THE UNIVERSITY OF CHICAGO

ARTIFICIAL INTELLIGENCE FOR BREAST CANCER RISK ASSESSMENT IN MAMMOGRAPHY AND METHODS FOR DATASET BALANCING AND DISTRIBUTION SAMPLING

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

NATALIE MARITA BAUGHAN

CHICAGO, ILLINOIS

JUNE 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I have been incredibly fortunate to have been surrounded by many wonderful friends and mentors who made this project possible. First, I want to thank Dr. Maryellen Giger, my thesis advisor, for her countless hours helping me conceptualize and complete this project. I am very thankful to have had the opportunity to work with such a well-respected and knowledgeable scientist. From her expertise in mentorship, she pushed me to achieve when she knew I could and always kept a clear vision of my projects, even when I struggled.

My thesis committee members have also been excellent mentors and guides in my scientific endeavors. Dr. Hania Al-Hallaq contributed her knowledge of breast imaging and clinical physics to ensure my project was practical and relevant to clinical problems. Dr. Isabelle Bedrosian generously shared her expertise as a breast surgical oncologist, which helped to motivate and guide the direction of this work as an extension of her previous work in genomics-based risk assessment. I also thank Dr. Bedrosian and her lab for all their assistance in image processing and transfer from the M.D. Anderson lab. Dr. Samuel Armato has been a strong medical physics graduate program leader and a source of knowledge in image biomarkers and the broader lens of MIDRC.

Beyond my committee, I have also worked with many collaborators who have been instrumental to this work. I have to first thank Dr. Hui Li and Li Lan for being excellent virtual office mates and teaching me about radiomics, deep learning, and image processing. I will miss catching up with you both on Thursday mornings. Dr. Heather Whitney and Dr. Kyle Myers, of my MIDRC TDP 3d team, I have to thank for their expertise in building the algorithms for sequestration and task-based sampling. Our larger MIDRC group, including Dr. Karen Drukker, Dr. Berkman Sahnier, Dr. Michael McNitt-Gray, Dr. Tingting Hu, Dr. Grace Kim, and Emily

Townley, I also thank for their valuable input and for giving me the opportunity to work with such a renowned group of physicists and scientists.

The Graduate Program in Medical Physics (GPMP) faculty have also been a great support to my goals and growth. I would like to particularly thank all faculty who assisted in teaching my courses, allowed me to rotate in their lab, or helped me to shadow in the clinic.

I would also like to thank the former summer students I have mentored, Dylan Tang, Jonathan Lopez, and Maya Ballard, for their assistance in my projects, allowing me to grow as a leader myself.

The GPMP student cohort was instrumental to my success in the program. A special thanks to my cohort of six: Hadley DeBrosse, Linnea Kremer, Mira Liu, Mena Shenouda, and Julian Bertini, for many great discussions about life and science. My lab mate, Lindsay Douglas, for always being there to relate and for her help in preparation for residency interviews. Many other GPMP students were also a joy to work alongside and learn from, and I am thankful for them.

This dissertation was completed with the financial support of the following: the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) grant numbers T32 EB002103, U01 CA195564, 75N92020C00008, and 75N92020C00021, the University of Chicago Comprehensive Cancer Center, and the Lawrence H. Lanzl Graduate Fellowship Award.

Lastly, I thank my family and friends for their love and support. To my college roommates, and now lifelong friends, I deeply appreciated our FaceTime calls to catch up on life and for giving me seven great reasons to take a weekend trip or share my tiny studio apartment for wonderful adventures. I thank my sister, Alex, for motivating me to achieve since I was a child through lots of friendly sibling rivalry and for being there to relate on both of our journeys to becoming Dr.

Baughan. I thank my parents for giving me a place to escape the stresses of school, and a pandemic, and for always reminding me of the important things in life.

ABSTRACT

Artificial intelligence (AI) has become a driving force in medical imaging, from applications in breast cancer screening to COVID-19. Within the field of breast cancer screening, AI systems using human-engineered radiomic features and deep learning extracted features have shown promising performance in breast imaging diagnosis, detection, and risk assessment. However, AI has not yet been applied to the investigation of a breast cancer field effect, in which histologically normal areas of the parenchyma show molecular similarity to the tumor. Identification of a cancer field effect in mammography has the potential to provide a novel approach to stratification of breast cancer risk in the general population. Furthermore, development of a temporal risk assessment model would expand upon the potential impact of utilizing AI-based tools to predict risk of future cancer from the breast parenchyma.

As a result of the explosion of machine intelligence algorithm development for understanding and characterizing a wide variety of diseases, including breast cancer and COVID-19, validation of algorithm performance and generalizability have become increasingly important. To ensure that AI systems are robust and generalizable, the data with which they are evaluated should be population-representative and independent of that used for training. The development of novel algorithmic methods for the creation of a large, common sequestered dataset and task-based sampling would enable robust evaluations of AI algorithms on representative datasets. A sequestered database for algorithm testing could also allow for expedited clinical implementation of algorithms developed for medical decision-making if accepted by regulating bodies.

Aim 1: Mammograms and mastectomy specimen radiographs of women with a malignant tumor were investigated using radiomic and deep learning based features to provide initial characterization of a breast cancer field effect in imaging. Features were extracted from four regions: within the tumor, near to the tumor, far from the tumor, and in the contralateral breast. Results found statistically significant correlations of feature values with the region's proximity to the tumor in intensity-based features and select structure-based features.

Aim 2: To improve upon conventional breast cancer risk assessment models, a method that analyzes prior mammography data to predict future occurrence of breast cancer was implemented. The long-short-term memory network (LSTM), a network that can incorporate AI-based features into a temporal model, was utilized and compared to classification using only a single time point. The resulting LSTM network was able to predict incidence of cancer in the subsequent year with performance significantly better than guessing.

Aim 3: Data used in the development and evaluation of AI models play a significant role in the robustness and generalizability of the model performance. To enable independent assessment of algorithms using a multi-institutional data commons, a first-of-its-kind sequestered commons was initiated using a developed method of multi-dimensional stratified sampling. To draw an independent sample for performance evaluation from the commons, a novel method of task-based distribution sampling was also developed. This aim was completed in collaboration with the Medical Imaging and Data Resource Center (MIDRC), a multi-institutional effort to accelerate machine intelligence research for COVID-19.

CHAPTER 1: INTRODUCTION

1.1 Artificial intelligence in medical imaging

Despite the recent explosion of artificial intelligence (AI) in medical imaging, facilitated by advances in deep learning networks and computing power, medical imaging AI research has been ongoing for decades. The first publications on the use of computers for cancer detection from radiographic images were published in the 1950s and 60s [1]. However, computational limitations and inadequate image quality prevented practical use of these methods. In the late 1980s and 1990s, AI tools for the detection of lung and breast cancer were revisited and developed, with the names CADe and CADx (for computer-aided detection and computer-aided diagnosis, respectively) to represent their role as an aid to the radiologist as opposed to a replacement [2,3]. The first observer study with mammography to compare radiologist performance with and without CADe was published by Chan et al. in 1990 [4,5]. The first use of deep learning in medical imaging using a convolutional neural network (CNN) was published by Zhang et al. in 1994 for detecting microcalcification clusters in mammograms and then incorporated into CADe commercial systems [6]. This first commercial CADe system was the ImageChecker M1000 (R2 Technology, now Hologic Inc., Bedford, MA), approved by the Food and Drug Administration (FDA) in 1998 to serve as a second reader to be used after a radiologist's initial review [4]. By 2008, CADe was used in 70% of mammographic screening studies at outpatient hospitals and in 81% of screenings at private practices [7].

The late 1990s and early 2000s brought about increased research in AI for breast cancer, especially for diagnostic tasks (i.e., CADx) with a focus on the use of human-engineered (radiomic) features in the task of distinguishing between malignant and benign lesions [8–13]. In reader studies, the impact of added computer-extracted attributes (i.e., features, graphical representations comparing to other cases, and machine-learning-driven lesion signatures indicating a likelihood of malignancy) to radiologists reading tasks in multiple modalities were evaluated [14–17]. CNNs were also investigated for the task of distinguishing biopsy-proven masses from normal tissue on mammograms, similar to other CADx applications [18].

While artificial intelligence (AI) has been developing for interpretation tasks associated with routine medical imaging exams, such as breast cancer screening, for decades, its potential to combat the subjective nature and improve the efficiency of human image interpretation is always expanding. The rapid advancement of computational power and deep learning has dramatically impacted AI research, with promising performance in detection and classification tasks across imaging modalities [19]. Many AI systems based on human-engineered or deep learning methods currently serve as concurrent or secondary readers, i.e., as aids to radiologists for a specific, well-defined task. In the future, AI may be able to perform multiple integrated tasks, making decisions at the level of or surpassing the ability of humans. AI may also serve as a partial primary reader to streamline ancillary tasks, triaging cases or ruling out obvious normal cases. However, before AI is used as an independent, autonomous reader, various challenges must be addressed, including ensuring repeatability and generalizability of algorithms so that AI can provide a significant clinical benefit to imaging tasks across all populations.

1.2 Breast cancer screening and risk assessment

Breast cancer is the most commonly diagnosed cancer for women in the United States, and is estimated to be diagnosed in approximately one in eight women in their lifetime [20]. Mortality of breast cancer has been declining since in the United States since 1989, which can partially be attributed to increased breast cancer screening with mammography [20]. There has an overall 43% reduction in mortality since that time, with annual reductions in mortality from approximately 3%

in the late 1990s to approximately 1% in recent years [20]. Screening mammography helps to reduce mortality by enabling detection at earlier stages, when treatment is more effective and generally less invasive [21,22]. The Society of Breast Imaging and the American College of Radiology recommend annual mammography screening beginning at age 40 for women at average risk, and additional screening is recommended for women at a higher risk [21]. However, guidelines for screening vary from different recommending bodies, as the World Health Organization recommends bi-annual screening for average risk women age 50 to 69 years old [22].

Recommendations in screening frequency and age for initiation of mammographic screening vary across national and international bodies due to the balancing of the harms and benefits of additional screening in the general population [22,23]. Current recommendations in the United States stratify average- from high-risk primarily on the factors of family or personal history of breast cancer, gene mutation status, and history of chest irradiation. However, risk factors derived from mammographic imaging may provide a more personalized method of risk stratification, without increasing the existing need for imaging or testing. Breast density is an example of one such factor that can be assessed from screening mammography and is associated with an increased risk of breast cancer. However, breast density alone provides marginal improvement to screening sensitivity, as over 40% of screening-age women have heterogeneously or extremely dense breasts on mammograms, while the five-year absolute risk of breast cancer for all screening-age groups is 1-2.5% [23,24]. As a result, mammographic density alone has not been found to have high enough specificity in predicting future cancer to be used independently in most accepted risk stratification schemes [21,23].

Breast cancer screening has evolved substantially over the past few decades due to advancements in both image acquisition systems and novel AI algorithms. While AI has been used in interpretation tasks for breast cancer screening since the 1980s, its potential impact is greater now due to the increased memory and computational power of typical clinical workstations and the growing need for enhanced efficiency of interpretation [2,3,7]. The benefit of a medical imaging exam depends on both the quality and interpretation of the image. Inherent limits to labor-intensive human interpretation include errors due to structure noise, incomplete visual search patterns, suboptimal image quality, and fatigue [25,26]. Previous studies have shown that AI tools, including human-engineered radiomic features and deep learning features extracted from mammograms, have additive value to other breast cancer risk assessment metrics, including breast density [27,28]. Radiomics and deep learning features offer ways to quantify a wide variety of parenchymal texture characteristics that are also fast and less subjective to radiologist judgment, as is the case for breast density. In addition, most studies have focused on the images from the year of diagnosis, potentially overestimating the classification ability of such methods for future risk stratification applications [29–31].

1.3 Need for diverse and generalizable data for AI evaluation

The broad adoption of AI in medical imaging research in recent years has prompted a number of new considerations for the clinical utility and ethical use of such methods [32]. While AI methods have been shown to be useful for a wide variety of tasks, the generalizability and robustness of these algorithms are highly dependent upon the data available for model development [33]. Consequently, many academic journals now encourage authors to make datasets and algorithms public when publishing so that algorithms can be compared and evaluated more meaningfully. Resources such as ImageNet, a database with millions of labeled natural images, have greatly accelerated the field of computer vision [34]. However, due to patient privacy regulations, large, well-curated datasets are particularly difficult to obtain in the field of medical imaging. As a result, many studies utilize small, single-institutional datasets, and calculated performance estimates do not readily generalize to other populations or imaging systems [35]. Furthermore, verification of algorithm performance for a specific task presents an ethical burden to both regulators and researchers to assure that data used for testing are independent of data used for development and that data are representative of the population that these algorithms will serve.

A recent and critical example of the impact of data in developing clinically relevant AI systems is the use case of COVID-19. Due to the urgency of biomedical research presented by the COVID-19 pandemic, numerous AI tools were rapidly developed for COVID-related image analysis [36–40]. However, many algorithms developed early in the pandemic were noted to have a high or moderate risk of bias in their reported performance. Multiple reasons were cited for the potential bias, including a lack of representative patients in control groups, non-standard exclusion of patients, and overfitting of models due to the limited available data [36,37,40]. As robustness and generalizability are critical aspects of the clinical utility of an algorithm, to improve upon future AI development in medical imaging of the COVID patient, access to large, population-representative datasets will be of key importance.

1.4 Research scope and goals

In this research, we propose to utilize AI to build an improved risk assessment model for breast cancer mammography screening and to develop a methodology for centralized algorithm performance assessment through the implementation of a sequestered database. AI-based image analysis methods, including human-engineered radiomic features and deep learning-based features, have shown additive value to current breast cancer risk assessment metrics. However, each of these methods, their potential interpretations, and the general challenges of the use of AI must be understood. Thus, this dissertation will begin by presenting the background and the potential limitations of these AI-based methods in Chapter 2.

Applying AI methods to mammography for quantitative risk assessment has the potential to personalize screening efforts and optimize the prevention of disease; however, such developments may be dependent upon the identification of a field effect in the breast parenchyma. Chapter 3 will summarize the characterization of computer-extracted feature relationships among tumor, near-tumor, and far from the tumor parenchymal patterns of women with at least one malignant tumor identified on mammograms and whose mastectomy specimens have been radiographed. To our knowledge, this is the first study to investigate a field effect in mammography and the first to evaluate radiomic features of specimen radiographs and their relation to mammographic features.

To fully understand the potential clinical utility of a mammography-based field effect, parenchymal characteristics of women with cancer must be compared to those of women at average risk. Further, characteristics of the breast parenchyma should be evaluated over time, to identify potential signatures of pre-diagnostic changes in imaging features, compared to images from women who did not develop cancer. Chapter 4 will describe the work accomplished by incorporating temporal imaging data extracted from multiple screening mammograms over time, to improve classification performance in the task of distinguishing women who develop breast cancer from those at average risk.

In recent years, the rapid development of AI models using small, independent datasets has greatly expanded the scope of AI applications in medical imaging but has caused a crisis of reproducibility and generalizability, restricting the clinical translation of these models. The development of a centralized, sequestered database for algorithm performance assessment that contains representative data from multiple institutions seeks to provide a solution to this problem.

Initiation of a sequestered database requires balancing of many demographic variables to create a database that is representative of the population. Additionally, since the proposed database will be used to evaluate various research claims, methods to sample the database for each task and create a custom, representative test set will be necessary. The development and evaluation of the methodology for the creation and sampling of a sequestered database for algorithm testing will be presented in Chapter 5.

A summary of the conclusions and potential future directions of this work will be discussed in Chapter 6. The proposed research aims to investigate methods of improved breast cancer risk assessment by utilizing AI tools such as radiomics and deep learning and integrating these tools within a temporal risk assessment model. We also aimed to develop a methodology for the initiation and utilization of a sequestered database for algorithm performance evaluation and acceleration of clinical translation. Overall, the proposed work in this dissertation will improve the management of cancer patients through the use of artificial intelligence.

CHAPTER 2: A REVIEW OF ARTIFICIAL INTELLIGENCE METHODS UTILIZED

Artificial intelligence refers broadly to the use of computers to learn and perform tasks typically conducted by humans. AI can be subcategorized by the extent of its scope or the learning capability of the system. Most currently available AI systems are considered to be limited learning or narrow AI. These systems perform a single, well-defined task, such as detection, diagnosis, or segmentation, learned from a labeled set of information directly related to the task. On a broader scale, future implementations could potentially enable AI to perform many integrated tasks at an organizational or societal level and to make decisions at the level of or surpassing the ability of humans [32,41].

Machine learning is a subset of AI that utilizes specific programs to identify patterns from an input and learns to make inferences without direct intervention from humans. Conventional machine learning methods in medical imaging use human-engineered radiomic features to characterize an image. These features are extracted from images and can be used as inputs into simple classifiers, i.e., random forest or support vector machines [33,42]. Image features can also be extracted from deep learning networks, a subset of machine learning that directly learns image features from pixel- or voxel-level data. However, these networks contain many learned parameters and components, necessitating large datasets for the training of the network, which are frequently difficult to obtain in medical imaging applications [33,42,43]. Machine learning can be further categorized as supervised or unsupervised learning. In supervised learning, the data on which an algorithm is trained are labeled; in unsupervised learning, the data are unlabeled [44]. Most medical imaging tasks use supervised learning to perform classification, training a network on a set of ground truth labels and then applying the network to a new set of data. Supervised algorithms must be trained on large datasets of well-annotated images to "learn" complex patterns and relationships within the data. As such, many supervised learning tasks in medical imaging are limited in their performance as a result of a lack of annotated data for training due to patient privacy standards [19]. Whereas, unsupervised learning is commonly used for clustering or dimensionality reduction, as it implicitly or explicitly learns underlying probability distributions of the dataset [45–47].

