further Investigations: Classification and Visualization using Multiple Instance Learning

Using attention-based deep multiple instance learning (MIL) for classification and visualization in the task of COVID-19 diagnosis on CXR was investigated. Each CXR image, after U-Net segmentation and cropping described in Section 4.3.1, was divided into 32×32 overlapping patches, with a step size of 16 pixels. The theory of the attention-based deep MIL is introduced in Section 1.4.3. In this application, each patch is considered an instance, and each CXR image is considered a bag. Bag-level labels, i.e., COVID-19 status of the patient at the time of CXR acquisition, were available, while instance-level labels, i.e., whether a patch of CXR image contained abnormalities related to COVID-19, were not available. The MIL model produced bag-level predictions, and the attention scores were extracted to generate heatmaps that visualized how much each patch contributed to the classification prediction. Experiments were performed on model architectures, loading pretrained weights, and the portion of the architecture to freeze or retrain on the patches. The attention heatmap was shown in three formats, first with the attention scores multiplied with patches, second with the attention scores converted to smooth heatmaps displayed in color, and third with bounding boxes drawn of patches with highest attention scores overlaid on the CXR images. The evaluation of the classification performance followed Section 4.3.3.

4.5 Results

4.5.1 Early Diagnosis

The ROC curves for the classification tasks in the first two phases are shown in Fig. 4.4. Phase 1 and phase 2 yielded AUC values similar to recent publications on the same tasks using these datasets [143, 144, 147].



(a) (b)

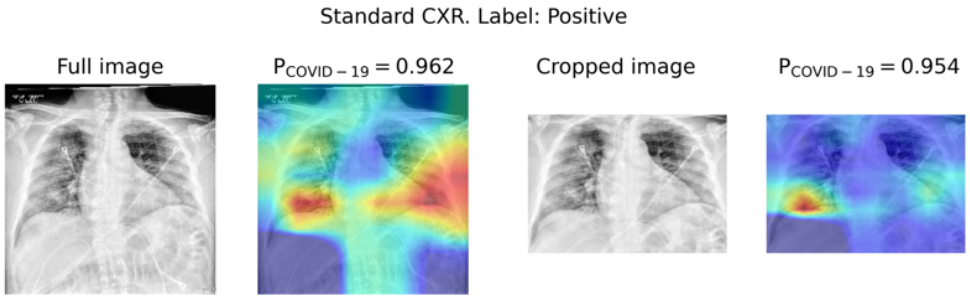
Figure 4.4: Fitted proper binormal ROC curves for classification tasks in the first two phases of training [19]. The legend gives the AUC with 95% CI for each classification task.

In phase 3, when full standard CXR images were used as input, the model achieved an AUC of 0.74 (95% CI: 0.70, 0.77), which was significantly lower than an AUC of 0.76 (95% CI: 0.73, 0.79) obtained when cropped standard CXR images were used, as shown in Table 4.4. Figure 4.5 shows examples of Grad-CAM heatmaps demonstrating that when full images were used as input, areas outside the patient body (e.g., the text label on the image in Fig. 4.5a) and areas outside the lungs (e.g., abdominal region and chest walls in Fig. 4.5a and shoulder and neck region in both Fig. 4.5a and Fig. 4.5b) contributed to the classification model's prediction. In contrast, influence from these irrelevant regions was eliminated or

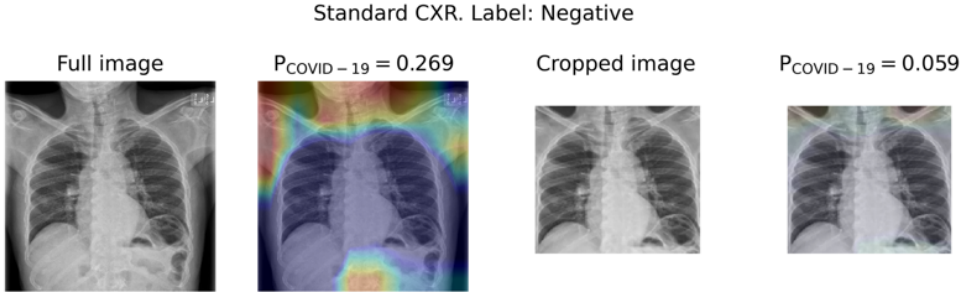
reduced by using cropped images. Due to the superior classification performance and the reduced influence from areas outside the lungs, cropped images were used in our subsequent analysis.

Table 4.4: AUC values for using full and cropped standard CXR, and the p-value and 95% CI of the difference in AUC values. Asterisks denote statistical significance.

	Full standard CXR	Cropped standard CXR
AUC [95% CI]	0.74 [0.70, 0.77]	0.76 [0.73, 0.79]
p-value	0.04*	
95% CI for ΔAUC	[0.001, 0.049]	



(a)



(b)

Figure 4.5: Example standard chest radiographs (CXR) and their Grad-CAM heatmaps overlays of (a) a COVID-19 positive case and (b) a COVID-19 negative case. The model prediction scores ($P_{COVID-19}$) are noted. Both examples show influence on model predictions from irrelevant areas outside the lungs when the full images were used, which was reduced when the cropped images derived from automatic lung segmentation were used.

For the three classification schemes, using cropped images for standard CXR, soft-tissue CXR, and the fusion of both, yielded AUC values on the held-out test set of 0.76 (95% CI: 0.73, 0.79), 0.73 (95% CI: 0.70, 0.76), and 0.78 (95% CI: 0.74, 0.81), respectively. The ROC

curves and comparison results are presented in Fig. 4.6 and Table 4.5. Using soft-tissue CXR yielded a significantly lower AUC value than using standard CXR and using the fusion of both types of CXR. Using the fusion of both types of CXR appeared to achieve a higher AUC value than when using standard CXR alone, but this improvement failed to reach statistical significance, and the performance was statistically equivalent to using standard CXR alone with an equivalence margin of $\Delta\text{AUC} = 0.05$.

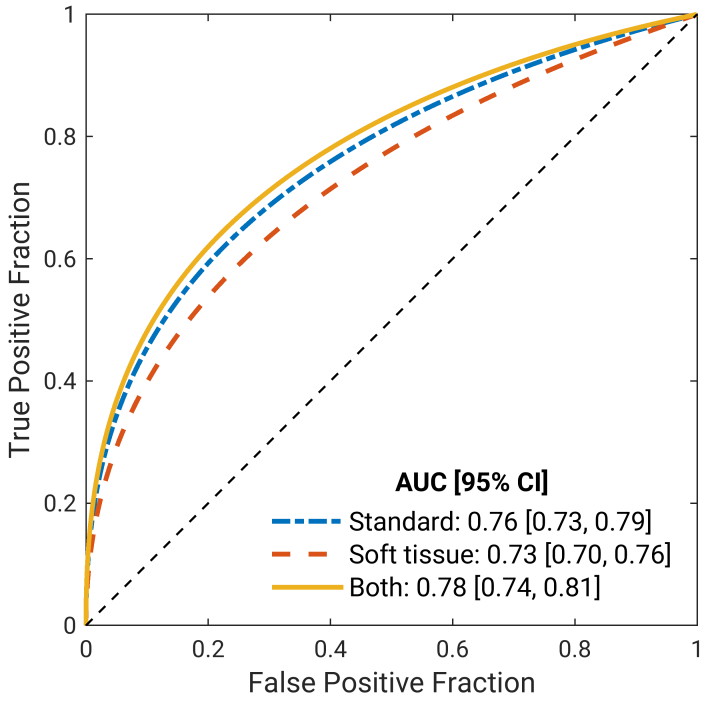


Figure 4.6: Fitted proper binormal ROC curves for the COVID-19 classification task for the held-out test set in the third phase when using cropped standard CXR and/or cropped soft-tissue CXR images.

The desired operating range is in the high sensitivity regime for diagnosing COVID-19 on CXR exams, not only due to the harm of having a false-negative diagnosis for COVID-19, but also because the RT-PCR test has moderate sensitivity and CXR exams are recommended for identifying COVID-19 positive patients who had a false-negative RT-PCR test. Table 4.6 presents the additional evaluation metrics on the held-out test set, including sensitivity, specificity, PPV, NPV, and F1 score at two sensitivity levels, 0.90 and 0.95, for the three

Table 4.5: Comparisons of classification performances using the DeLong test when standard CXR images, soft-tissue images, or fusion of both were used. The p-value and CIs of the difference in AUCs are presented for each comparison. The significance levels (α) and the widths of the confidence intervals are adjusted based on Bonferroni-Holm corrections. Asterisks denote statistical significance after correcting for multiple comparisons.

Comparison	p-value for ΔAUC	α	CI of ΔAUC
Standard vs soft-tissue	0.01*	0.017	98.3% CI: [-0.061, -0.001]
Standard vs Fusion	0.18	0.050	95% CI: [-0.008, 0.041]
soft-tissue vs Fusion	0.02*	0.025	97.5% CI: [0.001, 0.058]

methods in phase 3.

Table 4.6: Additional evaluation metrics for the COVID-19 classification task for the held-out test set in the third phase when using cropped standard CXR and/or cropped soft-tissue CXR images. The metrics are calculated at two sensitivity levels. The 95% CIs are shown in brackets.

Sensitivity	Input	Specificity	PPV	NPV	F1 score
0.95	Standard	0.15 [0.10, 0.26]	0.17 [0.16, 0.19]	0.95 [0.92, 0.97]	0.29 [0.28, 0.32]
	soft-tissue	0.11 [0.07, 0.20]	0.16 [0.16, 0.18]	0.92 [0.89, 0.96]	0.28 [0.27, 0.30]
	Fusion	0.18 [0.14, 0.29]	0.17 [0.17, 0.20]	0.95 [0.94, 0.97]	0.30 [0.29, 0.32]
0.90	Standard	0.30 [0.22, 0.39]	0.19 [0.17, 0.21]	0.94 [0.92, 0.95]	0.31 [0.29, 0.34]
	soft-tissue	0.23 [0.16, 0.33]	0.18 [0.16, 0.20]	0.93 [0.90, 0.95]	0.30 [0.28, 0.32]
	Fusion	0.34 [0.25, 0.43]	0.20 [0.18, 0.22]	0.95 [0.93, 0.96]	0.33 [0.30, 0.36]

Figure 4.7 shows the standard and soft-tissue CXR images in four example cases and their Grad-CAM heatmaps from the penultimate layer of their respective models. These examples were selected to illustrate the differences in model prediction and/or heatmaps that arose when the two types of CXR images were used. In both the positive case (Fig. 4.7a) and the negative case (Fig. 4.7b), using standard CXR images resulted in more accurate predictions, possibly due to undesirable alterations to the anatomy presentation when the soft-tissue images were generated by post-processing algorithms and/or the fact that the datasets used for pretraining do not contain soft-tissue images. On the other hand, the examples in Fig. 4.7c and Fig. 4.7d both show activations in the shoulder region when standard CXR images

were used, whereas in the soft-tissue images the bones were removed and hence did not contribute to the model prediction. In both cases, the soft-tissue model yielded accurate predictions with reasonable activation areas shown in the heatmaps. The influence from the shoulder bones led to a false-positive prediction in the negative case in Fig. 4.7d when standard CXR was used but did not greatly affect the prediction score in the positive case in Fig. 4.7c.

The Bland-Altman plot in Fig. 4.8a has a notable amount of points scattered outside of the $\pm 1.96\,SD$ lines, showing discrepancies between the model predictions based on standard and soft-tissue CXR images. It also shows that predictions, especially for COVID-19 positive cases, spanned a wide range. The patient visit status of COVID-19 positive patients is indicated by different colors. While COVID-19 early diagnosis on CXR scans is challenging in all categories, outpatient cases appear to be more challenging than other categories in our data. This observation is confirmed by Fig. 4.8b, the ROC curves for COVID-19 classification using both standard and soft-tissue CXR combined by feature fusion, presented by patient visit status. The outpatient category yielded the lowest AUC value, and the inpatient category yielded the highest AUC value.