Overall, AI can potentially improve both the efficacy and efficiency of medical imaging through quantitative, reproducible, and objective algorithms. AI techniques are capable of recognizing complex patterns that may be difficult to notice with the human eye; however, to do so, they should be developed to be robust to noise and generalizable to a variety of disease representations [33,48,49]. AI also has the potential to simultaneously interpret data from multiple streams, including images, genomics, and patient history [48]. Techniques for automatic longitudinal monitoring of images, such as sequential mammograms, could lead to personalized care decisions, particularly beneficial for high-risk screening populations. The benefits of improved detection rates, saved time, and profitability are currently challenged by the risk of increased recall rates, increasing costs, and less than favorable perceptions of AI [50]. However, further advances in AI systems could enhance the role of a radiologist by allowing them to focus on "value-added tasks," such as patient interactions and integrated care, rather than interpretation tasks [51].

2.1 Radiomics

In digital images, each pixel is represented in the computer by a numerical grayscale or color value. The resulting numerical representation of the image can then be stored as a matrix, with each row and column representing a pixel in the image and the corresponding numerical value stored in the matrix element. Matrix representations of image data allow for mathematical and computational analysis of medical images, providing the foundation for medical image quantification. Radiomics, specifically human-engineered radiomic features, utilize the numerical relationships of image patterns to characterize image features through a number of expert-generated formulas that may relate to clinically relevant aspects of the image. An example pipeline of the use of radiomics to perform image classification is shown in Figure 2.1. Radiomics, much like other types of AI and ML, can be used for many tasks, such as image characterization, classification, or prognostication.

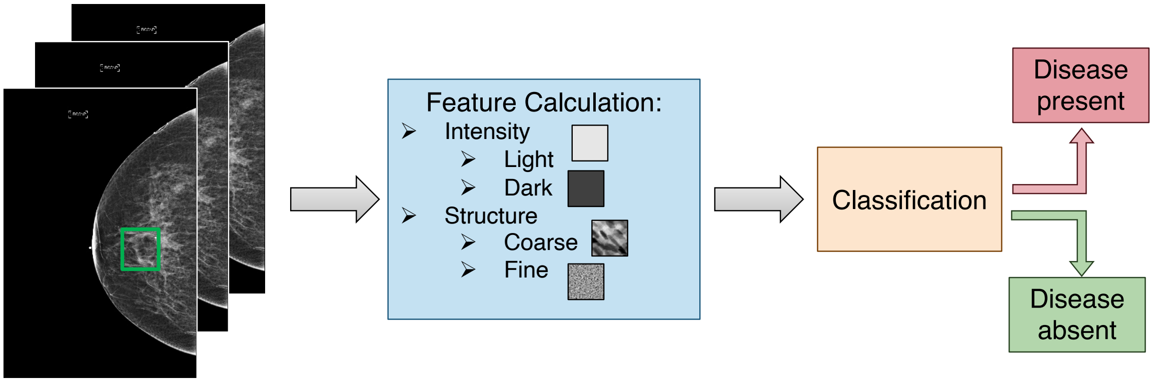


Figure 2.1. Typical pipeline of radiomics used in medical imaging.

Radiomics has been used in medical imaging for decades, with thousands of potential human-engineered features defined [52–54]. Potential feature categories include gray value histogram features, power spectral features, fractal analysis features, and features based on spatial relationships among gray levels, among many others. Commonly used formulas for calculating these features have been described in the literature [55–57]. The subsets of features that may be useful will vary based on the imaging modality and clinical task. A primary advantage of radiomics is that the features are human-defined; their values can be understood through the image characteristics and the feature's equation. This allows for a relatively simple explainability of feature relationships with the variations in the images used to generate the features. However, radiomic features have been found to be quite variable depending on their source definitions and location within an image [58,59]. Thus, the feature definitions utilized should be clearly stated, and region of interest locations should be verified to be robust to slight variations.

2.2 Deep learning

Deep learning is a subset of machine learning based on artificial neural networks with multiple hidden layers, and it has been widely adopted in medical imaging. Deep learning focuses on developing predictive models that can be applied to new, unseen images. In medical imaging, deep learning is used for a variety of tasks, including image classification, segmentation, registration, and generation.

2.2.1 Convolutional neural networks

Convolutional neural networks (CNNs) are a type of deep learning algorithm used in computer vision and image processing tasks. CNNs are designed to process data with a matrix structure, such as an image, and they are particularly well suited for tasks that require the extraction of spatial features and patterns within the data [19,45,60].

The fundamental building block of a CNN is the convolutional layer, which applies a set of filters to the input data to extract spatial features. The filters are learned during the training process and are designed to detect specific patterns or features in the data. After the convolutional layer, the data is usually processed through one or more additional layers, such as pooling layers, which reduce the dimensionality of the data, and fully connected layers, which make the final predictions based on the extracted features [45]. One of the key advantages of CNNs is that they are able to learn hierarchical representations of the input data, where each layer extracts increasingly complex and abstract features from the previous layer. The hierarchical representation learned by the CNN is capable of capturing both local and global features in the data, allowing the network to make robust predictions [19].

2.2.2 Transfer learning

Limitations in performance with deep networks due to dataset size have been partially alleviated in recent years through the use of transfer learning. Transfer learning utilizes networks that are pre-trained on other images, i.e., the millions of natural images (cats, dogs, etc.) available in ImageNet, that can then be used directly to extract generic features from medical images or subsequently fine-tuned to produce features specific to a medical imaging dataset [34,61,62].

When used for feature extraction, all weights of the pre-trained network are frozen at their learned values from the original dataset. Features can then be extracted from selected layers within the network using the new data as input. This process is illustrated in Figure 2.2. The extracted generic features can be used directly or reduced in dimension using common unsupervised learning methods such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), or uniform manifold approximation and projection (UMAP) to create “pseudo-features” that more specifically characterize your dataset [46,47,63–65]. These features can be used in a manner similar to radiomic features, as described previously, for classification or image characterization.

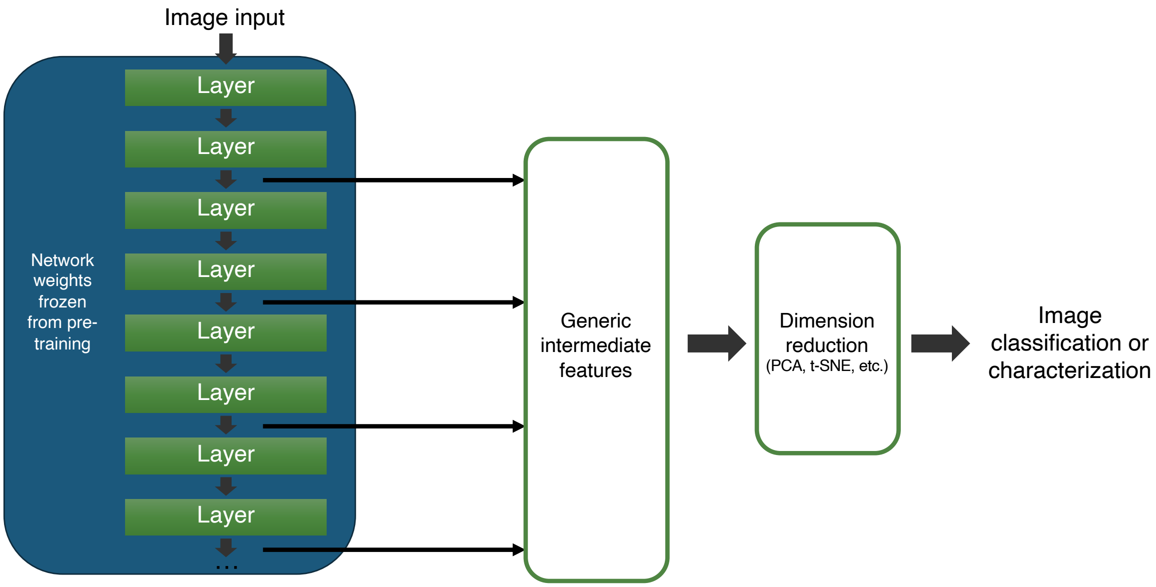


Image input
Layer
Layer
Layer
Layer
Layer
Layer
Layer
Layer
...
Network weights frozen from pre-training
Generic intermediate features
Dimension reduction
(PCA, t-SNE, etc.)
Image classification or characterization

Figure 2.2. General process of feature extraction using transfer learning.

2.2.3 Recurrent neural networks

Recurrent neural networks (RNNs) are a type of deep learning algorithm designed to process sequential data, such as a time series or sequences of words [45]. The unique component of an RNN is the recurrent hidden layer, which is designed to allow information to be passed from one step of the sequence to the next [45,66,67]. Within each step of training, the hidden layer receives information from the previous step and the current input and produces a combined output that is passed on to the next step. As a result, the weights are shared across multiple steps of network training. This allows the RNN to maintain a "summary" of information from all previous steps in the sequence within the hidden layer. RNNs are particularly well suited for tasks that require modeling of temporal dependencies and relationships within the data. This is particularly useful for tasks such as language modeling, where the meaning of a word can depend on the context provided by the preceding words [66,67].

Long-short term memory networks

Long-short term memory (LSTM) networks are a specialized form of RNN developed to improve the preservation of long-term dependencies in the data [68]. The LSTM architecture was motivated by error propagation in RNNs, which showed that backpropagated errors would rapidly grow or decay exponentially, termed the vanishing gradient problem, rendering the long-term dependencies inaccessible [45,68]. The vanishing gradient problem occurs when gradients used to update weights of the network become very small during backpropagation, reducing the network's ability to learn long-term dependencies.

Each LSTM "cell" contains several gates that control the flow of information into and out of the cell using sigmoid functions and element-wise multiplication [45,68]. The forget gate takes as input the current input and the previous hidden state and outputs a number between 0 and 1 for each element in the hidden state. These numbers are used to determine which elements of the previous hidden state should be "forgotten" and which should be retained. The input gate, similarly, takes the current input and the previous hidden state as input and outputs a number between 0 and 1 for each element in the hidden state. However, this output is used to determine how much new information should be added to the hidden state. The output gate takes the current input and the current hidden state as input and outputs a number between 0 and 1 for each element in the hidden state. These numbers are used to determine which elements of the hidden state should be output at the current time step. Finally, a "memory cell" stores the current hidden state. This cell is updated based on the output of the forget gate, the input gate, and the current input [45,66,68].

By using these gates, LSTM networks can selectively remember or forget information from previous time steps, which allows them to handle long-term dependencies more effectively than traditional RNNs. LSTM networks are particularly powerful for processing long sequences of data, such as natural language or time-series weather data, but they have been widely used in various applications, including speech recognition, machine translation, and stock price prediction [66].

2.3 Challenges of applying artificial intelligence to medical imaging

2.3.1 Explainability and interpretability

One critical challenge in AI is the "black-box” nature of algorithms; many physicians are hesitant to accept AI output when the decision-making processes are opaque. To reach full clinical potential, technology needs to be explainable, interpretable, and user-friendly [44,69]. Developers should also consider that various users, including clinicians, researchers, regulators, and insurance providers, will have differing interest in the system’s output, such as disease likelihood, pixel-level activation, data collection method, workflow efficiency, or cost [70,71]. Researchers have found some potential solutions for explainability in medical imaging through applications, which highlight pixels within an image used by the algorithm in its decision-making, i.e., Grad-CAM [72]. Correlating AI output with human descriptions can also help its interpretability. These applications can aid users in understanding why an AI algorithm may be failing in certain instances or populations. Nevertheless, the issue remains of how to trust and explain instances when an algorithm makes a prediction that does not align with the user's (i.e., radiologist's) interpretation of an image, such as highlighting areas outside the body [73].

2.3.2 Robustness and repeatability

Another key challenge focuses on the robustness and repeatability of AI algorithms. Due to the challenging nature of detection and diagnosis in medical images, the performance level of AI systems developed for these tasks may be very sensitive to slight variations in image data. As a result, the output of such algorithms could be perturbed by many factors, e.g., image acquisition parameters, segmentation selection, or biased training data. Robustness and repeatability challenges have been widely documented for systems that use conventional human-engineered radiomic features, as feature definitions and calculation methods can vary widely from system to system [58,59,74]. Deep learning AI methods are not immune to robustness challenges either, as the trained model and classifier performance can be impacted by the training data [75]. In addition, Whitney et. al demonstrated that by bootstrapping classifier prediction scores, robustness of the classifier prediction score may be reduced for cases that are in between the extremes of a prediction score range [75,76].

2.3.3 Generalizability, bias, and harmonization

Similar to the necessity for robustness and repeatability, AI algorithms should also be generalizable to new populations and imaging systems and as free from bias as possible [77]. Acquiring extensive and high-quality datasets for medical imaging is particularly challenging due to the stringent patient privacy regulations in place. As a result, many studies are based on small, single-institution datasets. For AI methods that rely on training with limited available data, performance estimates can result and may not readily generalize to other populations or imaging systems. A few publicly available image repositories, including the Medical Imaging and Data Resource Center (MIDRC) and The Cancer Imaging Archive (TCIA), aim to alleviate this challenge by providing equitable access to a diverse population of imaging studies for a variety of diseases and clinical tasks [78,79]. However, these resources require the initiative of researchers and clinicians to adopt an image-sharing culture for society to truly benefit from the power of AI in medical imaging. Conversely, to maintain useful and trusted outputs, it may be best to develop algorithms for specific tasks or acquisition systems rather than general systems [70]. Standardized training and testing protocols could be established to determine the generalizability of models, and it is important to evaluate the performance of the computer algorithm as well as the end users when they are interpreting images with and without the AI system [44,70].

2.3.4 Ethical implementation and integration

Other challenges include the ethical use and integration of AI systems into the clinical setting as most AI systems are not yet approved by the FDA or are approved for a narrow, specific application. The user has the ethical obligation to implement approved algorithms only as they are intended, including using an algorithm only with appropriate images and use cases and not for "off-label" applications. Also, clinical workflows may need to be modified to account for altered workflows, such as switching from reading cases manually to reading with an AI aid. Clinicians and hospitals may need to construct new billing codes for such AI tools, and future investigations should evaluate the clinical and financial impact of AI on radiologists and patients across healthcare systems [50].

2.4 Conclusions

In summary, AI is a rapidly developing tool widely used in medical imaging applications. This chapter reviewed the technical details of several AI-methods utilized in this work including radiomics, transfer learning, and recurrent neural networks. Despite challenges and concerns, AI has the potential to improve medical imaging tasks and further advance patient care decisions.

CHAPTER 3: BREAST CANCER FIELD EFFECT IN MAMMOGRAPHY USING HUMAN-ENGINEERED RADIOMIC FEATURES AND DEEP LEARNING

3.1 Introduction

To screen for breast cancer, annual mammography or digital breast tomosynthesis screening is recommended by the American College of Radiology starting at age 40 years for average-risk women [21,80]. As discussed in Chapter 1, current risk assessment models primarily use clinical factors and personal or family history of breast cancer. However, AI-based metrics from mammography images have been shown to potentially provide additive value to such risk assessment models. Many approaches for quantifying risk using AI metrics have been developed; however, it appears that none have investigated a potential cancer field effect in mammography as a signature of breast cancer risk.

In women with biopsy-proven breast cancer, histologically normal areas of the parenchyma within the ipsilateral (and the contralateral) breast have shown molecular similarity to the tumor, supporting a potential cancer field effect [81,82]. It is hypothesized that such an effect may be a precursor of malignancy or impact tumor recurrence [83]. A field cancerization that is identifiable via mammography, and confirmation of the distance to which this cancerization extends into the normal adjacent tissue for various breast cancer subtypes, has yet to be confirmed in the literature. Identification of a cancer field effect in mammography has the potential to provide a novel approach to the stratification of breast cancer risk in the general population by augmenting current risk assessment models. Radiomic texture analysis and deep learning are particularly well suited to identify and characterize potential signatures of a field effect in mammography due to their ability to quantify a multitude of image characteristics.

The work included in this chapter aims to characterize relationships among tumor, near-tumor, and far from the tumor parenchymal patterns of women with at least one malignant tumor identified in mammograms with corresponding radiographs of mastectomy specimens. Human-engineered radiomic features and deep learning features were used to identify and characterize texture signatures of a field effect in mammograms and specimen radiographs of women with biopsy-proven breast cancer. Results aim to characterize the parenchymal field in women with confirmed breast cancer at the time of diagnosis by comparing similarities and variations in parenchymal texture as a function of distance from the tumor. Analysis of specimen radiographs using texture analysis, while not clinically implementable, will provide a more fundamental understanding of parenchymal field relationships. Thus, this is the first study to investigate a field effect in mammography and the first to evaluate radiomic features of specimen radiographs and their relation to in-vivo mammogram features.

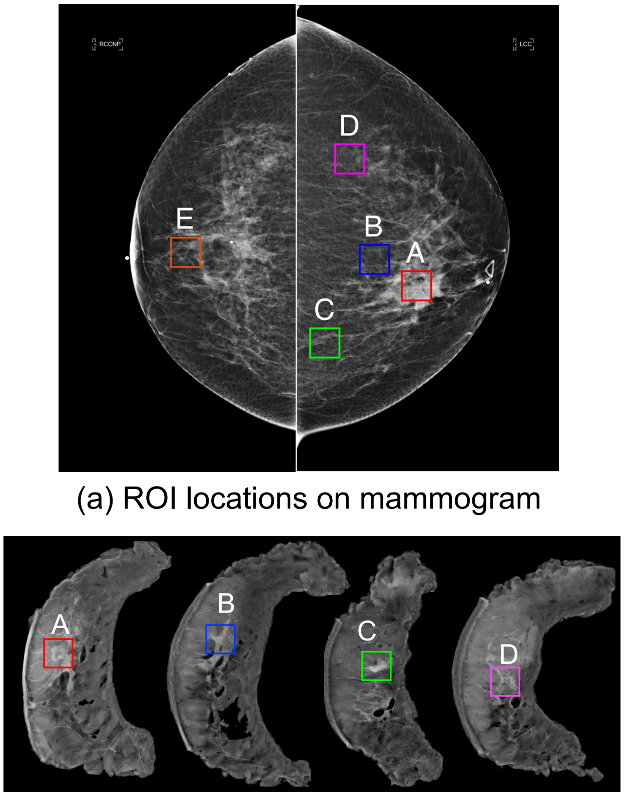
3.2 Methods

3.2.1 Database

The dataset consisted of 103 retrospectively collected patients with at least one identified malignant tumor. Inclusion in the initial cohort was specified by patients who were diagnosed with breast cancer and had undergone mastectomy for treatment of their breast cancer at MD Anderson Cancer Center between 2010 and 2017. Preoperative mammograms and intraoperative radiographs of the mastectomy specimens were retrieved under Health Insurance Portability and Accountability Act (HIPAA)-compliant Institutional Review Board (IRB) protocols. Patients with tumors occult on the craniocaudal (CC) view (n = 18), no preoperative mammogram available (n

= 8), preoperative mammogram not for presentation (n = 2), and breast region too small to fit ROIs (n = 1) were excluded. The remaining 74 patients were used in the analysis. In addition, a subset of 32 patients had also undergone intraoperative radiographic imaging of the mastectomy specimen. In a conventional clinical setting, specimen radiographs are used to verify removal of the targeted abnormality and to evaluate the margins of the resection. Although evaluating specimen radiographs is not typical clinical practice for risk assessment, radiomic features of the tissue in- and ex-vivo will allow for a deeper understanding of the relationships of tissue texture for a potential cancer field effect. Mammograms were acquired with a Hologic Lorad Selenia system (12-bit quantization, 70-micron pixels), and specimen radiographs were acquired with a Fuji imaging system (12-bit quantization, 50-micron pixels).

To characterize a potential field effect, regions of interest (ROIs) of 128x128 pixels were selected from four regions in the craniocaudal mammogram: within the tumor (A), near to the tumor (B), far from the tumor (C and D), and behind the nipple on the contralateral breast (E), as shown in Figure 3.1. Tumor locations were identified with the assistance of a research specialist with over 15 years of experience in mammography. For each paired mammogram and specimen radiograph analysis, corresponding 128x128-pixel ROIs were selected by a breast surgical oncologist with over 20 years of experience in the field from three regions across the serially sectioned specimen radiographs, as shown in Figure 3.1. Specimen radiograph ROI locations were selected by a breast surgical oncologist with over 20 years of experience in the field, and ROIs near to the tumor (B) were designated as less than 2 cm from the tumor, while ROIs far from the tumor (C and D) were designated as greater than 2 cm from the tumor.



(a) ROI locations on mammogram

(b) ROI locations on specimen radiograph

Figure 3.1. Examples of ROI locations for (A) tumor, (B) near to tumor, (C and D) far from tumor regions, and (E) in the central region behind the nipple on the contralateral breast depicted on (a) cranio-caudal-view mammogram and on (b) specimen radiograph. The specimen radiograph shows four serial sections of breast tissue from the same breast shown in (a).

3.2.2 Feature Extraction

An in-house AI workstation was used to automatically extract 45 radiomic texture features describing tissue contrast/intensity and structure in each breast region. Table 3.1 gives the category, name, and brief description of all 45 radiomic features calculated. More detailed feature descriptions and formulas can be found in the literature [55–57]. These features are based on (a) fractal analysis, including box-counting and Minkowski methods; (b) edge-frequency analysis; (c) gray-level histogram analysis; (d) Fourier transform analysis; (e) the neighborhood gray-tone difference matrix; (f) Powerlaw beta from power spectral analysis; (g) the gray-level co-occurrence matrix (GLCM). For deep learning-based features, a transfer learning approach was used. A VGG19 convolutional neural network (CNN) architecture was first pre-trained on ImageNet [34]. The generically trained network was then used with the mammogram and specimen radiograph ROIs as the input, and 1472 generic deep learning features were extracted from each of the five max-pooling layers, similar to the approach described by Antropova et al. [62]. To select only features relevant to each data set, deep learning features with zero variance or features in which greater than 50% of the values were zero were removed. To further reduce the number of features, principal component analysis (PCA) utilized to reduce dimensionality of the remaining features. The first 20 principal components (86.53% of the total variance for mammograms and 89.67% for specimen radiographs) were used as pseudo-features, i.e., principal components take to be characteristic features, for each region.

Table 3.1. Categories, names, and brief descriptions of 45 radiomic features calculated for each ROI.

Category	Feature Name	Description
(a) Fractal analysis, including box-counting and Minkowski methods	Boxcounting Dimension	Fractal dimension estimated based on box-counting method
	Boxcounting Dimension 1	Fractal dimension estimated based on box-counting method
	Boxcounting Dimension 2	Fractal dimension estimated based on box-counting method
	Boxcounting Dimension 3	Fractal dimension estimated based on box-counting method
	Boxcounting Dimension 4	Fractal dimension estimated based on box-counting method
	Boxcounting Dimension 5	Fractal dimension estimated based on box-counting method
	Minkod Global MD	Fractal dimension estimated based on Minkowski method

Table 3.1. (Continued) Categories, names, and brief descriptions of 45 radiomic features calculated for each ROI.

(b) Edge-frequency analysis	Edge Frequency: Mean Gradient	Average of edge gradient
	Edge Frequency: Max Gradient	Maximum of edge gradient
	Edge Frequency: Min Gradient	Minimum of edge gradient
	Edge Frequency: St. Dev. Gradient	Standard deviation of edge gradient
(c) Gray-level histogram analysis	Histogram Average	Average gray value within region of interest
	Histogram Max CDF	Gray level threshold yielding 95% of the area under the histogram of the region
	Histogram Min CDF	Gray level threshold yielding 5% of the area under the histogram of the region
	Histogram Balance	Ratio of (95% threshold-Average) to (Average-5% threshold)
	Histogram Seventy CDF	Gray level threshold yielding 70% of the area under the histogram of the region
	Histogram Thirty CDF	Gray level threshold yielding 30% of the area under the histogram of the region
	Histogram Quasi Balance	Ratio of (70% threshold-Average) to (Average-30% threshold)
	Histogram Skewness	Denseness measure used to characterize local tissue composition
(d) Features based on Fourier transform analysis	Fourier Root Mean Square (FRMS)	Root-mean-square variation based on Fourier transform analysis
	Fourier First Moment of Power Spectrum (FFMP)	First moment of power spectrum based on Fourier transform analysis
(e) Neighborhood gray-tone difference matrix	Coarseness	Coarseness measure calculated from neighborhood gray-tone difference matrix
	Contrast	Contrast measure calculated from neighborhood gray-tone difference matrix
(f) Powerlaw beta from power spectral analysis	Powerlaw Beta 1	Exponent beta estimated based on powerlaw spectrum analysis
	Powerlaw Beta 2	Exponent beta estimated based on powerlaw spectrum analysis
	Powerlaw Beta 3	Exponent beta estimated based on powerlaw spectrum analysis

Table 3.1. (Continued) Categories, names, and brief descriptions of 45 radiomic features calculated for each ROI.

(f) Powerlaw beta from power spectral analysis (Continued)	Powerlaw Beta 4	Exponent beta estimated based on powerlaw spectrum analysis
	Powerlaw Beta 5	Exponent beta estimated based on powerlaw spectrum analysis
	Powerlaw Beta 6	Exponent beta estimated based on powerlaw spectrum analysis
	Powerlaw Beta 7	Exponent beta estimated based on powerlaw spectrum analysis
	Powerlaw Beta 8	Exponent beta estimated based on powerlaw spectrum analysis
(g) Gray-level co-occurrence matrix (GLCM)	GLCM Contrast	Measure of local image variations
	GLCM Correlation	Measure of image linearity
	GLCM Difference Entropy	Measure of the randomness of the difference of neighboring pixels' gray-levels
	GLCM Difference Variance	Measure of variations of difference of gray-levels between pixel-pairs
	GLCM Energy	Measure of image homogeneity
	GLCM Entropy	Measure of the randomness of the gray-levels
	GLCM Homogeneity	Measure of the image homogeneity
	GLCM Information measure of correlation 1 (IMC1)	Measure of nonlinear gray-level dependence
	GLCM Information measure of correlation 2 (IMC2)	Measure of nonlinear gray-level dependence
	GLCM Max Correlation Coefficient	Measure of nonlinear gray-level dependence
	GLCM Sum Average	Measure of the overall image brightness
	GLCM Sum Entropy	Measure of the randomness of the sum of gray-levels of neighboring pixels
	GLCM Sum Variance	Measure of the spread in the sum of the gray-levels of pixel-pairs distribution
	GLCM Variance	Measure of the spread in the gray-level distribution

3.2.3 Statistical Analysis

To assess correlation of features between ROI regions, the Kendall’s Tau-b correlation test was used. This test allows for quantification of correlation between a categorical independent variable (ROI region) and a numerical dependent variable (feature values); thus, it was selected for evaluating correlations in mammograms and specimen radiographs separately. Kendall’s Tau-b is a nonparametric measure of the strength and direction of the association between two variables and is considered an alternative to the Spearman rank order correlation coefficient for data with many numerical ties in each group [84].

To evaluate correlations in features between mammograms and specimen radiographs, the Pearson correlation test was used [84,85]. Pearson’s Rho is a commonly used measure of linear correlation between two variables [85]. This test allows for quantification of correlation between two numerical variables, which is why it was selected to evaluate feature correlation between both modalities. For calculation of Pearson’s Rho, only matched pairs of patients and corresponding ROI regions with both mammograms and specimen radiographs were used (n = 32 patients, 118 ROIs).

The Pearson correlation test was also used to evaluate correlation between radiomic and deep learning features (only from the mammogram) for a more comprehensive understanding of the deep learning features evaluated. Strong correlations between radiomic and deep learning features may indicate that the deep learning feature described similar characteristics of the parenchyma as the radiomic feature, such as intensity or structure.

Both the test statistics of Kendall’s Tau-b and Pearson’s Rho are bounded between -1 and 1, with values of zero indicating no correlation and one indicating a strong correlation, with the sign indicating the direction of the relationship. All hypothesis tests were adjusted for multiple comparisons using the Benjamini-Hochberg correction [86]. This procedure controls for the false discovery rate (FDR), the proportion of significant results that are actually false positives. The Benjamini-Hochberg correction is recommended when the number of comparisons is large and is commonly used in exploratory procedures, such as identifying differentially expressed genes [86,87] In this correction, to be considered significant, the *p*-value must be less than the rank of said *p*-value (the smallest *p*-value would have a rank of 1, and the greatest *p*-value would have a rank of the total number of comparisons) divided by the total number of comparisons, multiplied by the selected FDR. Since this was completed for each set of tests, 45 was the total number of comparisons for radiomic features and 20 was the total number of comparisons for deep learning features. A FDR of 5% was selected to keep the number of potential false discoveries low, whereas FDRs of 10% to 25% are commonly used in genomic studies [87].

Preliminary analysis on the statistical similarity of features across the mammogram was also evaluated, as described in Appendix Section 1. From this analysis, it was determined that there was broad similarity in the feature distribution shapes using the Kolmogorov-Smirnov (KS) test. However, additional analysis evaluating the impact of shifting distributions to align the means before conducting a KS test revealed key differences in absolute (non-shifted) feature values between tumor and non-tumor regions. As such, the analysis described here was completed using absolute values of features, without shifting distributions.

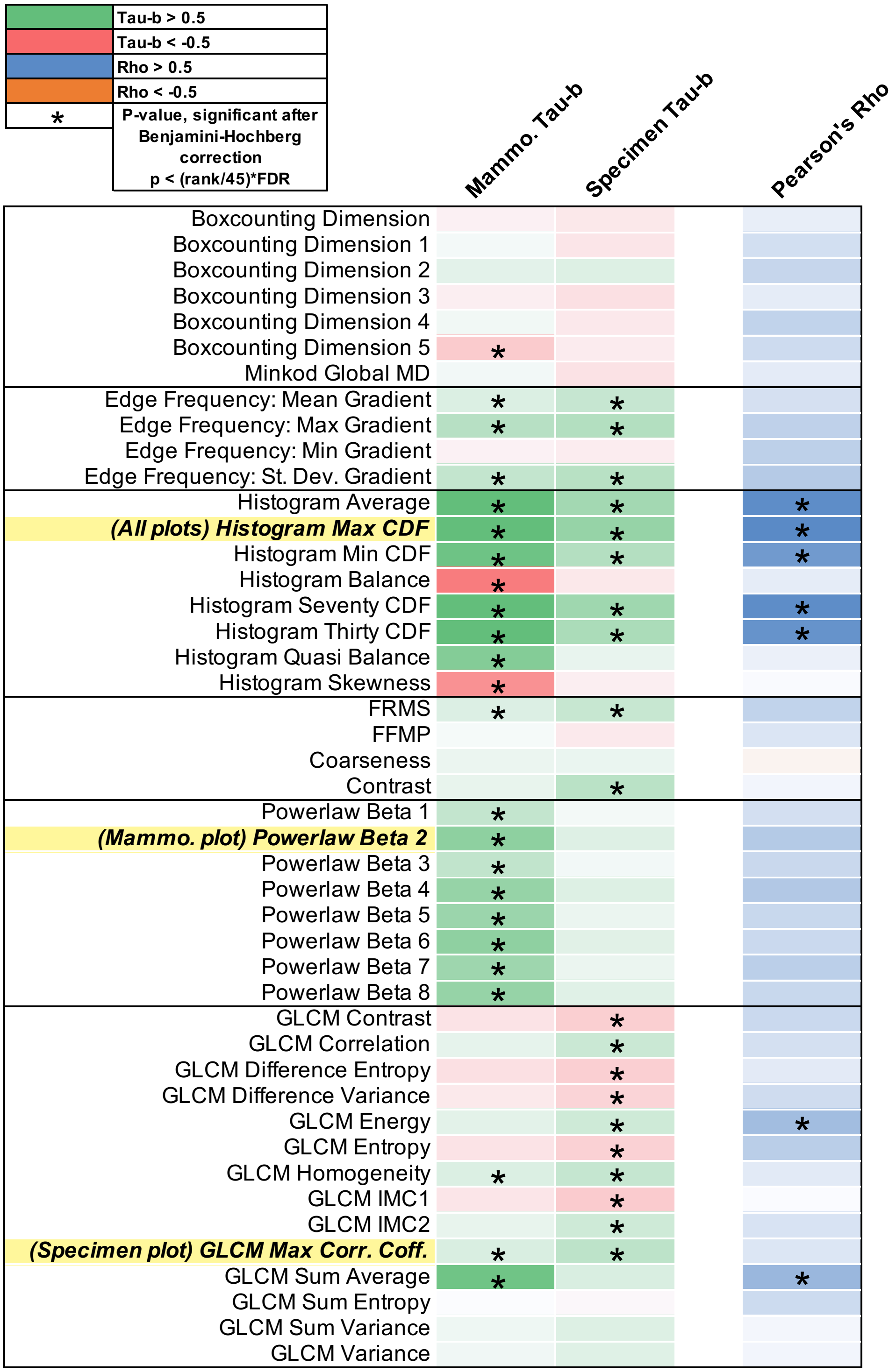
3.3 Results

3.3.1 Correlation of radiomic features between ROI regions and image modalities

Results of Kendall’s Tau-b and Pearson correlation tests for all calculated radiomic features are shown in Figure 3.2. Features were grouped into categories representing similar underlying characteristics. The color of cells for a given comparison represents the magnitude of the test statistic. For Kendall’s Tau-b, more saturated green cells represent stronger positive correlations, and more saturated red cells represent stronger negative correlations. Similarly, for Pearson correlation, more saturated blue cells represent stronger positive correlations, and more saturated orange cells represent stronger negative correlations. All color scales reached a maximum color saturation at a value of +/- 0.5 and are shown in white for test statistics equal to zero. Asterisks in each cell represent correlations considered significant after Benjamini-Hochberg correction using a 5% FDR. It is important to note that statistical significance here is for the purposes of discovery only, not to indicate a clinical difference between two groups. Results of this test emphasize changes in the absolute values of features across the ROI regions.