Since both portable exams and DES exams, which generated the soft-tissue images in fundamentally different ways, were included in this study, separate test set evaluation on the portable exam subset and the DES exam subset was performed and is presented in Table 4.7. AUC values were higher for the DES subset than the portable subset when using standard images or the fusion of standard and soft-tissue images, potentially attributed to the higher image quality in DES exams than portable or differences in the patient groups that received these two types of exams. However, AUC values were slightly lower for the DES subset than the portable subset when using just soft-tissue images, potentially because the model was trained mostly on synthetic soft-tissue images as the majority of the dataset was portable exams and thus performed better on this type of images. To study the utility

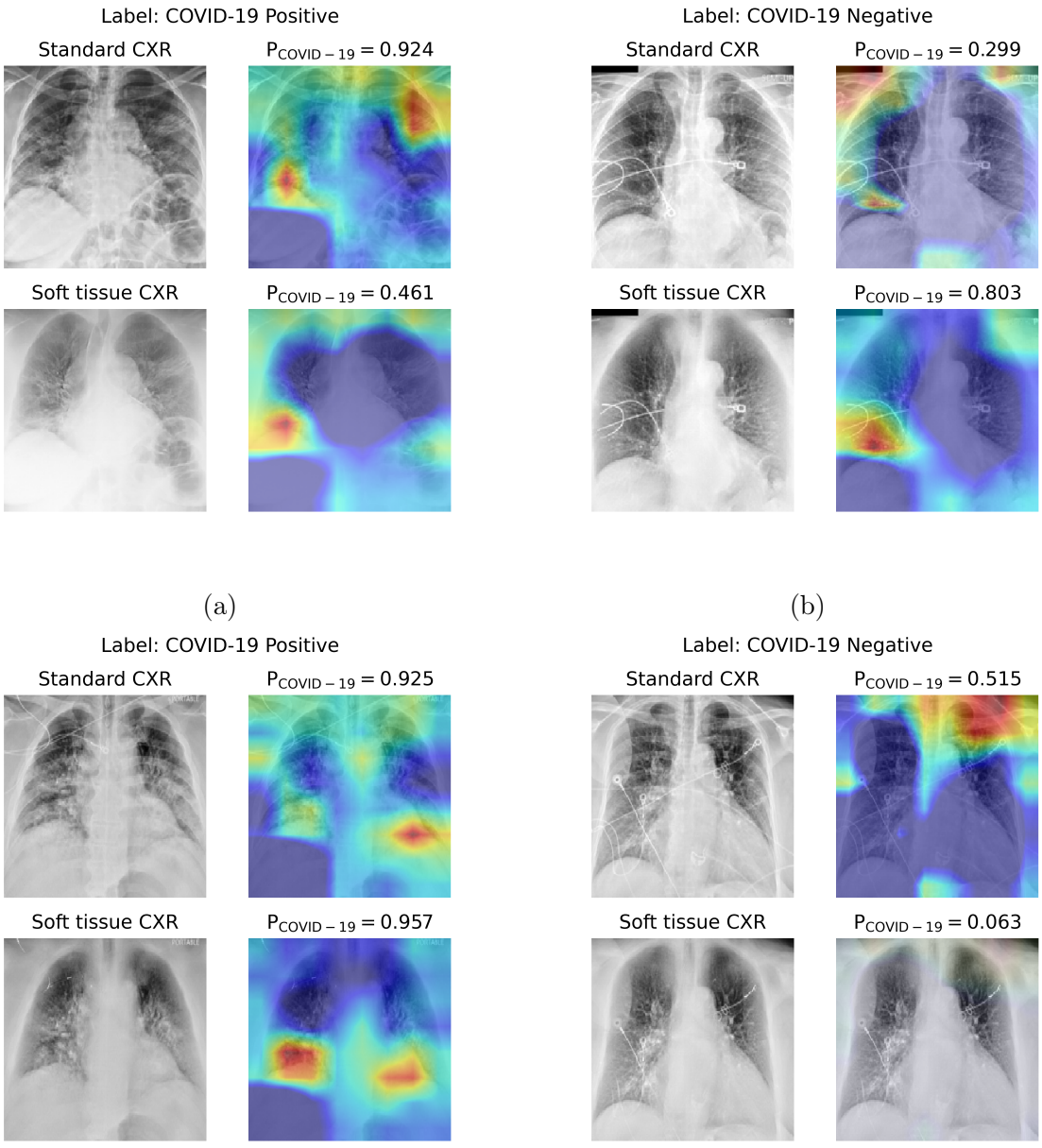
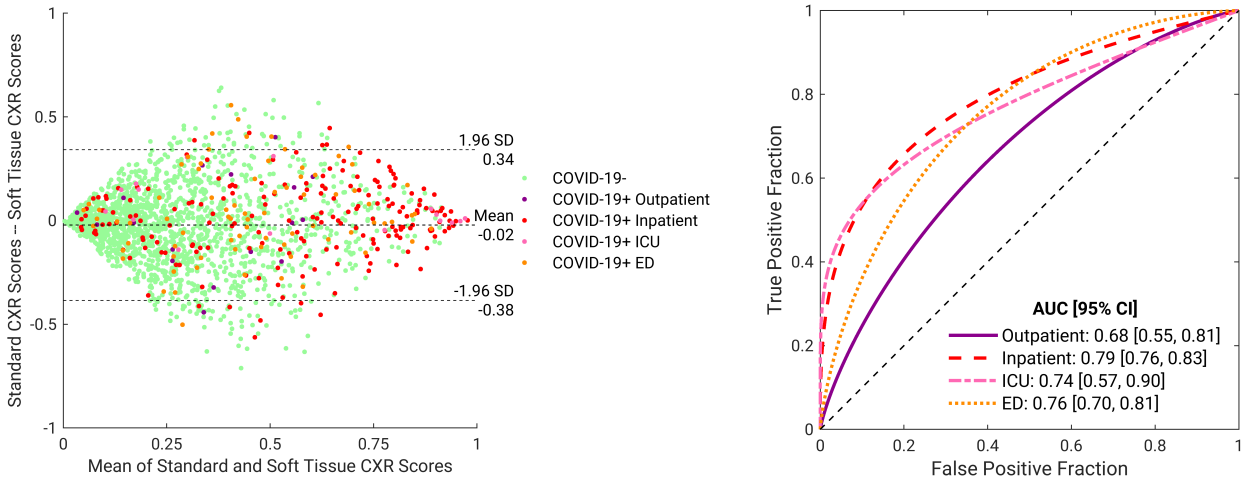


Figure 4.7: Standard and soft-tissue chest radiographs (CXR) of four example cases (post cropping) and their Grad-CAM heatmap overlays. The model prediction scores ($P_{\text{COVID-19}}$) are noted. In all four cases, model predictions and/or heatmaps show differences when the two types of CXR images are used.



(a) (b)

Figure 4.8: (a) Bland-Altman plot for the model predictions based on standard and soft-tissue CXR images. The patient visit status of COVID-19 positive patients are indicated by different colors. (b) ROC curves for COVID-19 classification using both standard and soft-tissue CXR combined by feature fusion, presented by patient visit status. ICU = intensive care unit, ED = emergency department.

and the contribution of soft-tissue images in these two types of CXR exams in COVID-19-related image interpretation, separate analyses and controlled experiments are needed in future work.

Table 4.7: COVID-19 classification performance by CXR exam type (portable or dual-energy subtraction [DES] exam). The 95% CIs are shown in brackets.

		Portable (80%)	DES (20%)	Overall
COVID-19 prevalence		16%	12%	15%
AUC	Standard	0.74 [0.70, 0.78]	0.86 [0.80, 0.91]	0.76 [0.73, 0.79]
	soft-tissue	0.73 [0.62, 0.80]	0.71 [0.70, 0.77]	0.73 [0.70, 0.76]
	Fusion	0.77 [0.73, 0.80]	0.83 [0.77, 0.89]	0.78 [0.74, 0.81]

4.5.2 Prognosis

The ROC curves for predicting COVID-19 patients' potential need for intensive care in 24, 48, 72, and 96 hours are shown in Fig. 4.9. The highest AUC [95% CI] of 0.77 [0.70, 0.84] was

achieved when predicting 24 hours in advance. Promising performances were also achieved when using earlier CXR for predictions: 0.73 [0.66, 0.80], 0.74 [0.67, 0.80], and 0.74 [0.67, 0.80] when predicting 48, 72, and 96 hours in advance, respectively.

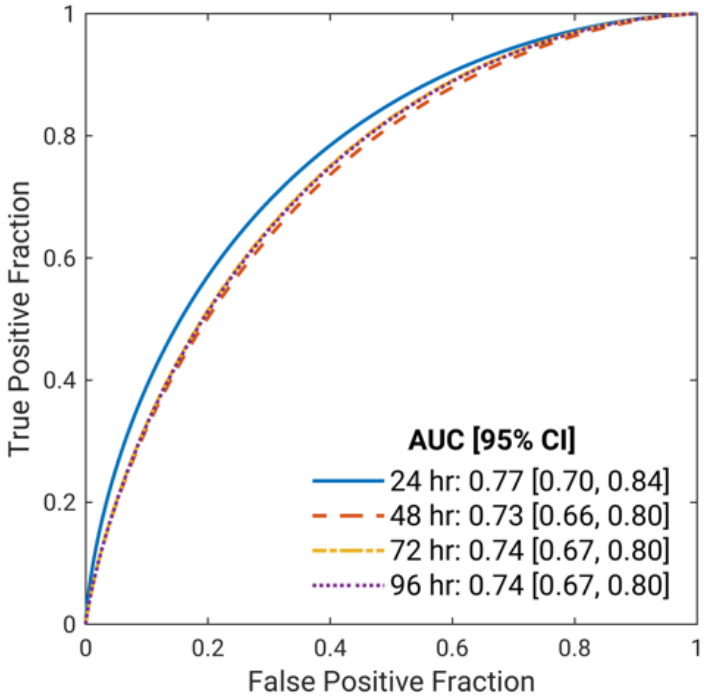
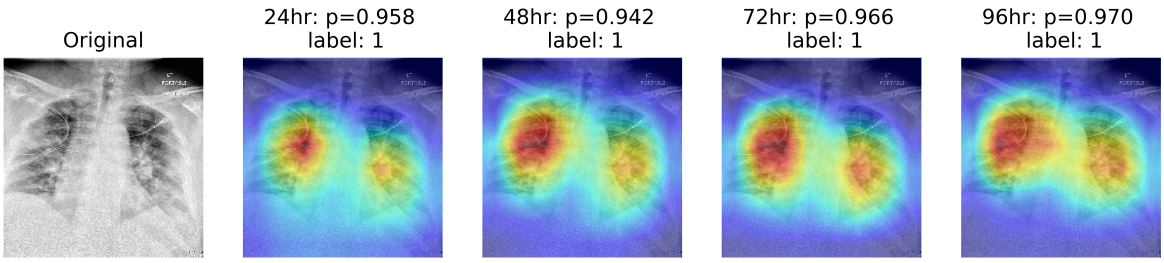
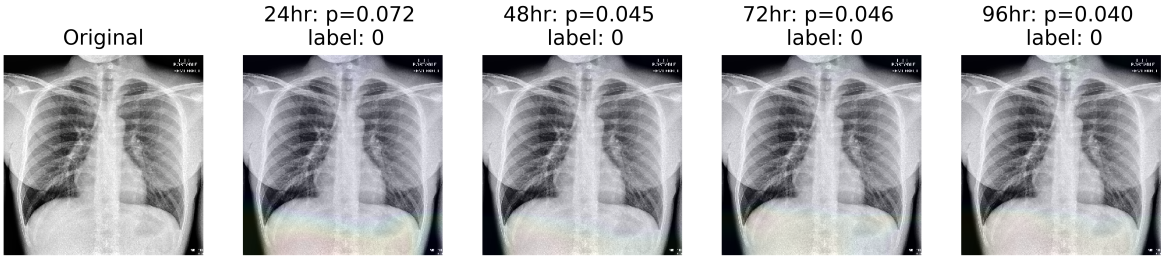


Figure 4.9: Fitted binormal ROC curves for classification tasks requiring intensive care or not within 24–96 hours from image acquisition. The legend gives the AUC with 95% CI for each task.

Figure 4.10 shows two examples, each with the original CXR image and the Grad-CAM heatmaps from the last batch normalization layer of the model overlaid on the CXR image. The patient shown in Fig. 4.10a was COVID-19 positive and was admitted to the ICU within 24 hours following the image shown here. The patient shown in Fig. 4.10b was COVID-19 negative and did not receive intensive care after the image shown here. The predictions for the four labels in both of these cases were accurate, and abnormalities in the lungs are highlighted by the Grad-CAM heatmaps.