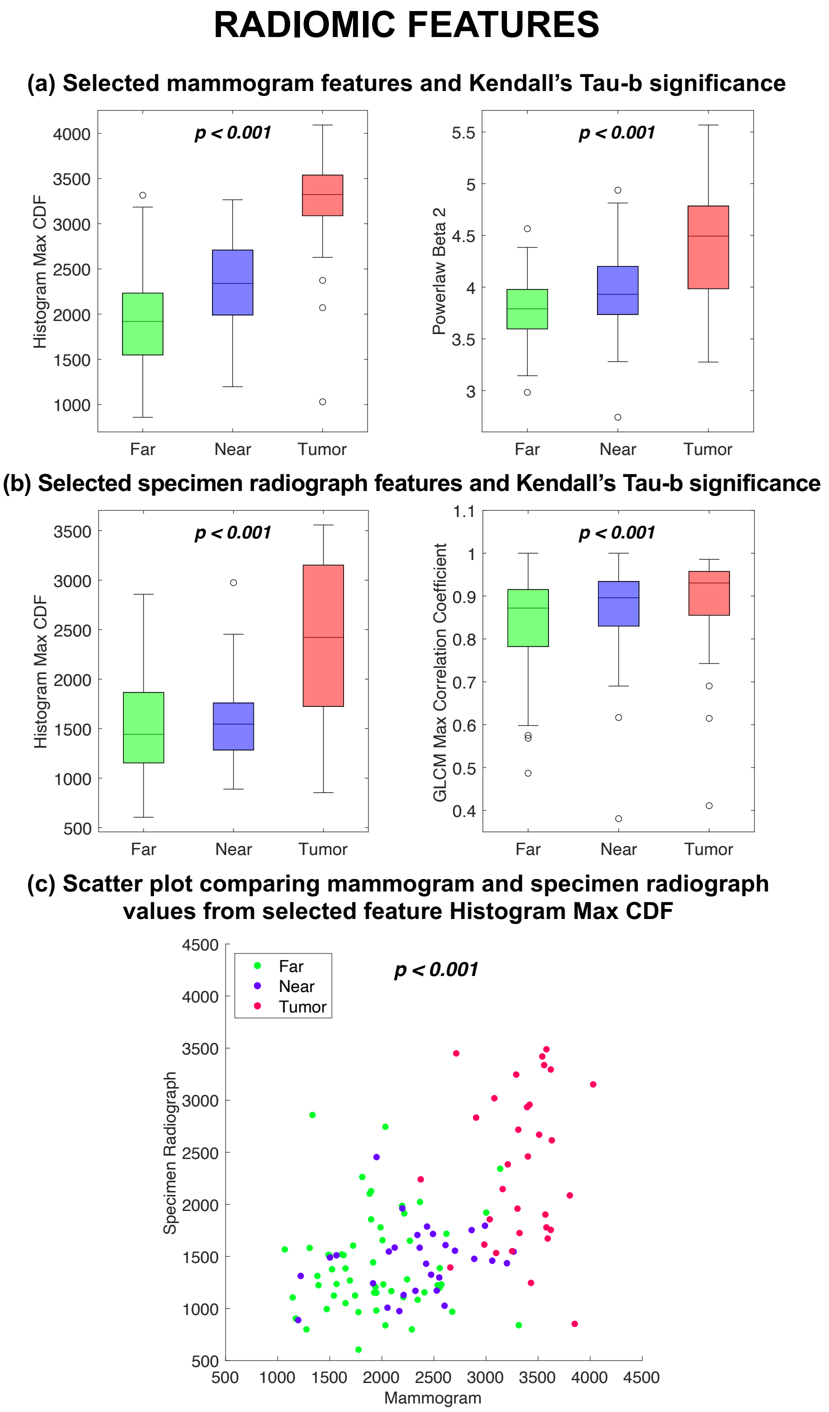
For radiomic feature analysis, Kendall’s Tau-b test results indicated a majority of statistically significant correlations between the tumor, near, and far regions in mammograms for intensity-based histogram features, edge frequency features, and Fourier-based power-law beta features. In the specimen radiographs, results indicated a majority of statistically significant correlations between intensity-based histogram features, edge frequency features, and GLCM features. Distinct subgroups showing statistical significance of power-law beta features and GLCM features in mammograms and specimen radiographs, respectively, demonstrate key differences in the results between the two modalities. Pearson correlation results identified a majority of statistically significant correlation in intensity-based histogram features between mammograms and specimen radiographs, presenting a strong relationship across both modalities' tumor, near, and far regions. This result seems reasonable, given that tumors have been found to be more dense and coarser in texture than parenchymal tissue, while indicating strong correlations across tumor and non-tumor tissue [27,88,89].

Highlighted features in Figure 3.2 were selected significant correlations from radiomic feature analysis and are plotted in Figure 3.3. Correlations indicated that radiomic features from ROIs closer to the tumor tended to show more similarity to the tumor than features from far ROIs and showed strong relationships of these features across the parenchymal field in in- and ex-vivo imaging. Features were selected as follows: (1) Histogram Max CDF – the strongest correlation for all mammogram and all specimen Kendall's Tau-b and Pearson tests, (2) Powerlaw Beta 2 – the strongest correlation in a subgroup of features where only the mammogram Kendall's Tau-b test indicated statistical significance in a majority of the features, (3) GLCM Max Correlation Coefficient – the strongest correlation in a subgroup of features where only the specimen Kendall's Tau-b test indicated statistical significance in a majority of the features.



	Mammo. Tau-b	Specimen Tau-b	Pearson's Rho
Boxcounting Dimension			
Boxcounting Dimension 1			
Boxcounting Dimension 2			
Boxcounting Dimension 3			
Boxcounting Dimension 4			
Boxcounting Dimension 5	*		
Minkod Global MD			
Edge Frequency: Mean Gradient	*	*	
Edge Frequency: Max Gradient	*	*	
Edge Frequency: Min Gradient			
Edge Frequency: St. Dev. Gradient	*	*	
Histogram Average	*	*	*
(All plots) Histogram Max CDF	*	*	*
Histogram Min CDF	*	*	*
Histogram Balance	*		
Histogram Seventy CDF	*	*	*
Histogram Thirty CDF	*	*	*
Histogram Quasi Balance	*		
Histogram Skewness	*		
FRMS	*	*	
FFMP			
Coarseness			
Contrast		*	
Powerlaw Beta 1	*		
(Mammo. plot) Powerlaw Beta 2	*		
Powerlaw Beta 3	*		
Powerlaw Beta 4	*		
Powerlaw Beta 5	*		
Powerlaw Beta 6	*		
Powerlaw Beta 7	*		
Powerlaw Beta 8	*		
GLCM Contrast		*	
GLCM Correlation		*	
GLCM Difference Entropy		*	
GLCM Difference Variance		*	
GLCM Energy		*	*
GLCM Entropy		*	
GLCM Homogeneity	*	*	
GLCM IMC1		*	
GLCM IMC2		*	
(Specimen plot) GLCM Max Corr. Coff.	*	*	
GLCM Sum Average	*		*
GLCM Sum Entropy			
GLCM Sum Variance			
GLCM Variance			

Figure 3.2. Color scale plot of Kendall’s Tau-b and Pearson correlation test results for radiomic features calculated from mammograms and specimen radiographs. The color of each cell represents the direction and strength of each correlation, as noted in the legend. The asterisks denote correlations considered significant after Benjamini-Hochberg correction with a 5% FDR.



(a) Selected mammogram features and Kendall's Tau-b significance

(b) Selected specimen radiograph features and Kendall's Tau-b significance

(c) Scatter plot comparing mammogram and specimen radiograph values from selected feature Histogram Max CDF

Figure 3.3. Boxplots of selected radiomic features and Kendall's Tau-b significance in mammograms (a) and specimen radiographs (b). Scatterplot of selected intensity-based histogram feature Max CDF, which had the strongest correlation between mammograms and specimen radiographs, and Pearson's Rho significance (c). Significant correlations among tumor, near, and far regions indicate relationships of feature values with proximity to the tumor.

3.3.2 Correlation of deep learning features between ROI regions and image modalities

Results of Kendall's Tau-b and Pearson correlation tests for all calculated deep learning features are shown in Figure 3.4. Since deep learning features cannot be easily categorized nor do they possess intuitive meanings as the radiomic features do, it is important to note that the deep learning features represent principal components and thus are listed in order of decreasing variance. The color scales and markers used to indicate correlation strength, direction, and significance are the same as described for radiomic features in Figure 3.2.

For the deep learning feature Kendall's Tau-b test, results indicated statistically significant correlations between the tumor, near, and far regions in mammograms for the first three features and feature 7. In specimen radiographs, results indicated a statistically significant correlation in only a single feature (feature 2). Pearson correlation results showed a statistically significant correlation in feature 1 between mammograms and specimen radiographs. These results seem reasonable, given that the first principal components/features will describe the majority of the variance in the dataset and fundamental characteristics of the images [46].

Highlighted features in Figure 3.4 were plotted in Figure 3.5 to demonstrate significant correlations from deep learning feature analysis. In agreement with the results from the radiomics feature analysis, correlations indicated that deep learning features from ROIs closer to the tumor tended to show more similarity to the tumor than features far from the tumor in both the in- and ex-vivo imaging.

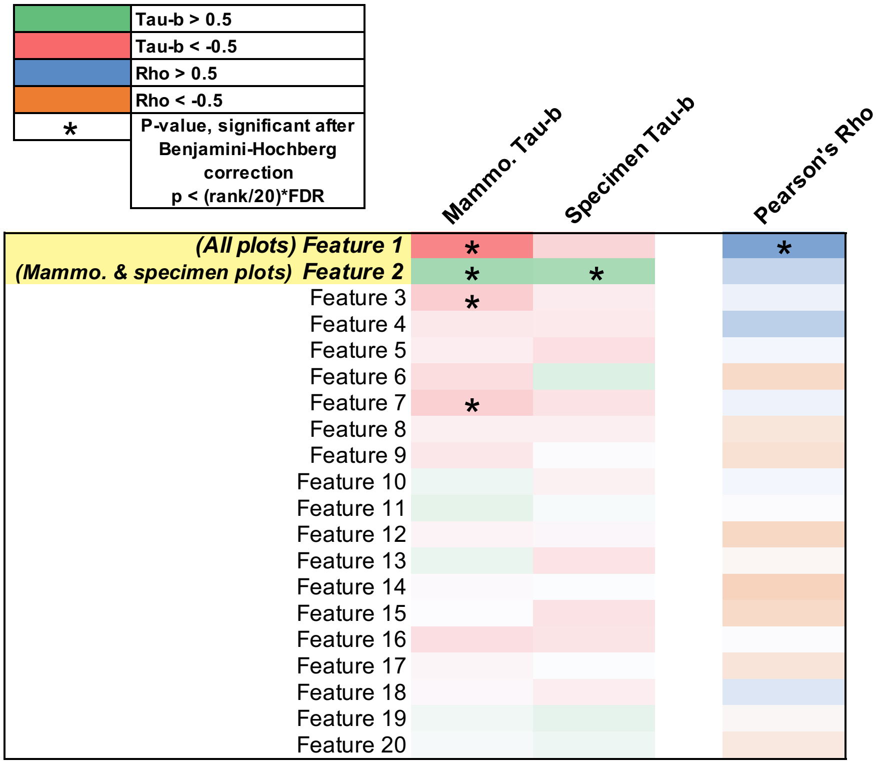
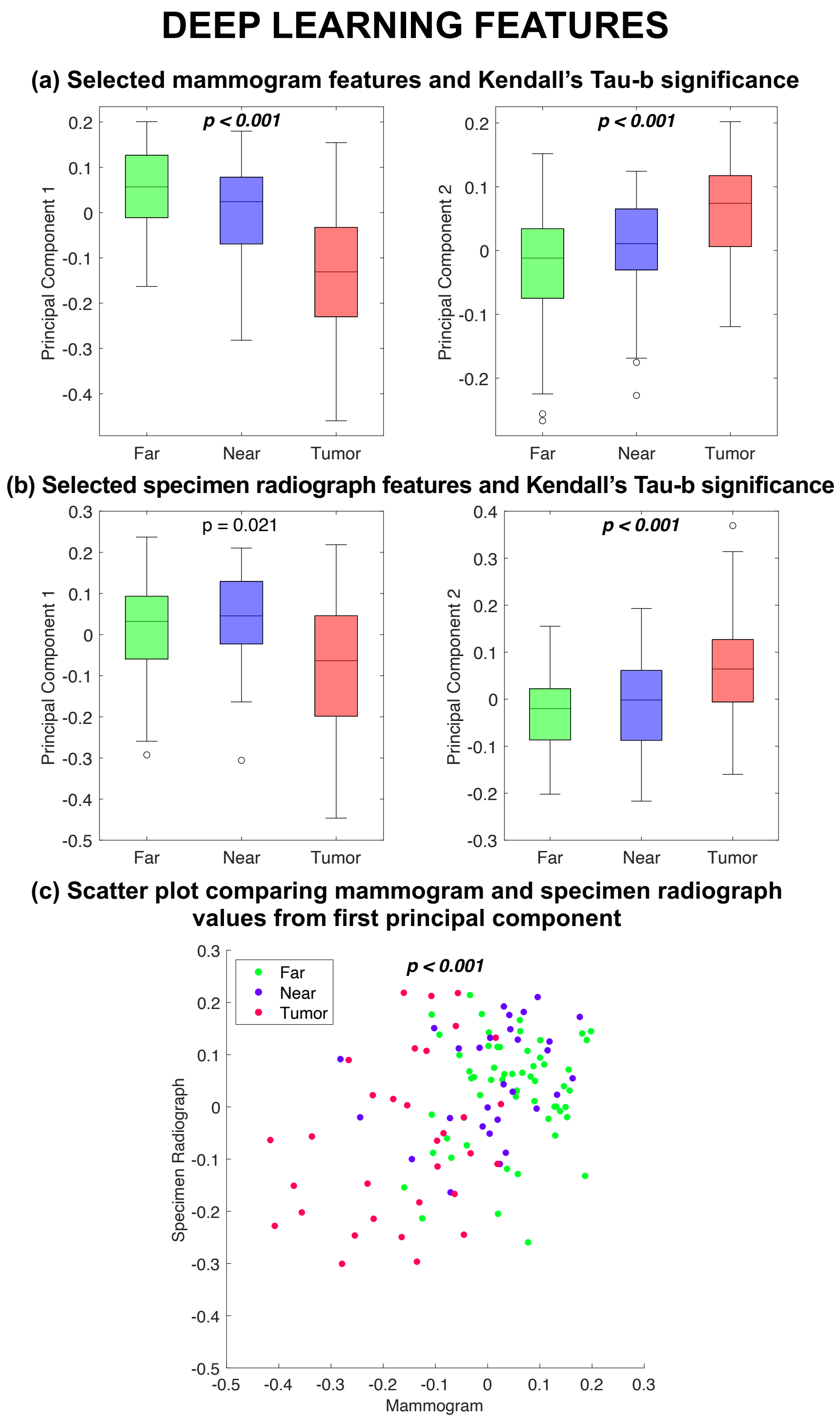


Figure 3.4. Color scale plot of Kendall's Tau-b and Pearson correlation test results for deep learning features calculated from mammograms and specimen radiographs. The color of each cell represents the direction and strength of each correlation, as noted in the legend. The asterisks denote correlations considered significant after Benjamini-Hochberg correction with a 5% FDR.



DEEP LEARNING FEATURES

(a) Selected mammogram features and Kendall's Tau-b significance

(b) Selected specimen radiograph features and Kendall's Tau-b significance

(c) Scatter plot comparing mammogram and specimen radiograph values from first principal component

Figure 3.5. Boxplots of selected deep learning features and Kendall's Tau-b significance in mammograms (a) and specimen radiographs (b). Scatterplot of the first principal component, taken to be feature 1, which had the strongest correlation between mammograms and specimen radiographs, and Pearson's Rho significance (c). Significant correlations among tumor, near, and far regions indicate relationships of feature values with proximity to the tumor.

3.3.3 Correlation between radiomic and deep learning features

The Pearson correlation test was used to evaluate correlation between human-engineered radiomic features and the 20 principal components, taken to be the deep learning pseudo-features. Results are shown in the color scale plot, Figure 3.6. Within the plot, values closer to negative one, representing stronger negative correlations, were shown in red; values closer to one, representing stronger positive correlations, were shown in green; and values in the middle (closer to 0.5) were represented in white. Colors were scaled linearly in proportion to their numerical values. Mammogram features from all ROI regions were combined for this analysis.

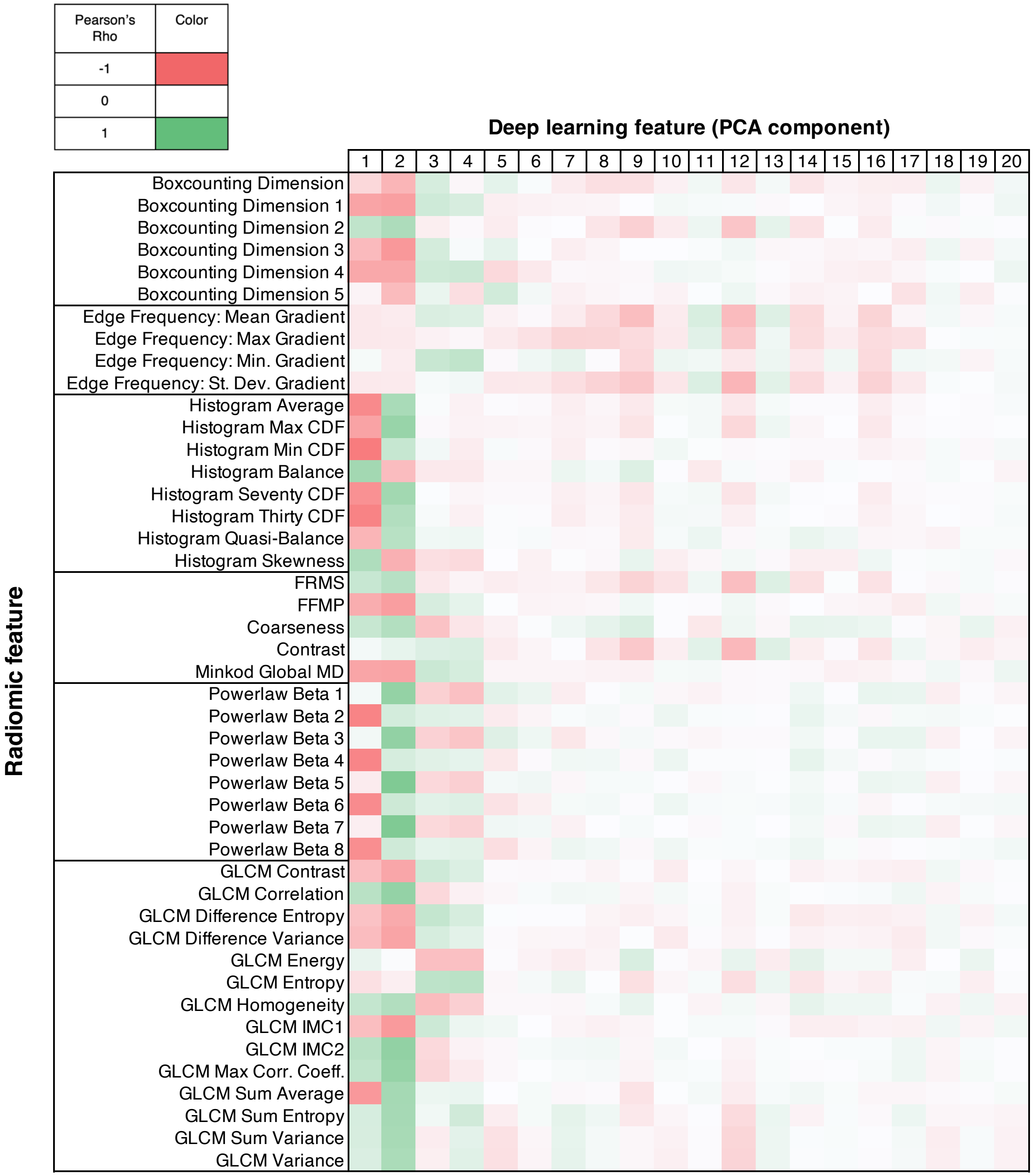


Figure 3.6. Pearson correlation results for all mammographic features between radiomic-based and deep learning-based features. The color of each cell represents the direction and strength of each correlation, as noted in the legend.

3.4 Discussion

3.4.1 Correlation of features between ROI regions and image modalities

Kendall’s Tau-b and Pearson correlation results from mammograms and specimen radiographs exemplify relationships of the parenchymal field in women with cancer. Understanding these features and relationships provides further important information in understanding a potential mammography-based cancer field effect. Correlation results from radiomic features and deep learning principal component features showed evidence of a relationship between feature values and ROI location with increasing distance from the tumor.

For radiomic features calculated within the mammogram, statistically significant correlations were primarily identified in histogram or intensity-based features, edge frequency features, and Fourier-based features using Kendall’s Tau-b. Similarly, for deep learning features calculated within the mammogram, the first three and seventh principal components indicated statistically significant correlations. Although the underlying characteristics of principal component features cannot be understood in the same way as radiomic features, the first principal components will represent the foundational characteristics of the object [46]. Thus, it could be reasonable to infer that the first principal components may also quantify the brightness of the pixels from a given ROI are, describing the tissue intensity as well. This relationship in intensity across the mammographic field may be related to underlying density of breast tissue in a given region, as the tumor tissue has been shown to be denser, and therefore brighter in a radiographic image [27,88,89]. Correlations across the field in Fourier-based features, as seen in the mammogram radiomic features, have not been well documented in the literature and should be investigated further to understand how these relationships may relate to a potential field effect and breast cancer risk.

Kendall’s Tau-b correlation in radiomic features extracted from specimen radiographs showed slightly different results than those found in the mammograms. Similar to the mammogram results, the intensity-based and edge frequency features showed significant correlation with increasing distance from the tumor. The significant correlation in the intensity-based features in specimen radiographs could similarly be attributed to the underlying tissue density as represented in a brighter area in radiographic imaging, as the solid tumor mass tends to be denser, and therefore brighter [88]. However, the significant correlations in GLCM features, as opposed to the Fourier-based features seen in the mammogram analysis, show a different aspect of the tissue structure exhibiting the correlation with increasing distance from the tumor. This change could due to structural changes of tissue after excision from the body or due to imaging variations resulting from use of a different system. Only the second feature or principal component showed a statistically significant correlation with ROI region location for deep learning features extracted from specimen radiographs. Figure 3.5 shows that the first feature indicated lower values on average for tumor ROIs than near or far ROIs but did not show as linear a correlation with ROI region as feature 2 and was not statistically significant after multiple comparisons correction. Since these features do not have intuitive meanings in the same way that radiomic features do, the exact reasoning for this is not fully understood and may be investigated further in future study. However, the first principal component for specimen radiographs may describe a characteristic that is not strongly correlated with ROI regions, like many radiomic feature categories are also not strongly correlated with ROI regions. Although evaluation of radiographs of mastectomy specimens is not standard clinical practice for risk assessment, investigation of these features offers a unique opportunity to gain a fundamental understanding of the field effect in both in- and ex-vivo imaging.

Further, to our knowledge , this is the first study to evaluate radiomic features of mastectomy specimens.

Pearson correlation analysis between features calculated from mammograms and specimen radiographs extended the results and demonstrated statistically significant correlations in both feature types across in- and ex-vivo imaging. This was shown primarily in the intensity-based histogram features, indicating radiomic features describing tissue intensity were highly correlated between mammograms and specimen radiographs. Structure-based radiomic features did not show these significant correlations between modalities, which may be explained by changes in the tissue presentation when excised from the body or changes resulting from the use of a different imaging system. For deep learning features, only the first feature reached statistical significance. Given that the first feature is the first principal component, it describes the largest percentage of variance of all deep learning features, indicating a correlation between fundamental characteristics of the mammographic and specimen radiograph deep learning features.

3.4.2 Correlation between radiomic and deep learning features

Correlation analysis of deep learning-based features and radiomic features revealed that the first two features, or principal components, of the deep learning features were strongly correlated with many of the radiomic feature values from all mammogram ROI regions. As mentioned previously, the first principal components are known to represent the foundational characteristics of the object [46]. It is somewhat unsurprising then, that the undefined, foundational characteristics of the ROI images are strongly correlated with human-engineered definitions of image texture and intensity. However, this result does aid our general understanding of deep learning-based features as they are not simply explainable with mathematical formulations as are radiomic features. Principal components are also generated to be orthogonal to each other in feature space, so it is expected to observe the strong distinction in the correlation directions between the first two deep learning features. This is particularly highlighted in the power-law beta features, which are Fourier-based features. The strong correlation was shown to nearly alternate between features one and two for each power-law beta feature, emphasizing how the deep learning features may be related to the spatial frequency components of the ROI images.

Strong correlation of the deep learning features with radiomic features was also notably reduced after the first two features. This result may also align with expectations, as the total variance of higher principal components tends to fall off rapidly after the first few components, depending on the complexity of the dataset [46].

3.4.3 Limitations and future work

Implications of these results and future studies may influence how patients are designated as high vs. average risk of breast cancer. This will require future studies that better describe the physical extent of the cancer field for each tumor subtype and that quantify the risk associated with the mammographically derived cancer field.

It is important to note key limitations of this work. One primary limitation was the focus only on the features of women with breast cancer. Future work will incorporate these findings into classification models of women with malignant tumors compared to those at low risk. However, this work aimed to characterize features of women with breast cancer in order to gain an understanding of potential signatures of a mammography-based field effect. Since this analysis only investigates features from women with confirmed breast cancer, density values were not controlled for, and the results likely represent an average distribution of breast densities, and this could be investigated further in future studies. The results also did not stratify findings by the molecular subtype of breast cancer present for each woman due to the low number of patients within each subtype. The dataset of 74 total patients included 26 hormone receptor-positive / HER-2 negative tumors, 25 HER-2 positive tumors, and 23 triple negative tumors. However, future analysis may find that feature relationships or presentation of a field effect may be more prevalent for specific molecular subtypes, just as the clinical profile and treatment of each molecular subtype varies.

3.5 Conclusions

The results of this study identified characteristics of a potential mammography-based cancer field effect using human-engineered radiomic and deep leaning-based features from women with breast cancer. Radiomic analysis within mammograms indicated that features in the subcategories of intensity-based, edge frequency, and Fourier-based features were significantly correlated with the parenchymal region in relation to the tumor location. In corresponding images of specimen radiographs showed similar results in intensity-based, edge frequency, and GLCM features. In deep learning features, similar associations were found in both mammograms and specimen radiographs within the first two principal components. Integration of novel data from specimen radiograph radiomic features showed strong relationships of intensity-based features across the parenchymal field in in- and ex-vivo imaging. These results provide potential support for the presence of a cancer field effect that is detectable from imaging studies alone and support the development of computerized analysis of mammographic parenchymal patterns to assess breast cancer risk.

CHAPTER 4: BREAST CANCER RISK ASSESSMENT IN MAMMOGRAPHY USING TEMPORAL NEURAL NETWORKS

4.1 Introduction

Despite variation in the frequency and age of mammogram initiation recommended by medical organizations, breast cancer screening guidelines recommend mammographic imaging at regular intervals during a woman's lifetime [21–23]. Screening exams produce a series of mammographic images over time, which can be compared by radiologists. Past studies have found that a radiologist's comparison of prior mammograms with a current exam may lower the recall rate in screening populations [90,91]. Such practice is important since changes in parenchymal characteristics may indicate the presence of a new suspicious lesion.

In the context of quantifying risk of future breast cancer, utilizing information from prior mammograms may also aid in classification performance of patients at high-risk from average-risk controls. However, most clinically used breast cancer risk models rely on genetic and clinical factors, such as germline mutation or family history of breast cancer [92,93]. Artificial intelligence (AI) based metrics from mammography have shown potential in providing additional value to these models, but most current work in mammography-based breast cancer risk assessment focuses on metrics calculated at the time of mammography screening in combination with clinical factors [28,94]. However, due to the presence of annual screening images, the calculation of AI-based metrics from prior mammograms offers an opportunity to provide imaging-based temporal data and potentially improve performance. Tan et al. investigated the classification performance of features calculated from prior mammograms, [95] but we hypothesize that combining this information through a temporal deep learning approach could provide further improvements to performance [95].

As a recurrent neural network, long-short term memory (LSTM) networks are able to retain information about previous time points in a series and use this information to inform decisions on present time points of that same series [68]. Given the relevance of serial imaging in the diagnostic interpretation of mammographic findings and the importance of incorporating temporal data for the classification of future disease state, we sought to investigate the additive value of multiple sequential antecedent mammograms to the classification of malignant vs. low-risk images. LSTM networks have been used in similar studies to incorporate the multiple acquisition times points within a dynamic imaging protocol; Antropova et al. demonstrated higher classification performance of lesions imaged with dynamic contrast-enhanced MRI using an LSTM-based classifier than using a fine-tuned feed-forward network at a single time point in predicting malignancy of breast lesions [96].

In this research, we aim to develop a model to assess a patient's future risk of developing cancer using AI-based parenchymal characteristics from sequential mammograms collected in the years prior to diagnosis. To appropriately quantify risk, the analysis utilizes only prior images from women who went on to develop cancer, or those who were confirmed to be cancer-free. This research will compare methods of classification using a single time point model and the additive value of multiple time points integrated into a temporal model using an LSTM classifier. This work will also seek to extend the research described in Chapter 3 by distinguishing characteristics of parenchymal patterns of women with biopsy-proven breast cancer from cancer-free controls, considered to have average risk for developing breast cancer.

4.2 Methods

4.2.1 Database

This analysis used a retrospectively collected database of patients who had undergone mammographic imaging at MD Anderson Cancer Center under Health Insurance Portability and Accountability Act (HIPAA)-compliant Institutional Review Board (IRB) protocols. All patients underwent at least three rounds of screening between 2008 and 2018. Note that this cohort is distinct from the cohort used in Chapter 3. Within this cohort, 193 eligible patients were identified with confirmed breast cancer who received treatment for this cancer at MD Anderson Cancer Center. To construct a corresponding cohort of cancer-free controls, patients whose images were classified by a radiologist as having a BIRADS category of 1 or 2 only and who were confirmed breast cancer-free in the two years following their final image were first identified as eligible controls. A matching algorithm was used to select controls for each cancer patient with the following matched characteristics: (1) race, (2) age at the final image within a margin, (3) year of the final image within a margin, and (4) an average time interval between screening exams of 10-14 months. The algorithm initially identified at least two potential matches for 93.3% of the 193 eligible cancer patients using a margin of 1 year and 2 years for patient age and year of final imaging, respectively. To identify potential matches for the remaining patients, the algorithm was applied iteratively while increasing each margin by 1 year. At least two potential matches were identified for all 193 eligible cancer patients within an age margin of 5 years and a final year of imaging margin of 5 years. From the identified potential cancer-free control matches, records were verified to confirm patients were breast cancer-free in the two years following their final image. Images were manually verified to be free of artifacts and all images were acquired with the same system (Hologic). This check included verifying image labels matched presentation, checking for unavoidable calcifications or scars, and the breast size was large enough for placement of a 512x512 pixel ROI. If potential cancer-free control matches were found to have a history of breast cancer or image issues, including calcifications or scars that could not be excluded from the ROI, they were removed from the cohort and the algorithm was re-applied to identify a new match if possible.

Using this sequential matching approach, 183 of the initial 193 eligible cancer patients had at least one cancer-free control match with acceptable image quality. Four patients were eliminated due to lack of a match, while three patients were eliminated due to small breast size/pervasive calcifications and another three due to having no available mammogram corresponding to the clinical record of treatment. Of the total 366 potential cancer-free control matches for the 183 cancer patients, 357 matches were identified (*n* = 9 cancer patients without a second match). A summary of the characteristics from the final selected cancer and corresponding cancer-free controls is shown in Table 4.1.

For the 183 cancer patients included in this study, cancer type and subtype for all patients are summarized in Table 4.2 and Table 4.3, respectively. Patients labeled as undefined or other for breast cancer type, Table 4.2, had synchronous bilateral breast cancer of different types or history of contralateral breast cancer of an undefined or different type than those identified with the most recent cancer. Patients labeled as undefined or other for breast cancer subtype, Table 4.3, had synchronous bilateral breast cancer of different subtypes or subtype classification could not be designated due to lack of biomarker data.

Table 4.1. Summary of the selected cohort of matched cancer and corresponding cancer-free control patients included in analysis. Age at last image is listed as mean (standard deviation). Control patients were matched to cancer patients using an approximate 2:1 ratio using matched (1) race, (2) age at final image within a margin, (3) year of final image within margin, and (4) having an average time interval between screening exams of 10-14 months.

	Cancer	**Control**
Total number of patients	183	357
Total number of images	1951	4626
Average number of exams per patient	5.6	6.6
Age at last image	64.8 (10.5)	64.5 (10.3)
	p = 0.79	
Race		
African American	8.7%	7.8%
Asian	5.5%	6.4%
Hispanic	5.5%	5.6%
Other/Unknown	0.6%	0.3%
White	79.8%	79.8%

Table 4.2. Distribution of breast cancer type from 183 cancer patients included in analysis.

Type	**Count**	**Percent**
DCIS	45	24.6%
Invasive	133	72.7%
Undefined or other	5	2.7%

Table 4.3. Distribution of breast cancer subtype from 183 cancer patients included in analysis.

Subtype	**Count**	**Percent**
ER+	153	83.6%
HER2	11	6.0%
TNBC	15	8.2%
Undefined or other	4	2.2%

For all patients included in the selected cohort (*n* = 540), all images were acquired on Hologic mammography systems (pixel size: 70 μm x 70 μm) and were processed according to the clinical standard at MD Anderson Cancer Center. Only the craniocaudal view was used. Figure 4.1 shows a histogram of the mammographic exams included in the study by months prior to the final or diagnosis image for cancer and cancer-free control patients. In this figure, the large spike at zero months before the final image represents all the available images as diagnosis or the final year of imaging for control patients that was used for matching, but not included in analysis. Along the x-axis from left to right represents going back in time (in months) relative to the final image for each cohort. Thus, the spike at approximately 12 months represents the first prior, the spike at approximately 24 months represents the second prior, and so on. From the figure, it can be seen that the overall number of exams tends to decrease as the months before the final image increases. This may align with expectation, as the further back in time one may search in the medical records, the sparser the information may become on average, as women may have only initiated screening a few years prior or changed institutions at which they attended screening exams. In addition, the spikes at the month markers representing each successive year become slightly more right skewed as the months before the final image increases. This may be a result of women tending to attend their screening exams on or after the date at which it is “due,” with less likelihood to attend the screening exam before one full year has passed . Figure 4.2 shows a histogram of the number of mammographic exams included in the study for cancer and cancer-free control patients. Note that mammographic exams at diagnosis or final time point were not included in the analysis to appropriately quantify risk of future breast cancer.