(a)



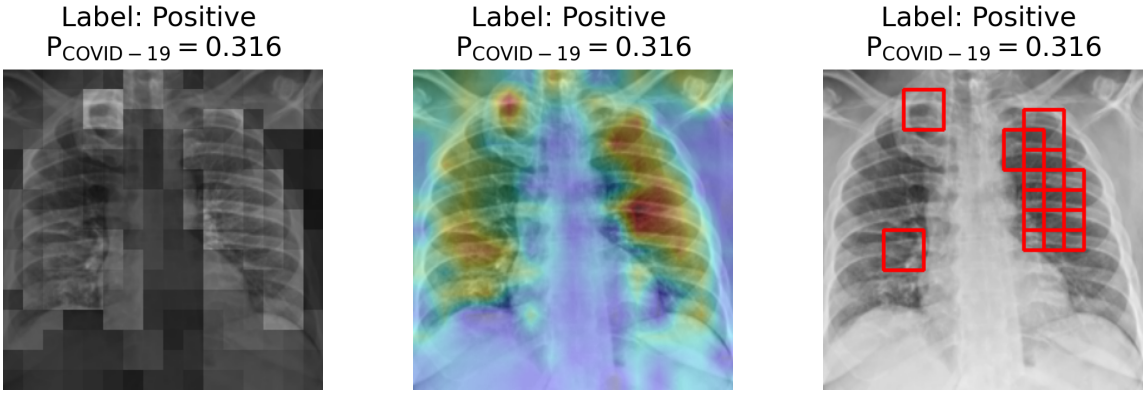
(b)

Figure 4.10: Two examples of portable chest radiography images overlaid with their Grad-CAM heatmaps for intensive care need prediction within 24, 48, 72, and 96 hours, respectively. The patient in the top example was admitted into ICU 4 hours after the image was acquired. The patient in the bottom example has not been hospitalized since the image was obtained.

4.5.3 Classification and Visualization using Multiple Instance Learning

The attention-based deep MIL was not able to yield high classification performance in the task of COVID-19 classification on initial CXR exams. The best performance was achieved when the DenseNet-121 model was used as the backbone CNN and the trained weights on cropped CXR images (as opposed to patches) from Section 4.5.1 were loaded and frozen; that is, only the MIL attention pooling mechanism and the final classification layers were updated during training. Training or fine-tuning a larger portion of the model did not improve the classification performance. A moderate AUC [95% CI] of 0.65 [0.62, 0.68] was achieved under this training scheme. A representative example of visualization heatmaps in three formats generated from the extracted attention scores is shown in Fig. 4.11. The classification prediction in this example was not accurate, and the highlighted locations,

while mostly within the lung region, only partially corresponded to abnormalities. Among all heatmaps generated on the test set, false-positive activations, i.e., highlighted areas that are not abnormalities related to COVID-19, included air space, clavicles, ribs, and medical devices.



(a) (b) (c)

Figure 4.11: Example heatmaps created by attention-based deep multiple instance learning algorithm in the task of COVID-19 classification on initial CXR exam.

4.6 Discussion and Conclusions

In summary, a large CXR database has been curated during the COVID-19 pandemic, and a sequential transfer learning curriculum was designed to pretrain and fine-tune a model on increasingly specific and complex tasks with the final goal of distinguishing COVID-19 positive and negative patients using their initial CXR exam within two days of their initial RT-PCR test for COVID-19 and predicting whether a COVID-19 positive patient will potentially need intensive care in the next one to four days [19]. Automatic lung segmentation and cropping were incorporated in the classification pipeline to reduce the influence of irrelevant regions of the images on model predictions. The role of soft-tissue images in CXR exams was investigated in addition to the standard CXR images. An AUC [95% CI] of 0.78 [0.74, 0.81] was achieved using both types of CXR images combined via feature fusion

for diagnosing COVID-19 on CXR at patient presentation, which was equivalent to the performance obtained when using the standard CXR alone. In the task of predicting whether COVID-19-positive patients would require intensive care based on CXR images, an AUC [95% CI] of 0.77 [0.70, 0.84] was achieved when predicting 24 hours in advance, and at least 0.73 [0.66, 0.80] when using earlier CXR for predictions.

Besides using feature fusion with shared weights for the combined use of the two types of CXR images (described in Section 4.3.1), two alternative fusion methods were explored, and they achieved similar results as the feature fusion method. Averaging the prediction scores given by the standard CXR model and the soft-tissue CXR model for each case yielded an AUC of 0.76 (95% CI: 0.73, 0.79). Feature fusion without weight sharing between the two parallel models achieved an AUC of 0.77 (95% CI: 0.73, 0.80) and was more computationally expensive with twice as many parameters in the models as when weight sharing was employed.

It is worth noting that all patients in our dataset had medical indications for receiving a CXR exam, such as being symptomatic for possible COVID-19, receiving medical care related to pneumonia of unknown or known origin, or undergoing diagnosis or treatment for other diseases. As such, our dataset does not represent the entire population that undergoes RT-PCR testing for COVID-19. Also, some COVID-19 positive patients might have received intensive care due to other diseases they had, which was not always indicated in the clinical information available. These limitations increased the difficulty of our task. Many patients in our dataset indeed presented with non-COVID lung abnormalities, which made the distinction between COVID and non-COVID patients more challenging and added confounding factors to the intensive care prediction for COVID-19 patients than if most of the non-COVID patients had been healthy and had presented with no abnormal lung findings.

Another important note is that the pre-pandemic public databases used for pretraining in this study only contain standard CXR images, which might partly contribute to the inferior performance of soft-tissue images in the COVID-19 classification stage. When the

first and second pretraining phases are removed, standard images, soft-tissue images, and the fusion of both yielded similar AUC values in COVID-19 classification (0.73 [0.69, 0.67], 0.72 [0.69, 0.75], and 0.72 [0.68, 0.75]). The results show that pretraining stages led to larger improvements in AUC for standard images than soft-tissue images and significantly improved the fusion model's performance ($P < .001$).

While MIL did not demonstrate improvement in the classification of COVID-19 status or visualization of COVID-19 abnormalities on early CXR images, the technique may be useful in future studies for CXR image analysis on a different task. The task of recognizing evidence of COVID-19 on patients' initial CXR was very challenging, as some COVID-19 positive patients could have had no abnormalities presented in the lungs at the time of their initial CXR exams. The local patch-level analysis performed by MIL can also be combined with the global analysis of full CXR images in a multi-scale algorithm that can potentially enrich the features learned and improve the classification performance. The observations in this study showed MIL's potential utility for improved localization without pixel-level annotations by experts, as well as higher explainability and interpretability of deep-learning-based CAD methods.

There are several other limitations of this study. Firstly, the database was collected from a single institution. Medical centers including ours have been contributing to a multi-institutional databases, and independent evaluations should be perform when such datasets become available in the future to assess the robustness of the approach developed in this research. Such high-quality publicly available databases will also allow the research community to establish reference standards and compare performances. Secondly, training was performed on a combination of both DES CXR exams and portable CXR exams in our database. The soft-tissue image in a DES exam is obtained from two physically acquired images, while the synthetic soft-tissue image in a portable exam is generated from the standard image using post-processing algorithms. Previous work has shown reduced clinical utility of

synthetic soft-tissue images relative to DES soft-tissue images [148]. Preliminary findings reported in Section 4.5.1 did show notable differences in the model performance on these two subsets. Future work will evaluate the difference in utility and contribution of each type of soft-tissue image when using deep-learning-based methods in tasks such as COVID-19 detection, classification, and prognosis. Thirdly, a single CXR exam at patient presentation was used to diagnose evidence of COVID-19 for each patient. Performing temporal analysis to utilize previous CXR exams of suspected patients, instead of only using images at a single time point, may improve the model performance. Finally, while the Grad-CAM technique, which we used to visualize and explain model predictions, is one of the commonly used explainability techniques for convolutional neural networks, the created heatmaps are not intended for precise localization tasks, especially in the medical imaging domain. Future studies will investigate methods to more precisely localize COVID-19 presentations on CXR images.

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

This dissertation contributes to the advancement of AI-assisted medical image analysis in the context of breast cancer computer-aided diagnosis (CADx) based on multiparametric MRI (mpMRI) and COVID-19 early diagnosis and prognosis based on chest radiography (CXR). Methods to utilize and integrate information from multiple sequences in mpMRI were investigated for both human-engineered radiomic features and deep learning methods to improve the classification performance in the task of distinguishing benign and malignant breast lesions. Furthermore, deep learning methods that could efficiently and effectively leverage high-dimensional information in mpMRI were investigated, further advancing the capability of differential diagnosis for breast lesions on mpMRI. With the onset and development of the COVID-19 pandemic, this research also developed deep learning methods for automated COVID-19 diagnosis at patient presentation and for predicting patients' need for intensive care using CXR exams.

Chapter 2 proposed and evaluated radiomics methods that could leverage the complementary information provided by the DCE, T2w, and DWI sequences in mpMRI and integrate the multiple sequences to collectively improve performance over single-sequence radiomics methods in the task of distinguishing between benign and malignant breast lesions. The study was performed on both human-engineered radiomic features and features extracted by pretrained CNN models. Three mpMRI fusion approaches were proposed and evaluated: image fusion, i.e., fusing images from multiple MRI sequences into an RGB image to form the input to the CNN (for CNN-based methods only); feature fusion, i.e., concatenating features extracted from mpMRI sequences to form the classifier input; and classifier fusion, i.e., aggregating the probability of malignancy output scores from single-sequence classifiers via soft voting. These fusion methods integrated information derived from mpMRI sequences at three different levels in the image analysis process. When human-engineered features were

used, the feature fusion and classifier fusion methods were equivalent, and both achieved significantly higher classification performance than using any single sequence alone. When CNN features were used, the feature fusion method significantly outperformed using the DCE sequence alone, demonstrating superiority to the other fusion methods. These findings can potentially improve the current breast cancer CADx systems based on DCE-MRI.

Chapter 3 investigated methods to effectively utilize the high-dimensional information inherent in MRI exams when using deep transfer learning in the task of distinguishing between benign and malignant breast lesions. The feature MIP method, which globally max pools the features extracted from a lesion volume along the lesion's axial dimension within a CNN, was proposed for incorporating volumetric information in MRI and demonstrated superiority to the previously proposed image MIP method. For 4D sequences, namely DCE and DWI, the RGB channels of CNNs pretrained on natural images were utilized to incorporate the images acquired at different time points in DCE and at different diffusion weighting strengths in DWI. Applying the feature MIP method to three sequences in mpMRI and the feature fusion method from Chapter 2 to combine information extracted from the sequences, the high-dimensional mpMRI classifier achieved high classification performance that significantly outperformed using any single sequence alone. The method presented in this chapter can potentially enhance the performance of current deep-learning-based CADx systems for breast cancer differential diagnosis by addressing the problem of underutilizing high-dimensional information in medical images while maintaining reasonable computing intensity by using transfer learning.

Chapter 4 developed a deep learning model on a large CXR database curated during the COVID-19 pandemic to diagnose COVID-19 at patient presentation and predict patients' needs for intensive care. The model was sequentially pretrained and fine-tuned on increasingly specific and complex tasks, following a learning curriculum, with the final goal of COVID-19 diagnosis and prognosis. Automatic lung segmentation and cropping were in-

corporated in the classification pipeline and were shown to reduce the influence of irrelevant regions of the images on model predictions. The role of soft tissue images in CXR exams was investigated in addition to the standard CXR images. The utility of multiple instance learning was also examined for the classification of COVID-19 status and visualization.

Future research can expand upon this work and address the current limitations. The limitations and suggested future work are summarized as follows. All studies in this work were developed and evaluated on datasets from single institutions. Future work can expand the database to include images from multiple medical centers and diverse populations and evaluate the robustness and generalizability of the proposed methods across manufacturers, image acquisition protocols, and patient populations. Besides independent validation and generalizability evaluation, such multi-institutional datasets can also be used to investigate harmonization methods for different acquisition parameters.