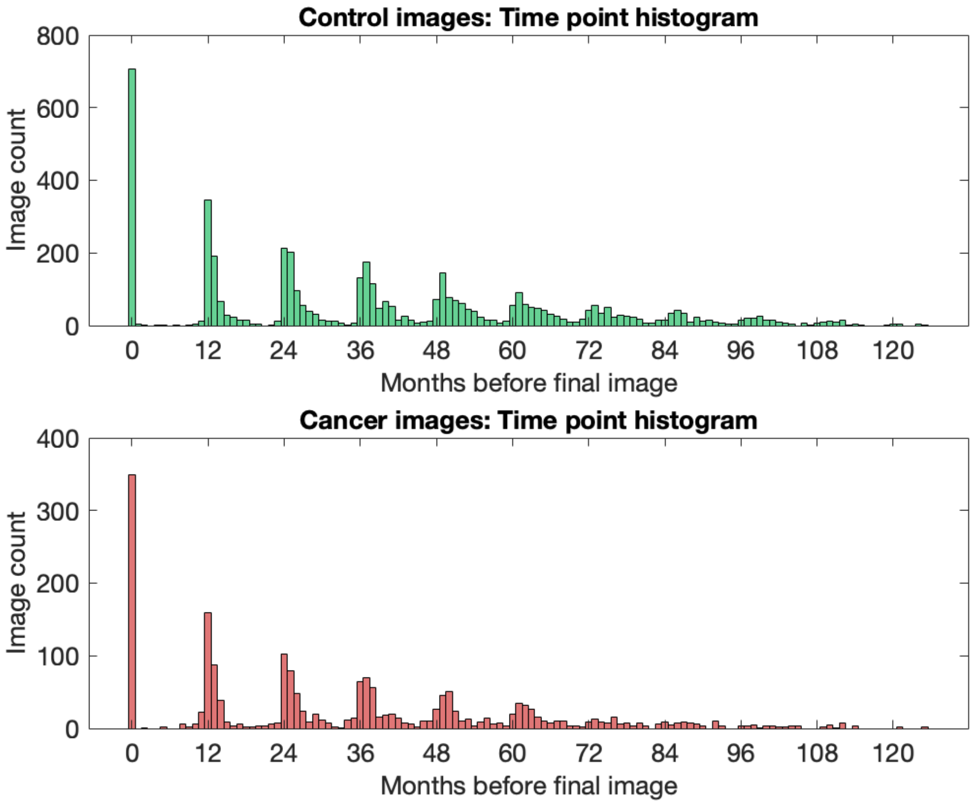


Figure 4.1. Histogram of mammographic exams included in the study by months prior to the final or diagnosis image for cancer and cancer-free control patients. Mammographic exams at diagnosis or final time point (months = 0) were not included in analysis to appropriately quantify risk of future breast cancer.

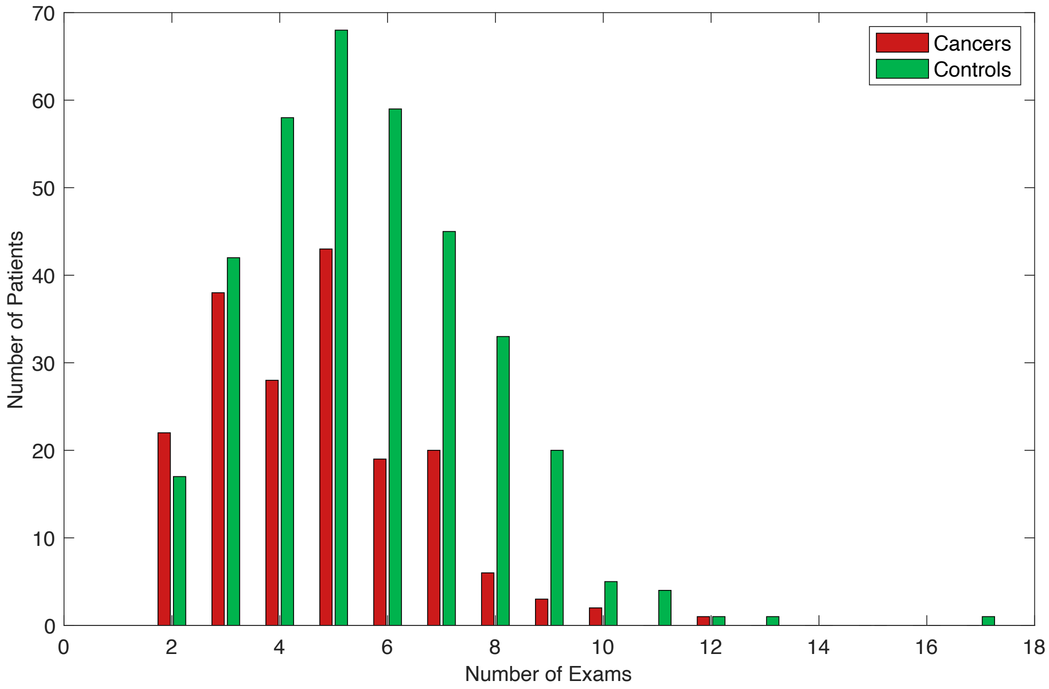


Figure 4.2. Histogram of the number of mammographic exams included in the study for cancer and cancer-free control patients. Mammographic exams at diagnosis or final time point were not included in analysis to appropriately quantify risk of future breast cancer.

4.2.2 Feature Extraction

Regions of interest (ROIs) of 512x512 pixels were manually selected from the central breast region posterior to the nipple on both lateralities of the craniocaudal mammogram, as shown in Figure 4.3. This ROI location has been found in previous studies to be robust to small variations in user placement and have improved classification performance when compared to other regions of the breast [89,97]. An in-house AI workstation was used to automatically extract 45 radiomic texture features, as previously listed in Table 3.1, describing tissue contrast/intensity and structure in each breast region. Feature formulas and descriptions can be found in the literature [55–57]. A transfer learning approach, similar to that described in Chapter 3, was used for deep learning-based features. A VGG19 convolutional neural network (CNN) architecture was first pre-trained on ImageNet [34]. The pre-trained network was then used with the mammogram ROI as the input, and 1472 generic deep learning features were extracted from each of the five max pooling layers. CNN feature extraction and network training were performed in Keras using a TensorFlow framework [98,99].

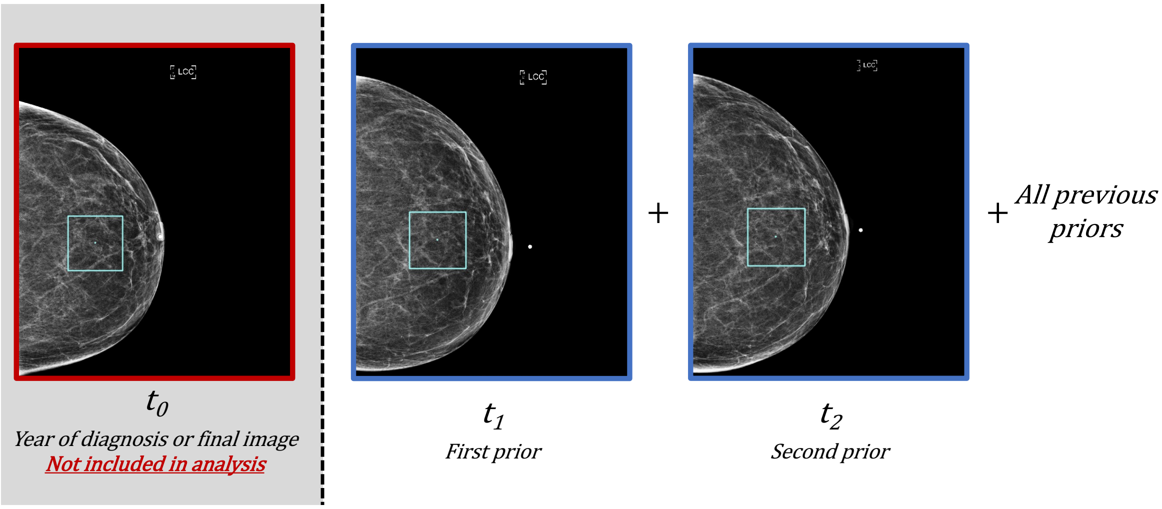


Figure 4.3. Example 512x512 pixel region of interest (ROI) placement on mammogram and priors with notation for the time point for each mammogram. Diagnosis or final (t_0) mammograms were not included in analysis.

4.2.3 Single time-point classification using support vector machine

To assess a patient's future risk of developing cancer using radiomics and deep learning-based parenchymal characteristics from a single time point, classification was performed on mammograms collected in the year prior to diagnosis (t_1). Features extracted from the first prior were used as the input to a support vector machine (SVM) classifier using five-fold cross validation [100]. A summary of the analysis pipeline is shown in Figure 4.4. To reduce the dimensionality of deep learning-based features input to the classifier, principal component analysis (PCA) was used to reduce the features to the first 20 principal components.

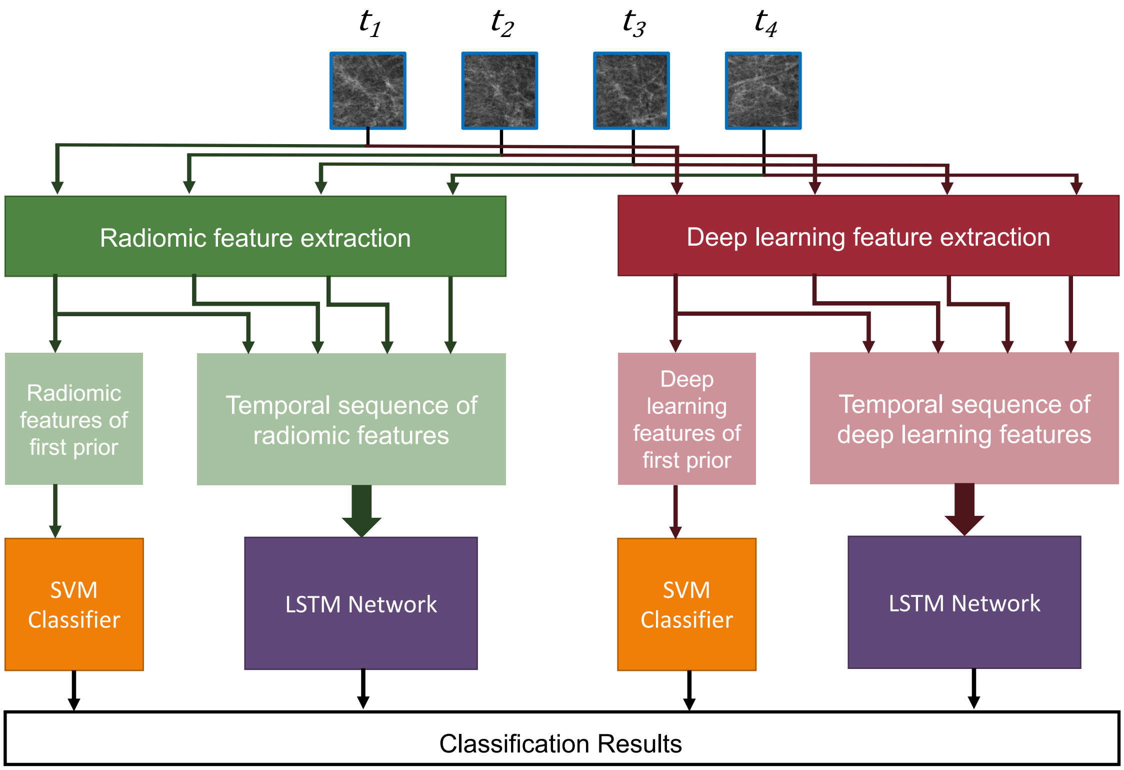

t_1
t_2
t_3
t_4
Radiomic feature extraction
Deep learning feature extraction
Radiomic features of first prior
Temporal sequence of radiomic features
Deep learning features of first prior
Temporal sequence of deep learning features
SVM Classifier
LSTM Network
SVM Classifier
LSTM Network
Classification Results

Figure 4.4. Summary of analysis methods used to classify patients who will be diagnosed with cancer from cancer-free controls using a single time point, representing the first prior (t_1) and an SVM classifier and a temporal sequence of time points (t_1, t_2, t_3, etc.) and an LSTM classifier.

4.2.4 Temporal classification using long-short term memory networks

Temporal classification was performed on mammograms collected in the years prior to diagnosis (t_1, t_2, t_3, etc.) to assess a patient's future risk of developing cancer using parenchymal characteristics from multiple time points. Features extracted from each image were organized into sequences, where each sequence is *n* time points long, representing *n* imaging exams for each patient. As such, features from one patient were only contained in a single fold for k-fold cross-validation. Analysis was completed with both a maximum sequence length of four time points and nine time points to compare the impact of sequence length on performance. The sequence length of four time points was selected as almost all patients had at least four time points, so this classifier would be trained with a relatively "full" dataset, and the sequence length of nine time points was selected to be inclusive of all available imaging exams. For evaluation with a maximum sequence length of four time points, imaging studies collected at least 10 months prior to diagnosis and no more than 58 months prior to diagnosis were included. These limits were selected as they represent approximate relative minima between sequential years, as shown in Figure 4.1. For evaluation with a maximum sequence length of nine time points, imaging studies collected 10 to 150 months prior to diagnosis were included to evaluate all possible images from prior years. Since patients had varying numbers of time points, sequences were post-padded with zeros to the maximum length of a sequence specified. Post-padding has been shown to be an effective method of sequence length standardization compared to other methods and was found in preliminary analysis to be the most effective method for this dataset when compared to pre-padding and padding with feature averages between time points with gaps greater than 18 months [101]. Sequences were used as the input to a long-short term memory (LSTM) network classifier using five-fold cross validation [68].

The LSTM network was trained using a standard stochastic gradient descent optimizer, and hyperparameters were determined through a limited sweep of learning rate, batch size, hidden dimensions, and epochs values [102]. Selected parameters included a learning rate of 10^{-3}, a batch size of 32, 512 hidden dimensions, and 25 epochs.

4.2.5 Statistical analysis

The performance of classifiers was assessed using receiver operating characteristic (ROC) analysis using the area under the ROC curve (AUC) as the figure of merit in the task of predicting cancer in the following year's mammogram. AUC values were compared using a 2-sample z-test, and p-values were corrected for multiple comparisons using the Holm-Bonferroni method [85,103]. A significance threshold of 0.05 was used and 27 total comparisons were corrected for.

4.3 Results

4.3.1 Receiver operating characteristic analysis performance

ROC analysis results for the task of predicting the future occurrence of breast cancer from the three classifiers are shown in Figure 4.5 and Table 4.4. These results show (1) the single time point SVM classifier using only the first prior, (2) the temporal sequence LSTM classifier using four prior time points, and (3) the temporal sequence LSTM classifier using nine prior time points. All AUC values were statistically compared to the baseline value of 0.5, which indicates performance equivalent to random guessing, and corrected for the 27 multiple comparisons using the Holm-Bonferroni method using a threshold of significance of 0.05 [103]. For the single time point SVM classifier, only the AUC of the merged feature classifier from the tumor side indicated statistical difference from 0.5 after multiple comparisons correction. All AUC values from the temporal sequence LSTM classifiers using four and nine prior time points were found to be statistically better than random guessing after multiple comparisons correction. As such, the temporal sequence LSTM classifiers generally outperformed the single time point SVM classifier on all feature types and image lateralities.

These results also show that the performance of from the temporal sequence LSTM classifiers did not vary much between the use of four maximum time points and nine maximum time points. The nine time point classifier slightly outperformed the four time point classifier when using radiomic features, the four time point classifier slightly outperformed the nine time point classifier when using deep learning features, and both classifiers performed similarly when using merged radiomic and deep learning features. Given that the highest overall performance resulted from the four time point classifier, it is the primary temporal sequence classifier discussed for the remainder of the results.

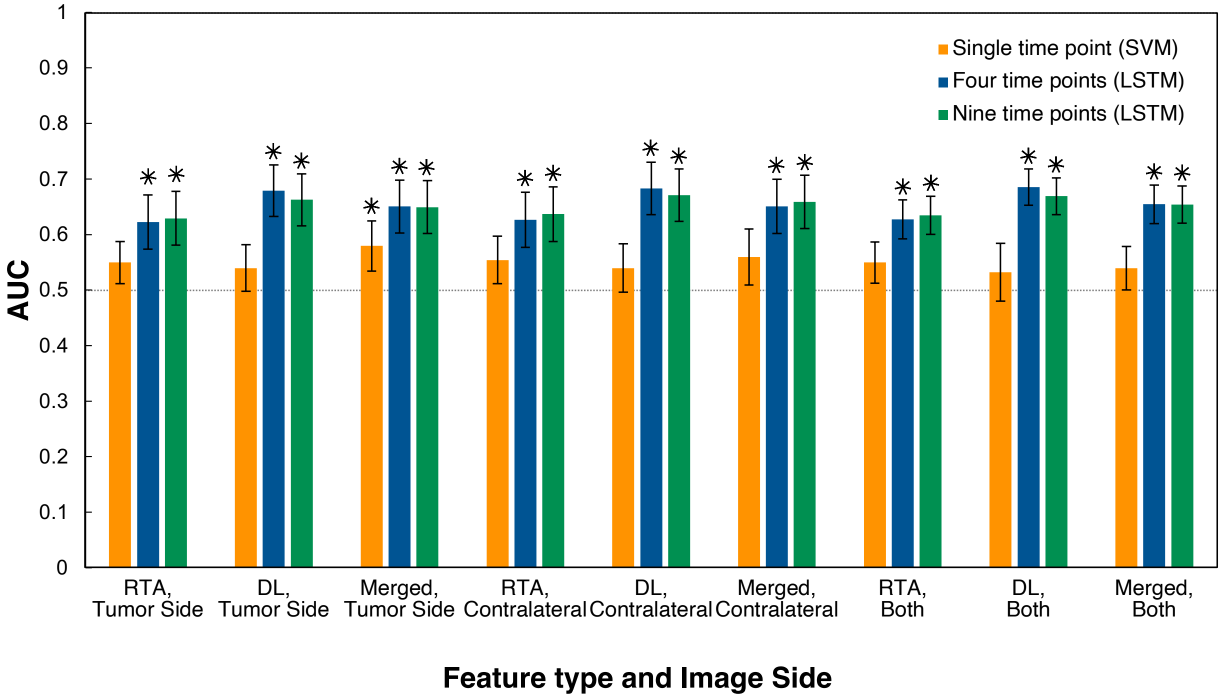


Figure 4.5. ROC analysis results for the single time point (SVM) classifier using only the first prior (t_1) and the temporal sequence (LSTM) classifiers using four and nine prior time points (t_1, t_2, t_3, etc.). Results are separated by feature type, radiomic (RTA) features and deep learning (DL) features, and the image side. An image side label of "Both" indicates the tumor side and contralateral predictions were combined into a single classifier. Error bars indicate one standard error, and the dashed line at AUC = 0.5 indicates performance of guessing.