Limitations to the newly curated COVID-19 datasets and truth labels were identified as we gained more understanding of this novel disease. The ground truth for COVID-19 positive or negative status was provided by single RT-PCR tests. It is known now that the RT-PCR test has moderate sensitivity, and repeated testing is recommended. Therefore, the ground truth can be revised in future work based on results from repeated RT-PCR tests when they are available. Moreover, the prognosis study was not able to be performed on a larger, more recently updated dataset because of ambiguity in the ground truth. To generate high-quality ground truth, it was necessary to extract information from clinical files in order to accurately estimate the time of intubation and ICU admission and to exclude patients who tested positive for COVID-19 but were intubated or admitted into the ICU for reasons not related to COVID-19. This process, however, was performed manually for each patient, which was not feasible to scale to much larger datasets. In future work, an automated process can be developed to extract relevant clinical information that helps generate ground truth labels. In addition, the soft tissue images in the dataset were from either dual-energy CXR

exams or portable CXR exams. The conclusion on the utility of soft tissue images may differ between these two subsets, but this study was not able to draw such conclusions due to confounding variables. Future work can continue to expand the dataset and investigate the contribution of each type of soft tissue image separately in COVID-19 patient management. Finally, while patients' previous CXR exams up to a year prior to their first RT-PCR tests were also curated in this database, predictions were based on single CXR exams in this work. Performing temporal analysis to utilize patients' previous CXR exams, instead of only using images at a single time point, may improve the diagnostic and prognostic performance and can be explored in the future.

There are a few limitations in the evaluation that can be addressed in future studies. First, the operating points at which sensitivity and specificity are reported were based on certain assumptions that may not align with the potential clinical scenario to which the algorithms may be applied. Without sufficient knowledge about the specific clinical use case, the selected optimal operating point may not be clinically optimal. A different threshold might be chosen, for example, if the relative cost of false-positive and false-negative diagnoses were known. Future studies can be conducted to evaluate the computational methods developed in this dissertation in specific clinical use cases. Similarly, the choice of equivalence margin was not a predetermined, clinically meaningful value either, since there are currently no widely used standards for establishing the equivalence margin in diagnostic performance studies. The methods can be evaluated with revised equivalence margins in the future when relevant guidelines are available. Moreover, while this dissertation is focused on the computational aspect of improving the standalone performance of diagnosis algorithms, reader studies can be performed in the future to assess the clinical significance of the algorithms when used as a secondary or concurrent reader for clinicians.

While this dissertation is focused on image analysis, many clinical tasks this research aims to achieve also benefit from other forms of data, such as demographic information,

vital signs, laboratory tests, symptoms, and electronic medical records. These data were intentionally excluded from the development of image analysis algorithms in this research, because they would have overshadowed the features that could be learned from images. For example, if patient age was provided as an additional variable to the classifier, then age may become the dominant feature that overwhelms the classifier's decision boundary, and little can be learned about the utility of various imaging features. Nonetheless, once image analysis algorithms have been developed, they can be combined with other forms of data by a multimodal algorithm to make comprehensive clinical assessments, a strategy that can be pursued in future work.

This dissertation demonstrates the strong potential of AI-assisted medical image analysis, which may enhance the accuracy and efficiency of radiologists' image interpretation in the future. Human-engineered and deep-learning-based radiomics can contribute to "virtual digital biopsy," allowing for the assessment of breast lesion malignancy when biopsies are not practical or necessary. Deep learning methods also have the potential to assist in image interpretation for COVID-19 diagnosis and preempt patient deterioration, helping healthcare professionals understand this novel disease and navigate the pandemic. Overall, this work demonstrates the capacity of AI applications in medical image analysis, which may ultimately contribute to higher quality healthcare and universal access to radiology expertise.

REFERENCES

[1] Maryellen L Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018.

[2] Bethany L Niell, Phoebe E Freer, Robert Jared Weinfurtner, Elizabeth Kagan Arleo, and Jennifer S Drukteinis. Screening for breast cancer. *Radiologic Clinics*, 55(6): 1145–1162, 2017.

[3] Ritse M Mann, Nariya Cho, and Linda Moy. Breast mri: state of the art. *Radiology*, 292(3):520–536, 2019.

[4] Edward A Sickles, Carl J D'Orsi, Lawrence W Bassett, Catherine M Appleton, Wendie A Berg, Elizabeth S Burnside, et al. Acr bi-rads® atlas, breast imaging reporting and data system. *Reston, VA: American College of Radiology*, pages 39–48, 2013.

[5] Habib Rahbar and Savannah C Partridge. Multiparametric mr imaging of breast cancer. *Magnetic Resonance Imaging Clinics*, 24(1):223–238, 2016.

[6] Savannah C Partridge and Elizabeth S McDonald. Diffusion weighted magnetic resonance imaging of the breast: protocol optimization, interpretation, and clinical applications. *Magnetic Resonance Imaging Clinics*, 21(3):601–624, 2013.

[7] Janet E Kuhlman, Jannette Collins, Gregory N Brooks, Donald R Yandow, and Lynn S Broderick. Dual-energy subtraction chest radiography: what to look for beyond calcified nodules. *Radiographics*, 26(1):79–92, 2006.

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[12] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[13] Qiyuan Hu, Heather M Whitney, and Maryellen L Giger. A deep learning methodology for improved breast cancer diagnosis using multiparametric mri. *Scientific reports*, 10 (1):1–11, 2020.

[14] Qiyuan Hu, Heather M Whitney, and Maryellen L Giger. Radiomics methodology for breast cancer diagnosis using multiparametric magnetic resonance imaging. *Journal of Medical Imaging*, 7(4):044502, 2020.

[15] Qiyuan Hu, Heather M Whitney, and Maryellen L Giger. Using resnet feature extraction in computer-aided diagnosis of breast cancer on 927 lesions imaged with multiparametric mri. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 1131411. International Society for Optics and Photonics, 2020.

[16] Qiyuan Hu, Heather M Whitney, Alexandra Edwards, John Papaioannou, and Maryellen L Giger. Radiomics and deep learning of diffusion-weighted mri in the diagnosis of breast cancer. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 109504A. International Society for Optics and Photonics, 2019.

[17] Qiyuan Hu, Heather M Whitney, Hui Li, Yu Ji, Peifang Liu, and Maryellen L Giger. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection mri in breast cancer diagnosis using dynamic contrast-enhanced mri. *Radiology: Artificial Intelligence*, 3(3):e200159, 2021.

[18] Qiyuan Hu, Heather M Whitney, and Maryellen L Giger. Transfer learning in 4d for breast cancer diagnosis using dynamic contrast-enhanced magnetic resonance imaging. *arXiv preprint arXiv:1911.03022*, 2019.

[19] Qiyuan Hu, Karen Drukker, and Maryellen L Giger. Role of standard and soft tissue chest radiography images in covid-19 diagnosis using deep learning. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, page 1159704. International Society for Optics and Photonics, 2021.

[20] Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017.

[21] Maryellen L Giger, Heang-Ping Chan, and John Boone. Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm. *Medical physics*, 35(12):5799–5820, 2008.

[22] Stephen SF Yip and Hugo JWL Aerts. Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150, 2016.

[23] Issam El Naqa, Masoom A Haider, Maryellen L Giger, and Randall K Ten Haken. Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. *The British journal of radiology*, 93(1106):20190855, 2020.

[24] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021.

[25] Kenneth GA Gilhuijs, Maryellen L Giger, and Ulrich Bick. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Medical physics*, 25(9):1647–1654, 1998.

[26] Weijie Chen, Maryellen L Giger, Li Lan, and Ulrich Bick. Computerized interpretation of breast mri: investigation of enhancement-variance dynamics. *Medical physics*, 31 (5):1076–1082, 2004.

[27] Weijie Chen, Maryellen L Giger, Ulrich Bick, and Gillian M Newstead. Automatic identification and classification of characteristic kinetic curves of breast lesions on dce-mri. *Medical physics*, 33(8):2878–2887, 2006.

[28] Weijie Chen, Maryellen L Giger, Hui Li, Ulrich Bick, and Gillian M Newstead. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(3):562–571, 2007.

[29] Weijie Chen, Maryellen L Giger, Gillian M Newstead, Ulrich Bick, Sanaz A Jansen, Hui Li, and Li Lan. Computerized assessment of breast lesion malignancy using dce-mri: robustness study on two independent clinical datasets from two manufacturers. *Academic radiology*, 17(7):822–829, 2010.

[30] Maryellen L Giger, Nico Karssemeijer, and Julia A Schnabel. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annual review of biomedical engineering*, 15:327–357, 2013.

[31] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[32] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.07.005. URL `http://www.sciencedirect.com/science/article/pii/S1361841517301135`.

[33] Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M Summers, and Maryellen L Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[35] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

[36] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[37] Daniel Truhn, Simone Schrading, Christoph Haarburger, Hannah Schneider, Dorit Merhof, and Christiane Kuhl. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri. *Radiology*, 290(2):290–297, 2019.

[38] Heather M. Whitney, Hui Li, Yu Ji, Peifang Liu, and Maryellen L. Giger. Comparison of breast mri tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods. *Proceedings of the IEEE*, 108(1):163–177, 2020. doi: 10.1109/JPROC.2019.2950187.

[39] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[40] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: a Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[41] Martha B Mainiero, Linda Moy, Paul Baron, Aarati D Didwania, Edward D Green, Samantha L Heller, Anna I Holbrook, Su-Ju Lee, Alana A Lewin, Ana P Lourenco, et al. Acr appropriateness criteria® breast cancer screening. *Journal of the American College of Radiology*, 14(11):S383–S390, 2017.

[42] Heidi D Nelson, Ellen S O'Meara, Karla Kerlikowske, Steven Balch, and Diana Miglioretti. Factors associated with rates of false-positive and false-negative results from digital mammography screening: an analysis of registry data. *Annals of internal medicine*, 164(4):226–235, 2016.

[43] Debra L Monticciolo, Mary S Newell, Linda Moy, Bethany Niell, Barbara Monsees, and Edward A Sickles. Breast cancer screening in women at higher-than-average risk: recommendations from the acr. *Journal of the American College of Radiology*, 15(3): 408–414, 2018.

[44] Christiane K Kuhl, Kevin Strobel, Heribert Bieling, Claudia Leutner, Hans H Schild, and Simone Schrading. Supplemental breast mr imaging screening of women with average risk of breast cancer. *Radiology*, 283(2):361–370, 2017.

[45] Dafydd Gareth Evans, Elaine F Harkness, Anthony Howell, Mary Wilson, Emma Hurley, Marit Muri Holmen, KU Tharmaratnam, Anne Irene Hagen, Y Lim, Anthony James Maxwell, et al. Intensive breast screening in brca2 mutation carriers is associated with reduced breast cancer specific and all cause mortality. *Hereditary cancer in clinical practice*, 14(1):1–8, 2016.

[46] Christiane K Kuhl, Simone Schrading, Kevin Strobel, Hans H Schild, Ralf-Dieter Hilgers, and Heribert B Bieling. 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(22): 2304–2310, 2014.

[47] Doris Leithner, Linda Moy, Elizabeth A Morris, Maria A Marino, Thomas H Helbich, and Katja Pinker. Abbreviated mri of the breast: does it provide value? *Journal of Magnetic Resonance Imaging*, 49(7):e85–e100, 2019.

[48] Jan CM van Zelst, Suzan Vreemann, Hans-Joerg Witt, Albert Gubern-Merida, Monique D Dorrius, Katya Duvivier, Susanne Lardenoije-Broker, Marc BI Lobbes, Claudette Loo, Wouter Veldhuis, et al. Multireader study on the diagnostic accuracy of ultrafast breast magnetic resonance imaging for breast cancer screening. *Investigative radiology*, 53(10):579–586, 2018.

[49] KH Haraldsdottir, Þ Jónsson, AB Halldórsdóttir, K-G Tranberg, and KS Ásgeirsson. Tumor size of invasive breast cancer on magnetic resonance imaging and conventional imaging (mammogram/ultrasound): comparison with pathological size and clinical implications. *Scandinavian Journal of Surgery*, 106(1):68–73, 2017.

[50] Eun Young Yoo, Sang Yu Nam, Hye-Young Choi, and Min Ji Hong. Agreement between mri and pathologic analyses for determination of tumor size and correlation with immunohistochemical factors of invasive breast carcinoma. *Acta radiologica*, 59 (1):50–57, 2018.