Table 4.4. Summary of ROC analysis results for single time point (SVM) classifier using only the first prior (t_1) and temporal sequence of timepoints (LSTM) classifier using four and nine prior time points (t_1, t_2, t_3, etc.). P-values indicate comparison to random guessing (AUC = 0.5). Values considered statistically significant after Holm-Bonferroni multiple comparisons correction are shown in italics.

	Single time point classifier (SVM)	**4-time point sequence classifier (LSTM)**	**9-time point sequence classifier (LSTM)**
	AUC [95% CI]	AUC [95% CI]	AUC [95% CI]
Radiomics, tumor side	0.55 [0.51,0.59] $p = 0.015$	*0.62 [0.57,0.67]* $p < 0.001$	*0.63 [0.58,0.68]* $p < 0.001$
Radiomics, contralateral	0.55 [0.51,0.60] $p = 0.018$	*0.63 [0.58,0.68]* $p < 0.001$	*0.64 [0.59,0.69]* $p < 0.001$
Radiomics, both lateralities	0.55 [0.51,0.59] $p = 0.013$	*0.63 [0.59,0.66]* $p < 0.001$	*0.64 [0.60,0.67]* $p < 0.001$
Deep learning features, tumor side	0.54 [0.50,0.58] $p = 0.071$	*0.68 [0.63,0.73]* $p < 0.001$	*0.66 [0.62,0.71]* $p < 0.001$
Deep learning features, contralateral	0.54 [0.50,0.58] $p = 0.079$	*0.68 [0.64,0.72]* $p < 0.001$	*0.67 [0.62,0.72]* $p < 0.001$
Deep learning features, both lateralities	0.53 [0.48,0.58] $p = 0.191$	*0.69 [0.65,0.72]* $p < 0.001$	*0.67 [0.64,0.70]* $p < 0.001$
Merged features, tumor side	*0.58 [0.53,0.63]* $p < 0.001$	*0.65 [0.60, 0.70]* $p < 0.001$	*0.65 [0.60, 0.70]* $p < 0.001$
Merged features, contralateral	0.56 [0.51,0.61] $p = 0.027$	*0.65 [0.60, 0.70]* $p < 0.001$	*0.66 [0.61, 0.71]* $p < 0.001$
Merged features, both lateralities	0.54 [0.50,0.58] $p = 0.054$	*0.66 [0.62,0.69]* $p < 0.001$	*0.65 [0.62,0.69]* $p < 0.001$

ROC curves for the single time point SVM classifier using only the first prior and the temporal sequence LSTM classifier using four prior time points for radiomic, deep learning, and merged deep learning and radiomic features are shown in Figure 4.6, Figure 4.7, and Figure 4.8, respectively.

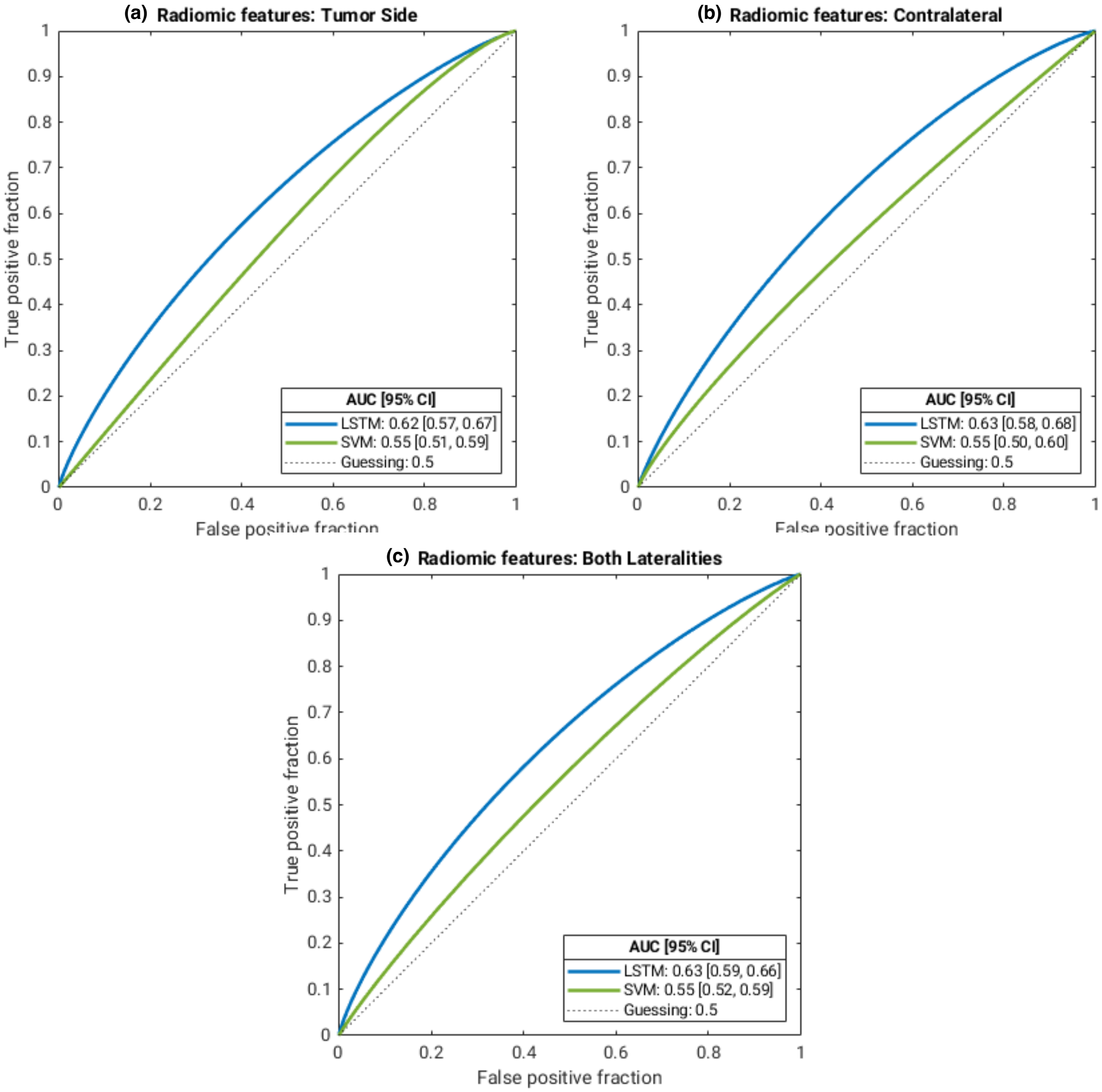


Figure 4.6. ROC curves for radiomic features, comparing performance between single time point (SVM) and four time points (LSTM) for tumor side (a), contralateral (b), and both lateralities (c). The dashed diagonal line indicates performance of guessing.

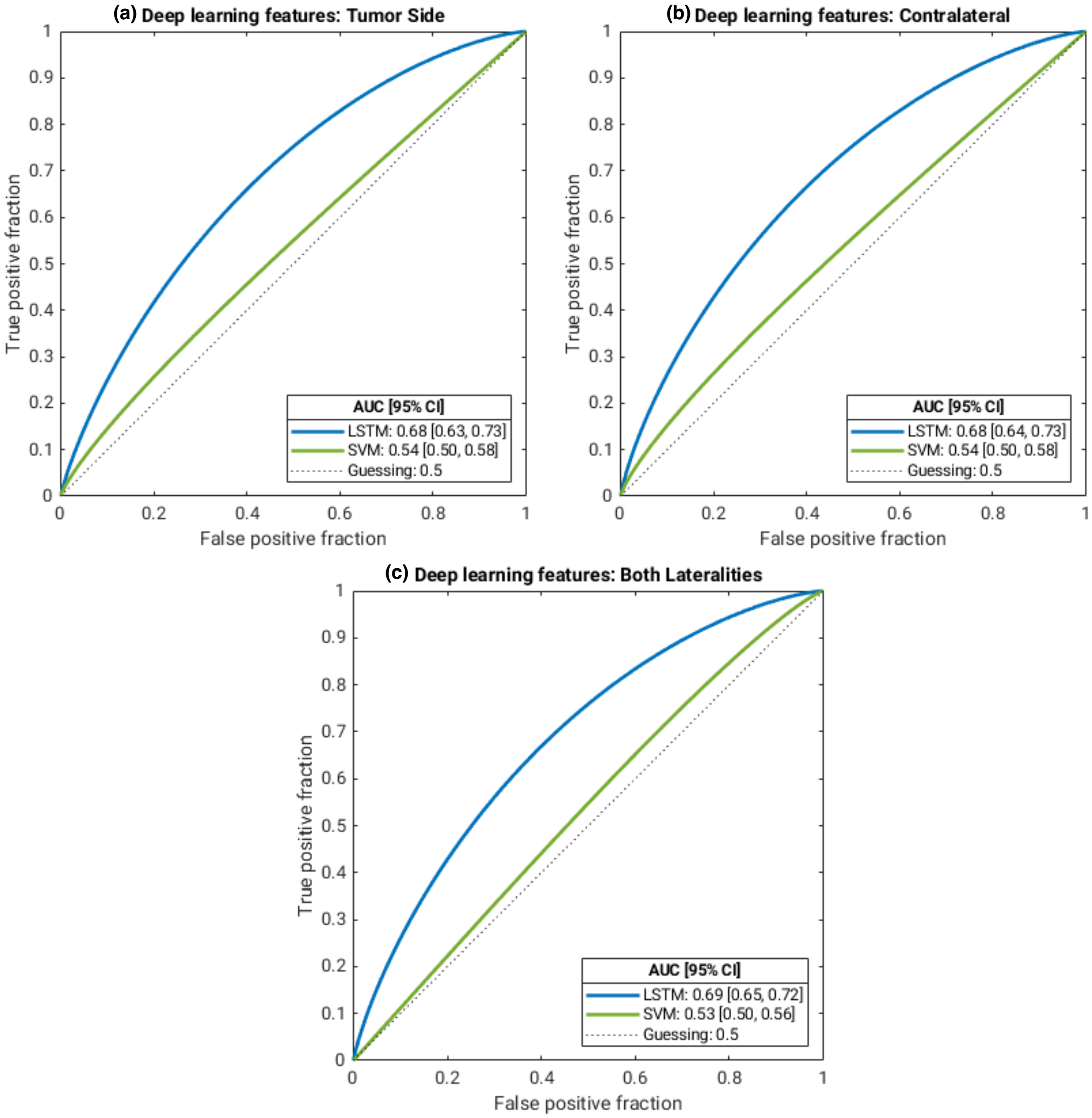


Figure 4.7. ROC curves for deep learning features, comparing performance between single time point (SVM) and four time points (LSTM) for tumor side (a), contralateral (b), and both lateralities (c). The dashed diagonal line indicates performance of guessing.

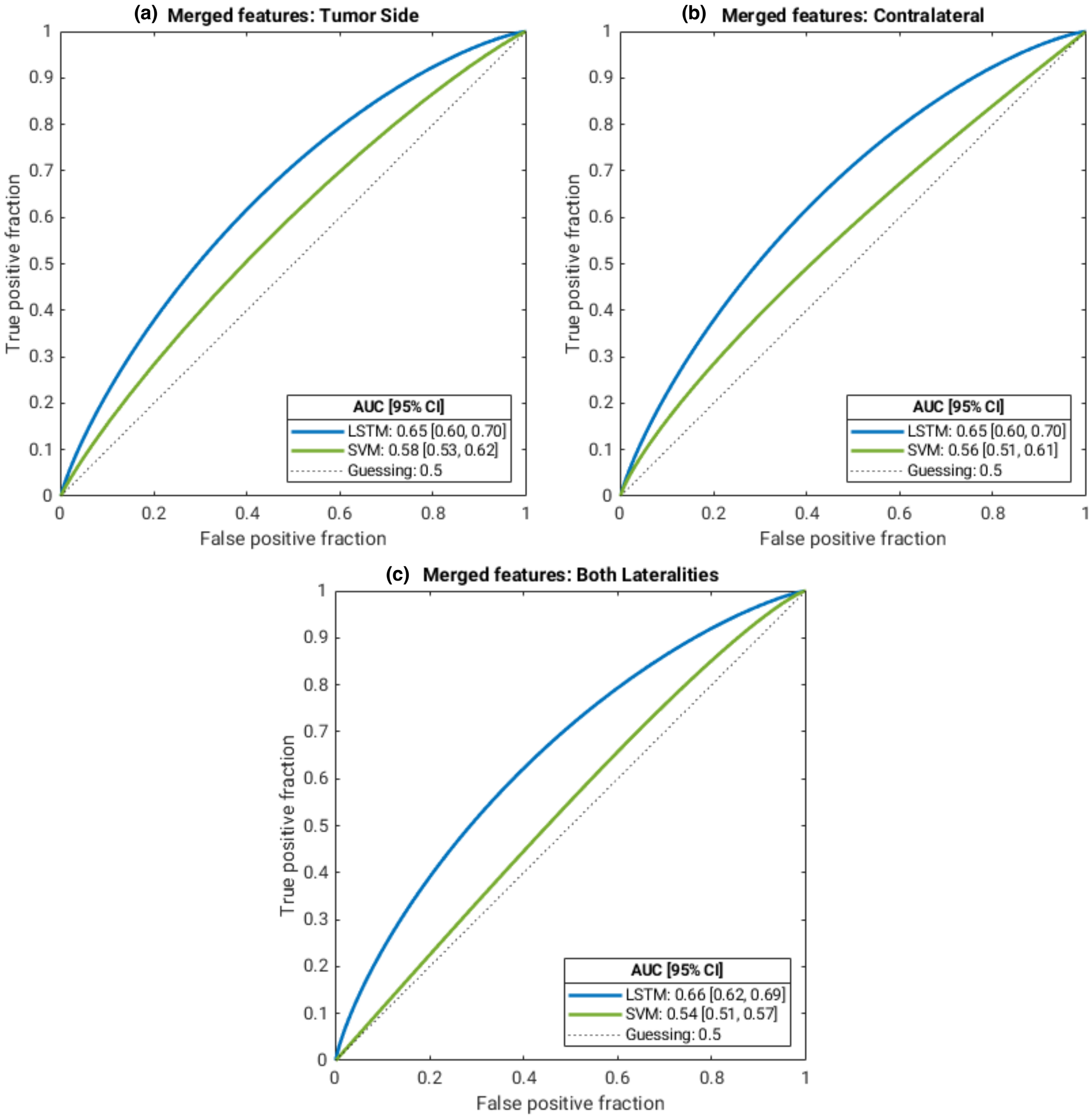


Figure 4.8. ROC curves for merged radiomic and deep learning features, comparing performance between single time point (SVM) and four time points (LSTM) for tumor side (a), contralateral (b), and both lateralities (c). The dashed diagonal line indicates performance of guessing.

4.3.2 LSTM network predictions

To gain a better understanding of the distributions of the LSTM classifier's predictions with respect to the patient subgroups, LSTM prediction scores were plotted as a function of the features from which they were generated in Figure 4.9. This figure shows a slight separation between clusters of cancer patients and cancer-free controls from both plots. The cancer patients tended to have higher prediction scores, plotting very slightly towards the upper right of the cluster. This distribution was further evaluated with respect to the cancer patients' future cancer type and subtype, as shown in Figure 4.10 and Figure 4.11, respectively. However, from these figures, no apparent clustering of cancer patients of a specific type or subtype could be visually identified, indicating no strong correlation of a select cancer subgroup that helped bolster or deteriorate classifier performance.