[51] Muhammad Asad Parvaiz, Peiming Yang, Eisha Razia, Margaret Mascarenhas, Caroline Deacon, Pilar Matey, Brian Isgar, and Tapan Sircar. Breast mri in invasive lobular carcinoma: a useful investigation in surgical planning? *The breast journal*, 22 (2):143–150, 2016.

[52] Valeria Selvi, Jacopo Nori, Icro Meattini, Giulio Francolini, Noemi Morelli, Diego Di Benedetto, Giulia Bicchierai, Federica Di Naro, Maninderpal Kaur Gill, Lorenzo Orzalesi, et al. Role of magnetic resonance imaging in the preoperative staging and work-up of patients affected by invasive lobular carcinoma or invasive ductolobular carcinoma. *BioMed research international*, 2018, 2018.

[53] Christiane K Kuhl, Kevin Strobel, Heribert Bieling, Eva Wardelmann, Walther Kuhn, Nikolaus Maass, and Simone Schrading. Impact of preoperative breast mr imaging and mr-guided surgery on diagnosis and surgical outcome of women with invasive breast cancer with and without dcis component. *Radiology*, 284(3):645–655, 2017.

[54] María Nieves Plana, Carmen Carreira, Alfonso Muriel, Miguel Chiva, Víctor Abraira, Jose Ignacio Emparanza, Xavier Bonfill, and Javier Zamora. Magnetic resonance imaging in the preoperative assessment of patients with primary breast cancer: systematic review of diagnostic accuracy and meta-analysis. *European radiology*, 22(1):26–38, 2012.

[55] Ritse M Mann, Claudette E Loo, Theo Wobbes, Peter Bult, Jelle O Barentsz, Kenneth GA Gilhuijs, and Carla Boetes. The impact of preoperative breast mri on the re-excision rate in invasive lobular carcinoma of the breast. *Breast cancer research and treatment*, 119(2):415, 2010.

[56] Marc BI Lobbes, Ingeborg JH Vriens, Annelotte CM van Bommel, Grard AP Nieuwenhuijzen, Marjolein L Smidt, Liesbeth J Boersma, Thijs van Dalen, Carolien Smorenburg, Henk Struikmans, Sabine Siesling, et al. Breast mri increases the number of mastectomies for ductal cancers, but decreases them for lobular cancers. *Breast cancer research and treatment*, 162(2):353–364, 2017.

[57] Michael L Marinovich, Nehmat Houssami, Petra Macaskill, Francesco Sardanelli, Les Irwig, Eleftherios P Mamounas, Gunter Von Minckwitz, Meagan E Brennan, and Stefano Ciatto. Meta-analysis of magnetic resonance imaging in detecting residual breast cancer after neoadjuvant therapy. *Journal of the National Cancer Institute*, 105(5): 321–333, 2013.

[58] John R Scheel, Eunhee Kim, Savannah C Partridge, Constance D Lehman, Mark A Rosen, Wanda K Bernreuter, Etta D Pisano, Helga S Marques, Elizabeth A Morris, Paul T Weatherall, et al. Mri, clinical examination, and mammography for preoperative assessment of residual disease and pathologic complete response after neoadjuvant chemotherapy for breast cancer: Acrin 6657 trial. *American Journal of Roentgenology*, pages 1376–1385, 2018.

[59] D Leithner, GJ Wengert, TH Helbich, S Thakur, RE Ochoa-Albiztegui, EA Morris, and K Pinker. Clinical role of breast mri now and going forward. *Clinical radiology*, 73(8):700–714, 2018.

[60] Luca A Carbonaro, Federica Pediconi, Nicola Verardi, Rubina M Trimboli, Massimo Calabrese, and Francesco Sardanelli. Breast mri using a high-relaxivity contrast agent: an overview. *American journal of roentgenology*, 196(4):942–955, 2011.

[61] John C Lindon, George E Tranter, and David Koppenaal. *Encyclopedia of spectroscopy and spectrometry*. Academic Press, 2016.

[62] Humairah S Cheung, M Tse Gary, Shuk-Yee Lai, and David KW Yeung. Relationship between lesion size and signal enhancement on subtraction fat-suppressed mr imaging of the breast. *Magnetic resonance imaging*, 22(9):1259–1264, 2004.

[63] Christiane Katharina Kuhl, Sven Klaschik, Peter Mielcarek, Jürgen Gieseke, Eva Wardelmann, and Hans H Schild. Do t2-weighted pulse sequences help with the differential diagnosis of enhancing lesions in dynamic breast mri? *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 9(2):187–196, 1999.

[64] Christine Westra, Vandana Dialani, Tejas S Mehta, and Ronald L Eisenberg. Using t2-weighted sequences to more accurately characterize breast masses seen on mri. *American Journal of Roentgenology*, 202(3):W183–W190, 2014.

[65] Otso Arponen, Amro Masarwah, Anna Sutela, Mikko Taina, Mervi Könönen, Reijo Sironen, Juhana Hakumäki, Ritva Vanninen, and Mazen Sudah. Incidentally detected enhancing lesions found in breast mri: analysis of apparent diffusion coefficient and t2 signal intensity significantly improves specificity. *European radiology*, 26(12):4361–4370, 2016.

[66] Denis Le Bihan and Mami Iima. Diffusion magnetic resonance imaging: what water tells us about biological tissues. *PLoS biology*, 13(7):e1002203, 2015.

[67] Ruo-yang Shi, Qiu-ying Yao, Lian-ming Wu, and Jian-rong Xu. Breast lesions: diagnosis using diffusion weighted imaging at 1.5 t and 3.0 t—systematic review and meta-analysis. *Clinical breast cancer*, 18(3):e305–e320, 2018.

[68] Scientific brief: Sars-cov-2 transmission. `https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/sars-cov-2-transmission.html`, 2021. Accessed: 2021-06-07.

[69] Jessica Watson, Penny F Whiting, and John E Brush. Interpreting a covid-19 test result. *Bmj*, 369, 2020.

[70] Fred A Mettler Jr, Walter Huda, Terry T Yoshizumi, and Mahadevappa Mahesh. Effective doses in radiology and diagnostic nuclear medicine: a catalog. *Radiology*, 248 (1):254–263, 2008.

[71] American College of Radiology et al. Acr-spr-str practice parameter for the performance of portable (mobile unit) chest radiography. *Am Coll Radiol*, 1076:1–8, 2017.

[72] Julio Lemos and Jeffrey S. Klein. Methods of examination, normal anatomy, and radiographic findings of chest disease. In William E Brant and Clyde A Helms, editors, *Fundamentals of diagnostic radiology*, chapter 12, pages 324–366. Lippincott Williams & Wilkins, 2012.

[73] Farheen Manji, Jiheng Wang, Geoff Norman, Zhou Wang, and David Koff. Comparison of dual energy subtraction chest radiography and traditional chest x-rays in the detection of pulmonary nodules. *Quantitative imaging in medicine and surgery*, 6(1): 1, 2016.

[74] Matthew Thomas Freedman, Shih-Chung Benedict Lo, John C Seibel, and Christina M Bromley. Lung nodules: improved detection with software that suppresses the rib and clavicle on chest radiographs. *Radiology*, 260(1):265–273, 2011.

[75] Geoffrey D Rubin, Christopher J Ryerson, Linda B Haramati, Nicola Sverzellati, Jeffrey P Kanne, Suhail Raoof, Neil W Schluger, Annalisa Volpi, Jae-Joon Yim, Ian BK

Martin, et al. The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society. *Chest*, 158 (1):106–116, 2020.

[76] ACR Radiology. Acr recommendations for the use of chest radiography and computed tomography (ct) for suspected covid-19. infection. *ACR website.*, 2020.

[77] Adam Jacobi, Michael Chung, Adam Bernheim, and Corey Eber. Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review. *Clinical imaging*, 2020.

[78] Ming-Yen Ng, Elaine YP Lee, Jin Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, et al. Imaging profile of the covid-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic Imaging*, 2(1):e200034, 2020.

[79] Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho-Tung Fong, Siu Ting Leung, Thomas Wing-Yan Chin, Christine Shing Yen Lo, Macy Mei-Sze Lui, Jonan Chun Yin Lee, Keith Wan-Hang Chiu, Tom Wai-Hin Chung, et al. Frequency and distribution of chest radiographic findings in patients positive for covid-19. *Radiology*, 296(2):E72–E78, 2020.

[80] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, 296(2):E32–E40, 2020.

[81] Elissa Driggin, Mahesh V Madhavan, Behnood Bikdeli, Taylor Chuich, Justin Laracy, Giuseppe Biondi-Zoccai, Tyler S Brown, Caroline Der Nigoghossian, David A Zidar, Jennifer Haythe, et al. Cardiovascular considerations for patients, health care workers, and health systems during the covid-19 pandemic. *Journal of the American College of Cardiology*, 75(18):2352–2371, 2020.

[82] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[83] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[84] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.

[85] Vladimir Vapnik, Isabel Guyon, and Trevor Hastie. Support vector machines. *Mach. Learn*, 20(3):273–297, 1995.

[86] Shun-ichi Amari and Si Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[87] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.

[88] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[89] Francois Chollet et al. *Deep learning with Python*, volume 361. Manning New York, 2018.

[90] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[91] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[92] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[93] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[95] Patrice Y Simard, Yann A LeCun, John S Denker, and Bernard Victorri. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer, 1998.

[96] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[97] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[98] Tyler Clark, Junjie Zhang, Sameer Baig, Alexander Wong, Masoom A Haider, and Farzad Khalvati. Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted mri using convolutional neural networks. *Journal of Medical Imaging*, 4(4):041307, 2017.

[99] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[100] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2): 31–71, 1997.

[101] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[102] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[103] Garry Choy, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E Samir, Oleg S Pianykh, J Raymond Geis, Pari V Pandharipande, James A Brink, and Keith J Dreyer. Current applications and future impact of machine learning in radiology. *Radiology*, 288(2):318–328, 2018.

[104] Natalia Antropova, Benjamin Q Huynh, and Maryellen L Giger. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical physics*, 44(10):5162–5171, 2017.

[105] Natalia O Antropova, Hiroyuki Abe, and Maryellen L Giger. Use of clinical mri maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. *Journal of Medical Imaging*, 5(1):014503, 2018.

[106] Yu Ji, Hui Li, Alexandra V Edwards, John Papaioannou, Wenjuan Ma, Peifang Liu, and Maryellen L Giger. Independent validation of machine learning in diagnosing breast cancer on magnetic resonance imaging within a single institution. *Cancer Imaging*, 19 (1):1–11, 2019.

[107] Neha Bhooshan, Maryellen Giger, Li Lan, Hui Li, Angelica Marquez, Akiko Shimauchi, and Gillian M Newstead. Combined use of t2-weighted mri and t1-weighted dynamic contrast–enhanced mri in the automated analysis of breast lesions. *Magnetic Resonance in Medicine*, 66(2):555–564, 2011.

[108] Mehmet U Dalmis, Albert Gubern-Mérida, Suzan Vreemann, Peter Bult, Nico Karssemeijer, Ritse Mann, and Jonas Teuwen. Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast mri protocol with ultrafast dce-mri, t2, and dwi. *Investigative radiology*, 54(6):325–332, 2019.

[109] Weijie Chen, Maryellen L Giger, and Ulrich Bick. A fuzzy c-means (fcm)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced mr images1. *Academic radiology*, 13(1):63–72, 2006.

[110] Karen Drukker, Alexandra V Edwards, Christopher Doyle, John Papaioannou, Kirti Kulkarni, and Maryellen L Giger. Breast mri radiomics for the pretreatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients. *Journal of Medical Imaging*, 6(3):034502, 2019.

[111] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[112] John Shawe-Taylor and Shiliang Sun. A review of optimization methodologies in support vector machines. *Neurocomputing*, 74(17):3609–3618, 2011.

[113] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[114] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[115] André Collignon, Frederik Maes, Dominique Delaere, Dirk Vandermeulen, Paul Suetens, and Guy Marchal. Automated multi-modality image registration based on information theory. In *Information processing in medical imaging*, volume 3, pages 263–274, 1995.

[116] John Ashburner and Karl J Friston. Rigid body registration. *Statistical parametric mapping: The analysis of functional brain images*, pages 49–62, 2007.

[117] Charles E Metz, Benjamin A Herman, and Jong-Her Shen. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in medicine*, 17(9):1033–1053, 1998.

[118] Charles E Metz and Xiaochuan Pan. "proper" binormal roc curves: theory and maximum-likelihood estimation. *Journal of mathematical psychology*, 43(1):1–33, 1999.

[119] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

[120] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[121] Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

[122] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[123] Soyeon Ahn, Seong Ho Park, and Kyoung Ho Lee. How to demonstrate similarity by using noninferiority and equivalence statistical testing in radiology research. *Radiology*, 267(2):328–338, 2013.

[124] Nancy A Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8, 2003.

[125] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[126] Karla Horsch, Maryellen L Giger, and Charles E Metz. Prevalence scaling: Applications to an intelligent workstation for the diagnosis of breast cancer. *Academic radiology*, 15 (11):1446–1457, 2008.

[127] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[128] NE Hawass. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *The British journal of radiology*, 70(832): 360–366, 1997.

[129] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[130] Lin Chen, Craig K Abbey, Anita Nosratieh, Karen K Lindfors, and John M Boone. Anatomical complexity in breast parenchyma and its implications for optimal breast imaging strategies. *Medical physics*, 39(3):1435–1441, 2012.

[131] Lin Chen, Craig K Abbey, and John M Boone. Association between power law coefficients of the anatomical noise power spectrum and lesion detectability in breast imaging modalities. *Physics in Medicine & Biology*, 58(6):1663, 2013.

[132] John W Garrett, Yinsheng Li, Ke Li, and Guang-Hong Chen. Reduced anatomical clutter in digital breast tomosynthesis with statistical iterative reconstruction. *Medical physics*, 45(5):2009–2022, 2018.

[133] Jing Li, Ming Fan, Juan Zhang, and Lihua Li. Discriminating between benign and malignant breast tumors using 3d convolutional neural network in dynamic contrast enhanced-mr images. In *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, volume 10138, page 1013808. International Society for Optics and Photonics, 2017.

[134] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[135] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12, 2020.

[136] Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 490, 2020.

[137] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna AA Damen, Thomas PA Debray, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020.

[138] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[139] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[140] Rsna pneumonia detection challenge. `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data`, 2018. Accessed: 2020-06-18.

[141] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[142] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[143] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[144] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

[145] Saman Motamed, Patrik Rogalla, and Farzad Khalvati. Randgan: randomized generative adversarial network for detection of covid-19 in chest x-ray. *Scientific Reports*, 11(1):1–10, 2021.

[146] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[147] Ian Pan, Alexandre Cadrin-Chênevert, and Phillip M Cheng. Tackling the radiological society of north america pneumonia detection challenge. *American Journal of Roentgenology*, 213(3):568–574, 2019.

[148] Feng Li, Roger Engelmann, Lorenzo L. Pesce, Kunio Doi, Charles E Metz, and Heber MacMahon. Small lung cancers: improved detection by use of bone suppression imaging—comparison with dual-energy subtraction chest radiography. *Radiology*, 261(3): 937–949, 2011.

[149] American College of Radiology et al. Acr–spr–str practice parameter for the performance of chest radiography, 2017.

[150] Christiane K Kuhl. Abbreviated magnetic resonance imaging (mri) for breast cancer screening: rationale, concept, and transfer to clinical practice. *Annual review of medicine*, 70:501–519, 2019.

[151] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

APPENDIX A

HUMAN-ENGINEERED RADIOMIC FEATURES FOR MULTIPARAMETRIC MRI

Table A.1: Radiomic features extracted from dynamic contrast-enhanced (DCE) sequence and their descriptions.

Category	Full name (unit)	Feature description
Geometry [25]	Volume (mm^3)	Volume of lesion
	Effective diameter (mm)	Greatest dimension of a sphere with the same volume as the lesion
	Surface area (mm^2)	Lesion surface area
	Maximum diameter (mm)	Maximum distance between any two voxels in the lesion
	Sphericity	Similarity of the lesion shape to a sphere
	Irregularity	Deviation of the lesion surface from the surface of a sphere
	Surface area/volume (1/mm)	Ratio of surface area to volume
Morphology [25]	Margin sharpness	Mean of the image gradient at the lesion margin
	Variance of margin sharpness	Variance of the image gradient at the lesion margin

Table A.1: DCE radiomic features (continued)

	Variance of radial gradient histogram	Degree to which the enhancement structure extends in a radial pattern originating from the center of the lesion
Texture [28]	Contrast	Location image variations
	Correlation	Image linearity
	Difference entropy	Randomness of the difference of neighboring voxels' gray-levels
	Difference variance	Variations of difference of gray-level between voxel-pairs
	Angular second moment (energy)	Image homogeneity
	Entropy	Randomness of the gray-levels
	Inverse difference moment (homogeneity)	Image homogeneity
	Information measure of correlation 1	Nonlinear gray-level dependence
	Information measure of correlation 2	Nonlinear gray-level dependence
	Maximum correlation coefficient	Nonlinear gray-level dependence
	Sum average	Overall brightness
	Sum entropy	Randomness of the sum of gray-level dependence
	Sum variance	Spread in the sum of the gray-levels of neighboring voxels

Table A.1: DCE radiomic features (continued)

	Sum of squares (variance)	Spread in the gray-level distribution
Kinetics [27]	Maximum enhancement	Maximum contrast enhancement
	Time to peak (s)	Time at which the maximum enhancement occurs
	Uptake rate (1/s)	Uptake speed of the contrast enhancement
	Washout rate (1/s)	Washout speed of the contrast enhancement
	Curve shape index	Difference between late and early enhancement
	Enhancement at first post-contrast time point	Enhancement at first post-contrast time point
	Signal enhancement ratio	Ratio of initial enhancement to overall enhancement
	Volume of most enhancing voxels (mm^3)	Volume of the most enhancing voxels
	Total rate variation ($1/s^2$)	How rapidly the contrast will enter and exit from the lesion
	Normalized total rate variation ($1/s^2$)	How rapidly the contrast will enter and exit from the lesion
Enhancement-variance kinetics [26]	Maximum enhancement-variance	Maximum spatial variance of contrast enhancement over time

Table A.1: DCE radiomic features (continued)

	Enhancement-variance time to peak (s)	Time at which the maximum variance occurs
	Enhancement-variance increasing rate (1/s)	Rate of increase of the enhancement-variance during uptake
	Enhancement-variance decreasing rate (1/s)	Rate of decrease of the enhancement-variance during washout
Gray-level statistics [110]	Mean voxel value pre-contrast	Average gray-level intensity within the lesion prior to contrast injection
	Mean voxel value post-contrast injection	Average gray-level intensity within the lesion at first post-contrast injection time point
	Standard deviation of voxel value distribution pre-contrast	Variation in gray-level intensity within the lesion prior to contrast injection
	Standard deviation of voxel value distribution post-contrast	Variation in gray-level intensity within the lesion at first post-contrast injection time point
	Maximum voxel value pre-contrast	Maximum gray-level intensity within the lesion prior to contrast injection

Table A.1: DCE radiomic features (continued)

Maximum voxel value post-contrast	Maximum gray-level intensity within the lesion at first post-contrast injection time point
Minimum voxel value pre-contrast	Minimum gray-level intensity within the lesion prior to contrast injection
Minimum voxel value post-contrast	Minimum gray-level intensity within the lesion at first post-contrast injection time point
Kurtosis of voxel value distribution pre-contrast	Tailedness of gray-level intensity distribution within the lesion prior to contrast injection
Kurtosis of voxel value distribution post-contrast	Tailedness of gray-level intensity distribution within the lesion at first post-contrast injection time point
Skewness of voxel value distribution pre-contrast	Asymmetry of gray-level intensity distribution about the mean within the lesion prior to contrast injection
Skewness of voxel value distribution post-contrast	Asymmetry of gray-level intensity distribution about the mean within the lesion at first post-contrast injection time point

Table A.2: Radiomic features extracted from T2-weighted sequence and their descriptions.

Category	Full name (unit)	Feature description
Morphology [25, 107]	Margin sharpness	Mean of the image gradient at the lesion margin
	Variance of margin sharpness	Variance of the image gradient at the lesion margin
	Variance of radial gradient histogram	Degree to which the enhancement structure extends in a radial pattern originating from the center of the lesion
Texture [28, 107]	Contrast	Location image variations
	Correlation	Image linearity
	Difference entropy	Randomness of the difference of neighboring voxels' gray-levels
	Difference variance	Variations of difference of gray-level between voxel-pairs
	Angular second moment (energy)	Image homogeneity
	Entropy	Randomness of the gray-levels
	Inverse difference moment (homogeneity)	Image homogeneity
	Information measure of correlation 1	Nonlinear gray-level dependence
	Information measure of correlation 2	Nonlinear gray-level dependence

Table A.2: T2-weighted radiomic features (continued)

	Maximum correlation coefficient	Nonlinear gray-level dependence
	Sum average	Overall brightness
	Sum entropy	Randomness of the sum of gray-level dependence
	Sum variance	Spread in the sum of the gray-levels of neighboring voxels
	Sum of squares (variance)	Spread in the gray-level distribution
Gray-level statistics [107]	Mean voxel value	Average gray-level intensity within the lesion
	Variance of voxel value	Variation in gray-level intensity within the lesion

Table A.3: Radiomic features extracted from the apparent diffusion coefficient (ADC) map derived from diffusion-weighted imaging (DWI) sequence and their descriptions.

Category	Full name (unit)	Feature description
ADC map statistics [16]	Mean ADC	Average ADC within the lesion
	Standard deviation of ADC distribution	Variation in ADC within the lesion
	Maximum ADC	Maximum ADC within the lesion
	Minimum ADC	Minimum ADC within the lesion
	Range of ADC distribution	Range of ADC distribution within the lesion
	Skewness of ADC distribution	Asymmetry of ADC distribution about the mean within the lesion

APPENDIX B

COMPARATIVE RADIOMICS EVALUATION OF PAIRED CONVENTIONAL DCE-MRI AND ABBREVIATED MRI FOR BREAST CANCER DIAGNOSIS

B.1 Introduction

Breast MRI offers the highest cancer detection rate of all breast imaging modalities, and screening with MRI has shown evidence to benefit women at high risk and average risk, as introduced in Chapter 1.2.1. However, the use of conventional, full-protocol breast MRI to screen the large number of average-risk women with dense breasts will be neither practical nor cost-effective. Abbreviated breast MRI has been introduced to reduce the complexity and cost of MRI by reducing image acquisition and interpretation time and hence improve access to breast MRI [46]. Multiple studies have confirmed equivalent diagnostic accuracy of abbreviated breast MRI with full MRI protocols [150]. The work presented in this chapter compared the diagnostic performance of radiomics analysis on dynamic contrast-enhanced (DCE)-MRI and abbreviated MRI in the task of distinguishing between benign and malignant breast lesions.

B.2 Methods

A dataset consisting of 1188 unique breast lesions (271 benign and 917 malignant) from 877 women (age range 23-89) who had undergone conventional breast DCE-MR exams was retrospectively collected under HIPAA-compliant Institutional Review Board-approved protocols. This dataset was from the same database as described in Section 2.2 but is larger since exams were not required to contain T2-weighted or diffusion-weighted imaging sequences.

The methods for this comparative analysis are illustrated in Fig. B.1. Lesions were

automatically segmented on the conventional DCE-MRI using a fuzzy C-means method, and 50 radiomic features were extracted as described in Section 2.3.1 and A. To mimic an abbreviated MRI sequence and provide paired comparisons, a "mock abbreviated MRI database" was simulated by only considering the pre-contrast and first post-contrast time points of the DCE-MRI. The lesion segmentation method initially developed for DCE-MRI was applied on the mock abbreviated MRI, and 42 radiomics features were extracted from the mock abbreviated MRI. Eight of the DCE-MRI kinetics-related features involve wash-out and thus are not relevant for abbreviated MRI. A hybrid analysis of the DCE and abbreviated MRI was performed by segmenting lesions from DCE-MRI and extracting the 42 abbreviated radiomic features. The hybrid analysis was only for the purpose of comparative analysis and did not correspond to a realistic scenario. The dataset was randomly split by patient into 80%/20% training/test sets that had the same class prevalence. Three support vector machine (SVM) classifiers were trained on the 50 features extracted from DCE-MRI, the 42 features extracted from mock abbreviated MRI, and the hybrid features, respectively.

Dice similarity coefficient was used to evaluate the agreement between segmentations from DCE-MRI and abbreviated MRI [151]. Diagnostic performance in the task of classifying lesions as malignant or benign was evaluated using receiver operating characteristic (ROC) analysis, and the area under the ROC curve (AUC) served as the figure of merit [118]. Bonferroni-Holm corrections were used to account for multiple comparisons [122]. $P < .05$ was considered to indicate a statistically significant difference.

B.3 Results and Discussion

The Dice coefficient comparing the segmentation between DCE-MRI and abbreviated MRI had a median and 95% CI of 0.86 [0.52, 0.96], with the distribution shown in Fig. B.2. Figure B.3 include examples comparing segmentation results based on full DCE-MRI and abbreviated MRI.

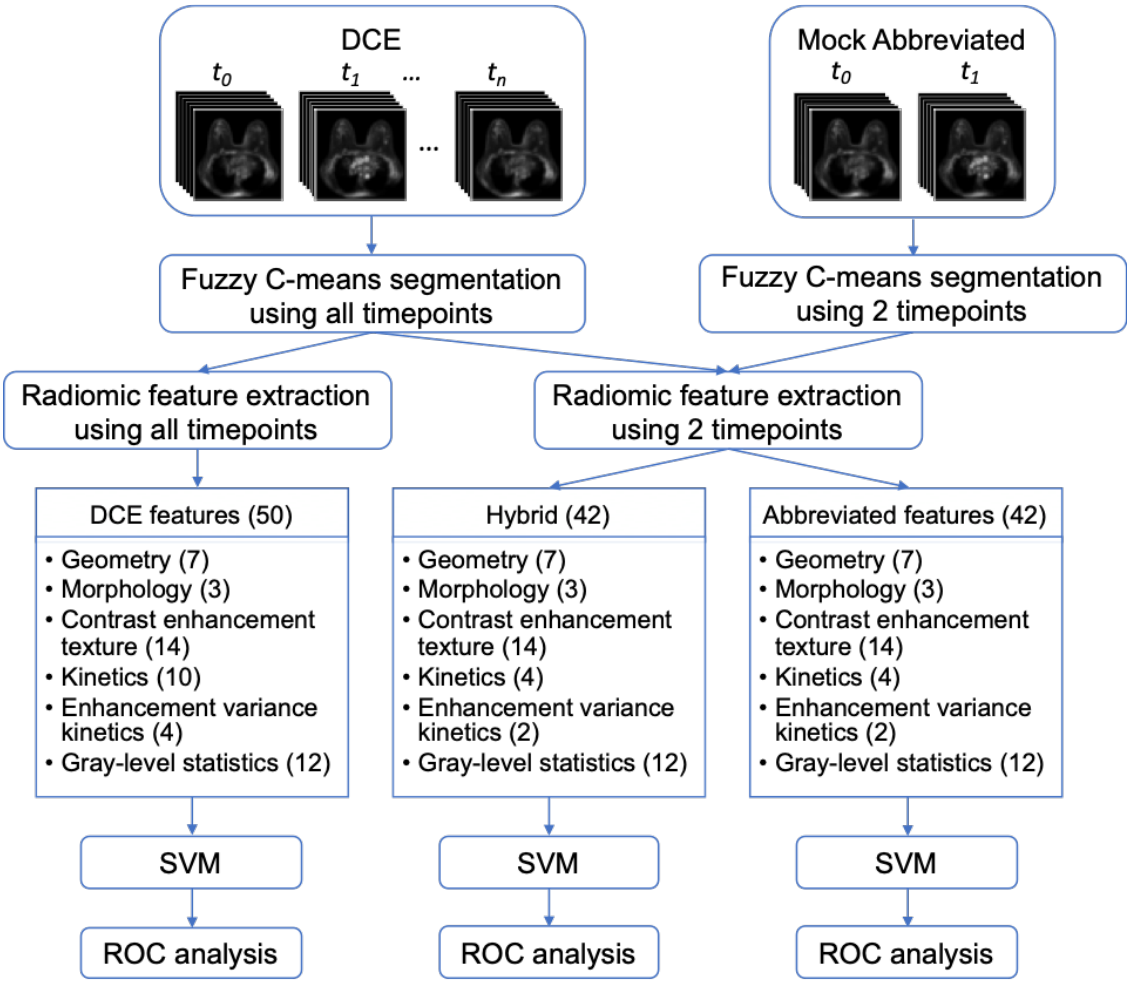


Figure B.1: Flowchart for the comparative radiomic analysis of paired DCE-MRI and abbreviated MRI.

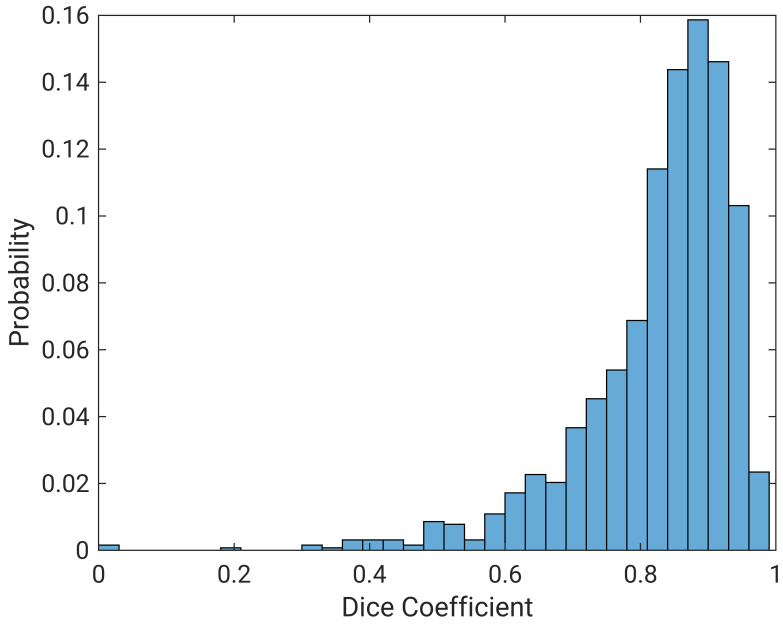
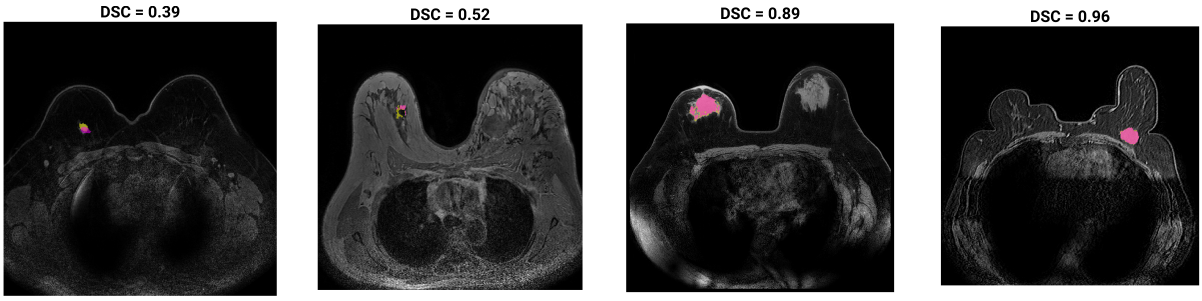


Figure B.2: Distribution of Dice coefficient comparing segmentation on DCE-MRI and abbreviated MRI.

Classification of benign and malignant lesions using radiomic features extracted from DCE-MRI, abbreviated MRI, and hybrid analysis yielded AUC values [95% CI] of 0.87 [0.85, 0.90], 0.84 [0.81, 0.86], and 0.86 [0.85, 0.89], respectively, as shown in Fig. B.4. The statistical analysis comparing the abbreviated MRI classifier and the hybrid classifier is presented in Table B.1. The abbreviated MRI classifier yielded a statistically significantly lower AUC than the DCE-MRI classifier, but the hybrid classifier failed to demonstrate a significant difference from the DCE classifier and yielded equivalent performance within a margin of $\Delta\text{AUC} = 0.05$.

Moderate disagreement between predictions based on radiomic features extracted from DCE and abbreviated MRI is observed in the Bland-Altman plot in Fig. B.5. More benign cases are in the lower half of the figure, and more benign cases are in the upper half of the figure, showing that the classifier using features from DCE-MRI more accurately assigned malignant cases with higher probabilities of malignancy (PMs) and benign cases with lower PMs than the classifier using featured from abbreviated MRI.



(a) (b) (c) (d)

Figure B.3: Examples of segmentation results based on DCE-MRI (yellow) and abbreviated MRI (magenta).

Table B.1: AUC values for classifiers using radiomic features from DCE-MRI, abbreviated MRI, and hybrid analysis, as well as the p-value and 95% CI of the difference in AUC values. Asterisks denote statistical significance after multiple comparison corrections.

	AUC [95% CI]	Compared with DCE
DCE	0.87 [0.85, 0.90]	—
Abbreviated	0.84 [0.81, 0.86]	$P < .001^*$ 95% CI ΔAUC = [0.021, 0.049]
Hybrid	0.86 [0.85, 0.89]	$P = .11$ 95% CI ΔAUC = [-0.002, 0.015]

Figure B.6 shows the difference in the two classifiers' PMs in various Dice coefficient ranges. In the low Dice coefficient range, the predictions based on radiomic features extracted from DCE-MRI tended to be more accurate than those based on abbreviated MRI, i.e., lower PMs for benign lesions and higher PMs for malignant lesions, and there tended to be larger variations in the difference in PMs. This observation and the fact that the hybrid classifier yielded equivalent classification performance as the DCE classifier suggest that the difference in lesion segmentation quality likely contributed to the difference in downstream classification performance between classifiers based on DCE-MRI and abbreviated MRI.

Future work will develop new segmentation methods for abbreviated MRI, aiming to improve the segmentation quality. Then abbreviated MRI radiomics classification performance will be evaluated using improved lesion segmentation. Validation on independent datasets

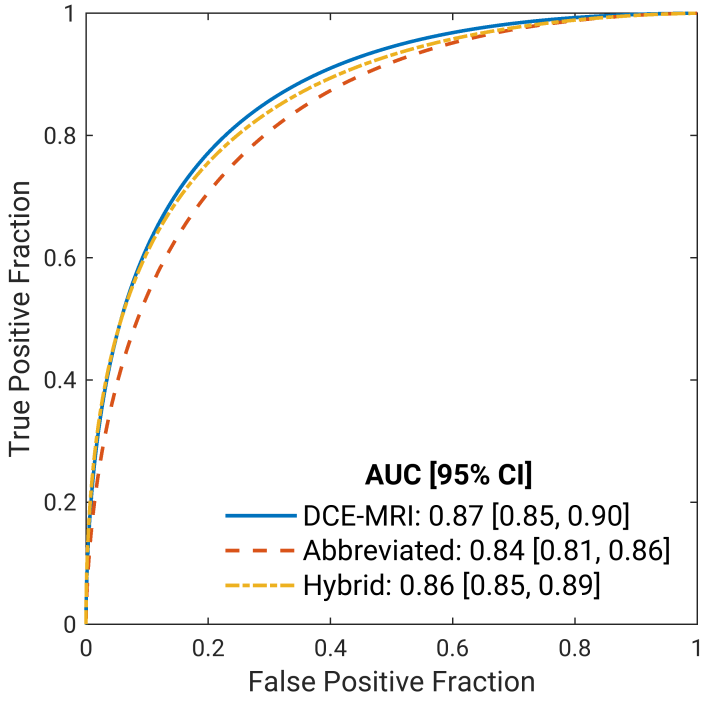


Figure B.4: Fitted binormal receiver operating characteristic (ROC) curves for the classification task breast lesions using DCE-MRI, abbreviated MRI, and the hybrid analysis. The legend gives the area under the ROC curve (AUC) with the 95% confidence interval (CI) for each classifier.

acquired using actual abbreviated MRI protocol can be performed as well.

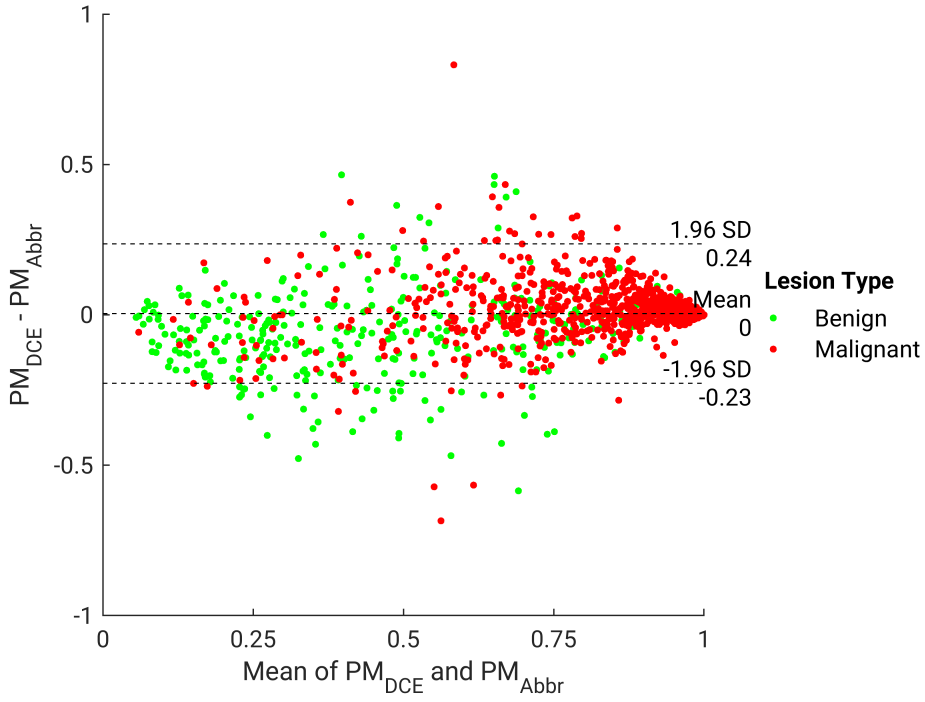


Figure B.5: Bland-Altman plot illustrating classifier agreement between the classifiers trained on dynamic contrast-enhanced (DCE) features and abbreviated MRI features. PM = Probability of malignancy.

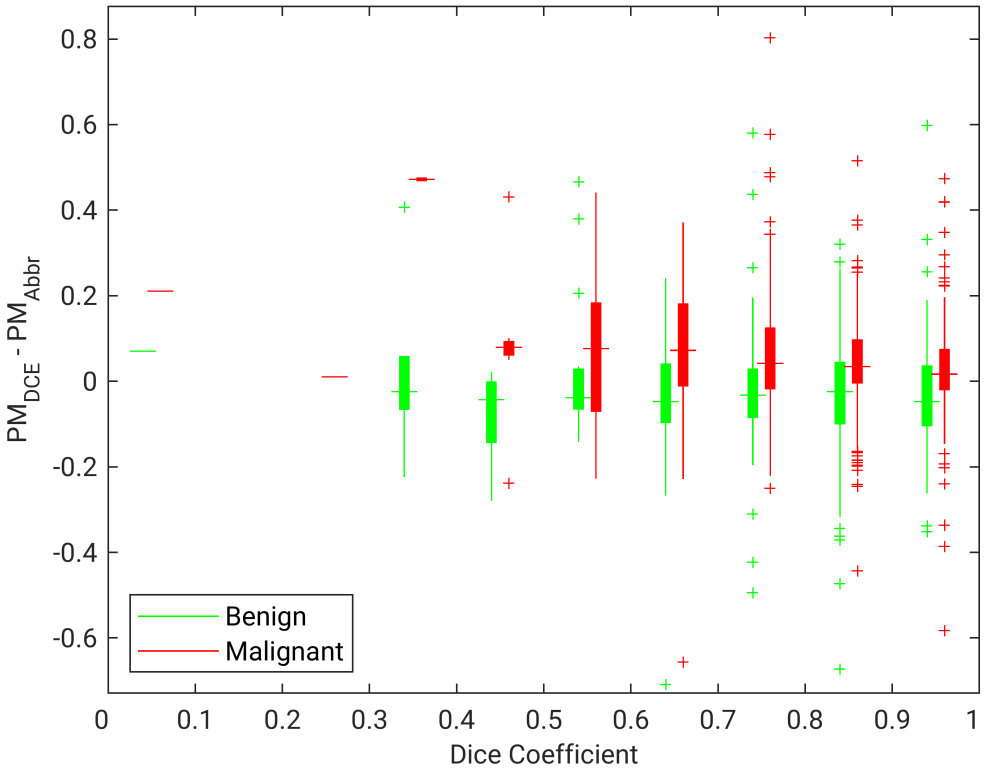


Figure B.6: Difference in the two classifiers' PMs in various Dice coefficient ranges.

LIST OF PUBLICATIONS AND PRESENTATIONS

Peer-Reviewed Publications

Hu Q, Drukker K, Giger ML. The role of standard and soft tissue chest radiography images in deep-learning-based early diagnosis of COVID-19. (under review at Journal of Medical Imaging).

Hu Q, Giger ML. Chapter on clinical AI applications: breast imaging, Radiologic Clinics of North America issue on Artificial Intelligence and Radiology. (in press).

El Naqa I, Li H, Fuhrman JD, Hu Q, Gorre N, Chen W, Giger ML. Lessons learned in transitioning to AI in the medical imaging of COVID-19. (under review at Journal of Medical Imaging)

Fuhrman JD, Gorre N, Hu Q, Li H, El Naqa I, Giger ML. A review of explainable and interpretable AI with application to medical imaging of COVID-19. (under review at Medical Physics)

Hu Q, Whitney HM, Li H, Ji Y, Liu P, Giger ML. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection MRI in breast cancer diagnosis using dynamic contrast-enhance MRI. Radiology: Artificial Intelligence. 2021 Feb 24;3(3):e200159.

Hu Q, Whitney HM, Giger ML. Radiomics methodology for breast cancer diagnosis using multiparametric magnetic resonance imaging. Journal of Medical Imaging. 2020 Aug;7(4):044502.

Hu Q, Whitney HM, Giger ML. A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. Scientific reports. 2020 Jun 29;10(1):1-1.

Gong H, Hu Q, Walther A, Koo CW, Takahashi EA, Levin DL, Johnson TF, Hora MJ, Leng S, Fletcher JG, McCollough CH. Deep-learning-based model observer for a lung nodule detection task in computed tomography. Journal of Medical Imaging. 2020 Jun;7(4):042807.

Proceedings Papers and Extended Abstracts

Hu Q, Drukker K, Giger ML. Role of standard and soft tissue chest radiography images in COVID-19 diagnosis using deep learning. InMedical Imaging 2021: Computer-Aided Diagnosis 2021 Feb 15 (Vol. 11597, p. 1159704). International Society for Optics and Photonics. (Runner-up, Computer-Aided Diagnosis Paper Award. Runner-up, Robert F. Wagner All-Conference Best Student Paper award.)

Bhattacharjee R, Douglas L, Drukker K, Hu Q, Fuhrman JD, Sheth D, Giger M. Comparison of 2D and 3D U-Net breast lesion segmentations on DCE-MRI. InMedical Imaging 2021: Computer-Aided Diagnosis 2021 Feb 15 (Vol. 11597, p. 115970D). International Society for Optics and Photonics.

Hu Q, Drukker K, Giger ML. Predicting the feed for intensive care for COVID-19 patients using deep learning on chest radiography. The 34th Neural Information Processing Systems Conference, Medical Imaging meets NeurIPS Workshop 2020.

Hu Q, Whitney HM, Giger ML. Using ResNet feature extraction in computer-aided diagnosis of breast cancer on 927 lesions imaged with multiparametric MRI. InMedical Imaging 2020: Computer-Aided Diagnosis 2020 Mar 16 (Vol. 11314, p. 1131411). International Society for Optics and Photonics.

Hu Q, Whitney HM, Giger ML. Transfer learning in 4D for breast cancer diagnosis using dynamic contrast-enhanced magnetic resonance imaging. arXiv preprint arXiv:1911.03022. 2019 Nov 8.

Hu Q, Whitney HM, Edwards A, Papaioannou J, Giger ML. Radiomics and deep learning of diffusion-weighted MRI in the diagnosis of breast cancer. InMedical Imaging 2019: Computer-Aided Diagnosis 2019 Mar 13 (Vol. 10950, p. 109504A). International Society for Optics and Photonics.

Gong H, Walther A, Hu Q, Koo CW, Takahashi EA, Levin DL, Johnson TF, Hora MJ, Leng S, Fletcher JG, McCollough CH. Correlation between a deep-learning-based model observer

and human observer for a realistic lung nodule localization task in chest CT. InMedical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment 2019 Mar 4 (Vol. 10952, p. 109520K). International Society for Optics and Photonics.

Yu L, Hu Q, Koo CW, Takahashi EA, Levin DL, Johnson TF, Hora MJ, Dirks S, Chen B, McMillan K, Leng S. A virtual clinical trial using projection-based nodule insertion to determine radiologist reader performance in lung cancer screening CT. InMedical Imaging 2017: Physics of Medical Imaging 2017 Mar 9 (Vol. 10132, p. 101321R). International Society for Optics and Photonics.

Oral Presentations

Hu Q, Drukker K, Giger ML. Role of standard and soft tissue chest radiography images in COVID-19 diagnosis using deep learning. SPIE Medical Imaging 2021.

Bhattacharjee R, Douglas L, Drukker K, Hu Q, Fuhrman JD, Sheth D, Giger M. Comparison of 2D and 3D U-Net breast lesion segmentations on DCE-MRI. SPIE Medical Imaging 2021, by Bhattacharjee R.

Hu Q, Whitney HM, Giger ML. Using ResNet feature extraction in computer-aided diagnosis of breast cancer on 927 lesions imaged with multiparametric MRI. SPIE Medical Imaging 2020.

Hu Q, Whitney HM, Edwards A, Papaioannou J, Giger ML. Multiparametric breast MRI radiomics in distinguishing between benign and malignant breast lesions. American Association of Physicists in Medicine (AAPM) 2019. (Selected oral presentation in the Science Council session. Highlighted as best of AAPM at American Society for Radiation Oncology 2019 annual meeting.)

Gong H, Walther A, Hu Q, Koo CW, Takahashi EA, Levin DL, Johnson TF, Hora MJ, Leng S, Fletcher JG, McCollough CH. Correlation between a deep-learning-based model observer and human observer for a realistic lung nodule localization task in chest CT. SPIE Medical

Imaging 2019, by Gong H.

Yu L, Hu Q, Koo CW, Takahashi EA, Levin DL, Johnson TF, Hora MJ, Dirks S, Chen B, McMillan K, Leng S. A virtual clinical trial using projection-based nodule insertion to determine radiologist reader performance in lung cancer screening CT. SPIE Medical Imaging 2017, by Yu L.

Poster Presentations

Schlafly G, Hu Q, Li F, Drukker K, Fuhrman JD, Giger ML. Automatic lung field segmentation on chest radiographs of COVID-19 patients to improve diagnostic deep learning. AAPM 2021, by Schlafly G.

Hu Q, Drukker K, Giger ML. Predicting the need for intensive care for COVID-19 patients using deep learning on chest radiography. The 34th Neural Information Processing Systems Conference (NeurIPS), Medical Imaging meets NeurIPS Workshop 2020.

Hu Q, Papaioannou J, Whitney HM, Edwards A, Giger ML. Comparative radiomics evaluation of paired conventional DCE-MRI and abbreviated MRI for breast cancer diagnosis. Scientific poster presentation, Radiological Society of North America 2020.

Hu Q, Whitney HM, Giger ML. Transfer learning in 4D for breast cancer diagnosis using dynamic contrast-enhanced magnetic resonance imaging. The 33rd Neural Information Processing Systems Conference, Machine Learning for Health Workshop 2019.

Hu Q, Whitney HM, Edwards A, Papaioannou J, Giger ML. Radiomics and deep learning of diffusion-weighted MRI in the diagnosis of breast cancer. SPIE Medical Imaging 2019.