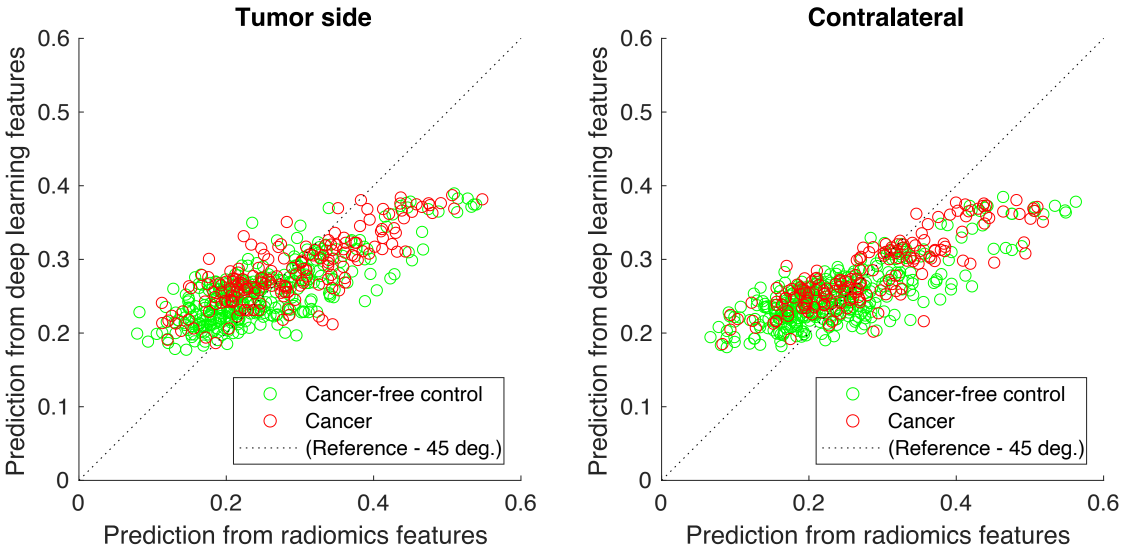


Figure 4.9. LSTM prediction scores from radiomic features and deep learning features for all patients analyzed. Slight clustering of the cancer patients towards the upper right away from the primary cancer-free control cluster indicates differences between the two cohorts.

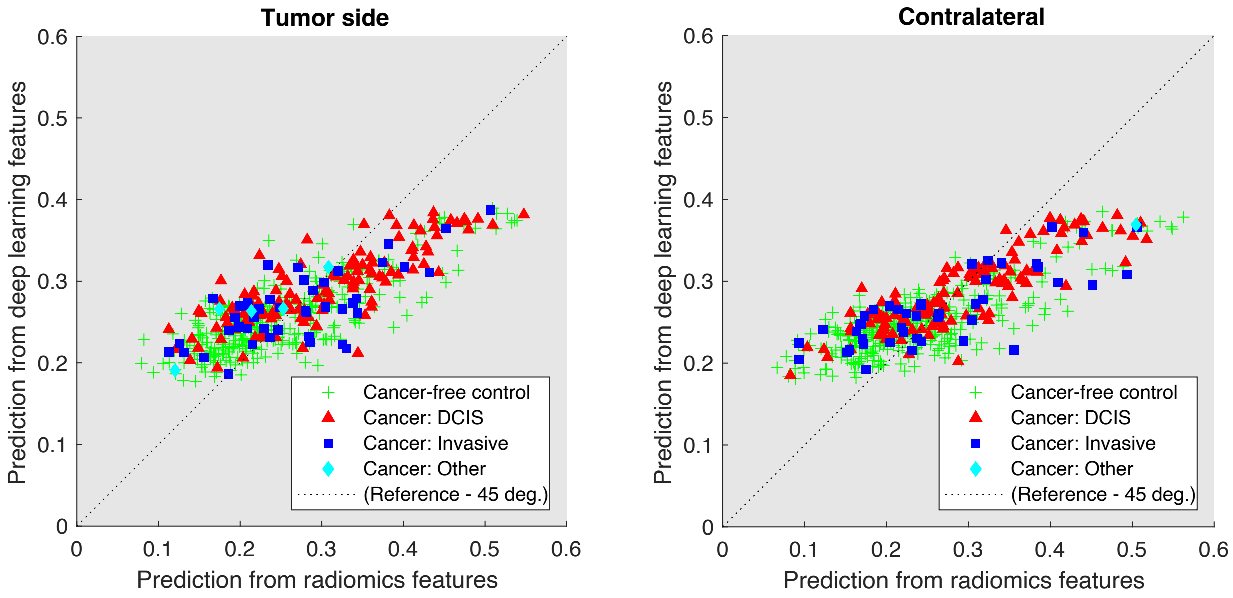


Figure 4.10. LSTM prediction scores from radiomic features and deep learning features with all cancer patients labeled by cancer type. No distinct clustering of a cancer type (DCIS vs. invasive) can be noted from their distributions.

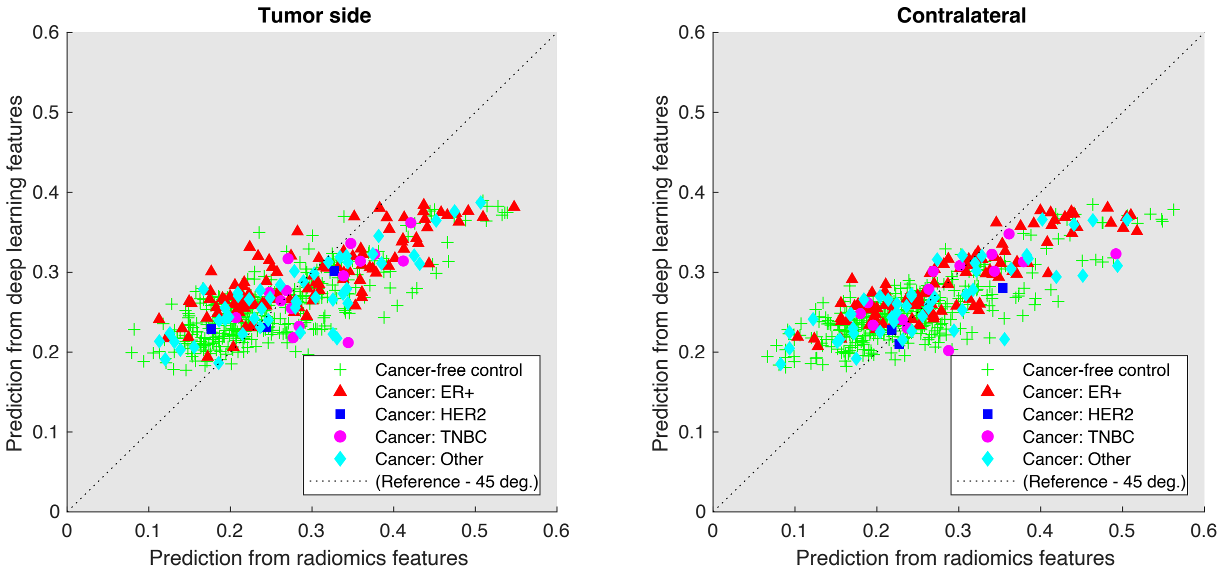


Figure 4.11. LSTM prediction scores from radiomic features and deep learning features with all cancer patients labeled by cancer subtype. No distinct clustering of a cancer subtype can be noted from their distributions.

4.4 Discussion

For the task of assessing a patient's future risk of developing cancer using AI-based parenchymal characteristics from mammograms, a single time point SVM classifier using only the first prior image and temporal sequence LSTM classifiers using multiple prior images were investigated. Results failed to show statistical differences of the single time point SVM classifier using only the first prior image from random guessing. However, results indicated statistical differences of the temporal sequence LSTM classifiers performance from guessing. While the performance of this classifier was modest, results show a small but significant ability to classify women who were diagnosed with cancer in the following year versus those with no evidence of cancer. The temporal LSTM classifier performance is also comparable to other risk models based on prior imaging, as found in the literature [104,105].

An important component of this study, in contrast to other investigations of breast cancer risk in the literature, is the emphasis on constructing and verifying cancer and cancer-free control cohort that were matched on the demographic factors of age, race, and year of imaging. Furthermore, this study utilized manual methods of image review and ROI selection, ensuring that data included in the study did not contain any erroneous markers that could influence classifier performance. Increased age at imaging has been cited to be significantly associated with breast cancer diagnosis and should be adjusted for or controlled within the dataset, as performed in this study [31,105]. Although race is not typically a demographic factor controlled for in breast cancer risk studies, recent work by Gichoya et al. indicates the ability of AI deep learning models to predict self-reported race as a potential bias learned by AI models for other tasks, such as predicting breast cancer [106]. Future work should also investigate integrating clinical factors into the classification model to improve performance.

An interesting result from this study included the similarity of classifier performance in predicting future unilateral cancer from both the tumor side and contralateral breast. Across all classifiers and feature types investigated, AUC values were not systematically higher for tumor side or contralateral images. In the context of predicting future cancer, this is an important finding, as one will not know in which side cancer will develop. However, this finding also supports the hypothesis of a broad cancer field effect, extending across both the tumor side and contralateral breast. While the tumor side breast may present changes in texture in the immediate tumor area, results of this study support that changes are also present in the contralateral breast that may be indicative of future cancer.

In addition, similar AUC values were found between LSTM classifiers on this dataset using a maximum of four or nine time points across all feature types and both lateralities. One potential reason for this may be that as the months prior to the final image increased, as shown in Figure 4.1, the overall number of patients with additional exams decreased. As such, the LSTM classifier with a maximum of nine time points contained sequences with more zero-padding than is present in sequences truncated at four time points, as most patients in the dataset had at least three prior exams. This zero padding would not be expected to degrade performance but would produce relatively less populated time points available to train the model. Contribution from images at very early time points also may not have added substantially to the model. Future work should seek to comprehensively evaluate classification performance with fewer and additional time points to understand how much data is needed to reach satisfactory performance.

LSTM network prediction distributions did not indicate clustering of predictions from radiomics or deep learning features for any specified cancer type or subtype investigated. However, for the variable of cancer subtype, it is important to note that the database contained a higher prevalence of ER+ cancers (83.6%, Table 4.3) than is typical (60-70%) [107]. As a result, other cancer subtypes were underrepresented in this database and may not have had the statistical power necessary to demonstrate variation within the current model.

A limitation of this study is that the database used in this investigation only included a single institution and single mammography system (Hologic). Future work should seek to investigate these findings in patients from multiple institutions and imaging systems.

4.5 Conclusions

This research focused on the development of an AI model to evaluate a patient's future risk of developing cancer, based on parenchymal characteristics from sequential mammograms collected prior to diagnosis. Our analysis only utilized radiomic and deep learning features from pre-cancerous and control images to accurately quantify risk. Results failed to show statistical difference of a single time point SVM classifier, using only the first prior image, from random guessing but indicated the performance of temporal sequence LSTM classifiers to be statistically better than guessing and comparable to other risk-based models in the literature. Further, the similarity of classifier performance across both lateralities may support the hypothesis of a cancer field effect, extending the work described in Chapter 3.

CHAPTER 5: STATISTICAL METHODS FOR DEMOGRAPHIC BALANCING AND POPULATION SAMPLING IN A MULTI-INSTITUTIONAL DATA COMMONS

5.1 Introduction

The recent explosion of AI development has led to an increased emphasis on the need for algorithm performance validation and generalizability. While AI methods have shown great utility for many tasks, the availability and quality of data used to develop models have been shown to affect their generalizability and robustness significantly. Reliance on small, single-institution datasets may produce performance estimates that will not generalize to other populations. However, obtaining large, well-curated datasets in medical imaging is challenging due to patient privacy regulations. Furthermore, verification of algorithm performance presents an ethical burden to regulators and researchers. The medical imaging AI applications of COVID-19 highlight the impact of this challenge on developing clinically relevant AI systems, as the pandemic has presented an urgent and critical public health crisis with many essential biomedical research and development needs to address.

Since early 2020, the COVID-19 pandemic has stimulated a rapid research effort in AI model development for COVID-19 applications [108]. Many AI tools for medical imaging of the COVID patient were developed, including early detection and differential diagnosis of COVID-19, prognosis and assessment of response to treatment, and monitoring of the post-COVID patient [37–40,71,109,110]. Many studies have shown promising performance of AI models for various applications. However, the potential impact of algorithm bias and lack of clinical utility have been noted as major shortcomings in AI for COVID-19 medical imaging [36,77]. In response to this urgent public health need, the Medical Imaging and Data Resource Center (MIDRC; MIDRC.org) was established in August of 2020 to accelerate medical imaging machine intelligence research for COVID-19.

MIDRC is a multi-institutional research collaboration between the American College of Radiology (ACR®), the Radiological Society of North America (RSNA), and the American Association of Physicists in Medicine (AAPM) created to address critical gaps in resources and technology for AI in medical imaging. Through the work of five technology development projects and twelve collaborating research projects, MIDRC is providing processes for data intake, de-identification, quality assessment, and distributed public access in addition to organizing research challenges and curated datasets to support high-quality research methods. The aim of MIDRC is to accelerate machine intelligence research for COVID-19 and eventually for other diseases that utilize medical imaging in, for example detection, diagnosis, or prognosis. One primary component of MIDRC is the development of a publicly available image repository (data.MIDRC.org), as well as a sequestered database for performance evaluation and benchmarking of algorithms.

While the majority of the de-identified data (both images and metadata) submitted to MIDRC are open, i.e., accessible to the public, approximately 20% of the data are being sequestered from public use for the purpose of machine intelligence algorithm evaluation. These sequestered data will act as a large base from which task-based samples, or “test-sets,” can be drawn to provide an estimate of an algorithm’s performance or generalization ability, without ever releasing the data publicly or giving users direct access to the cases used for testing, thus maintaining the integrity of the test set. While MIDRC aims to provide a platform with a wide array of diverse data representative of the American population, gaps and biases may inadvertently arise. Purposeful selection of patients for the sequestered database will be a useful tool for the future assessment of the bias of the database itself, as well as the algorithms developed from it. Potential bias and lack of generalizability in AI algorithms have been key shortcomings of the clinical utility of AI, and multiple studies have found demographic differences to have a profound impact on the performance of medical image classifiers [106,111]. To ensure both datasets are similarly representative of the population, we developed a methodology, described in Section 5.2.1, to balance demographic characteristics, or variables, across the sequestered and public data. This process is implemented for incoming batches of data on an ongoing basis.

Balancing multiple variables among subgroups is a widely studied topic in the field of clinical trial development [112,113]. However, similar approaches have rarely been applied in the field of machine intelligence due to the use of typical train-test splitting in datasets that are often small or moderate in size. In both clinical trials and image analysis studies , there are similar patient variables, but imaging contributes additional complexity including imaging machine type, protocol, etc., as the data collection process is more varied. Furthermore, the process must balance variables within each batch of incoming imaging data on an ongoing basis.

To utilize a data commons for algorithm development or performance assessment tasks, each user will need to select a sample that is specific to their task and target population (e.g., all COVID-19 positive images from CT, or 50:50 COVID-19 positive and negative in chest radiographs, with a demographic profile matching that of the U.S. Census). Matching data from a public repository to a specified demographic group is a conventionally arduous task that may require hours of manual modification to datasheets, which are also frequently subject to human error. Section 5.2.2 will outline the process developed for large-scale sampling of the MIDRC data commons to create task-based samples matched to a specified target demographic population.

In this research, we describe the developed sequestration method of multi-dimensional stratified sampling to separate incoming data batches into “open” and “sequestered” commons, evaluate the performance of this sequestration method in terms of similarities between the two commons, and describe a developed method of task-based distribution sampling. These methods were all developed for direct application and implementation in MIDRC; however, they are intended to be generalizable to other databases and fields of research.

5.2 Methods

5.2.1 Sequestration by stratified sampling

To describe the developed method of sequestration, we will first begin by describing the associated data pipeline for receiving batches of data to sequester. An overview of this pipeline is shown in Figure 5.1. Before sequestration, de-identified clinical data of patients are submitted to MIDRC through data input portals hosted by ACR and RSNA. The quality of submitted medical imaging studies is assessed, and the associated metadata are harmonized for representation within the MIDRC data model at data.MIDRC.org. Subsets of the incoming data then undergo separation and are designated as “open” or “sequestered” on an ongoing basis in batches created at regular submission time intervals.

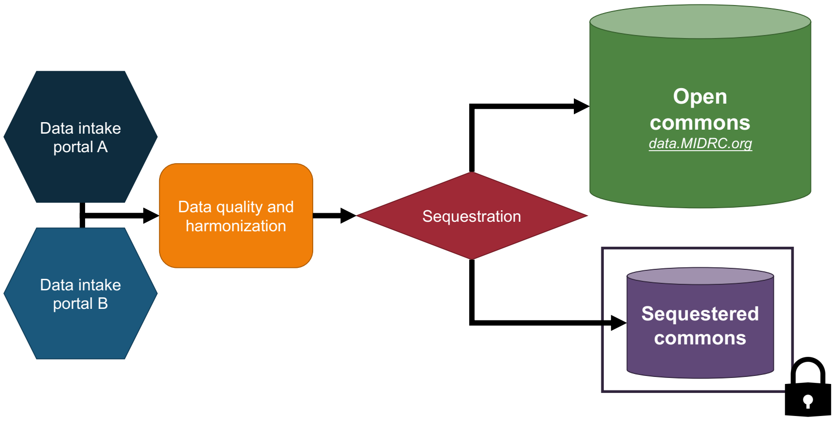


Figure 5.1. Simplified pipeline of data intake and sequestration in MIDRC.

To sequester approximately 20% of an incoming data batch, first, de-identified patient IDs are compared across all previously processed batches. If a patient ID already exists in either the public or sequestered commons, the incoming data for this patient are placed in the relevant commons. This process ensures data are placed into the open or sequestered commons at the patient level, and all images from longitudinal studies of a given patient are contained in only one of the two commons. Following this longitudinal data check, the data of remaining patients in the intake batch are sequentially separated into multiple strata based upon the anonymized clinical site ID, image modality, COVID status (whether a patient ever tested positive for COVID), and reported patient race, age, sex at birth, and ethnicity. A diagram of the sequestration process is shown in Figure 5.2. Within each resulting bin or strata, i.e., a group of patients with a particular combination of characteristics, the patients are randomly assigned to the open dataset or the sequestered dataset with proportions of approximately 80% and 20%, respectively. Thus, for *n* variables of interest, the balance of the *n*-dimensional distribution of variable combinations can be controlled.

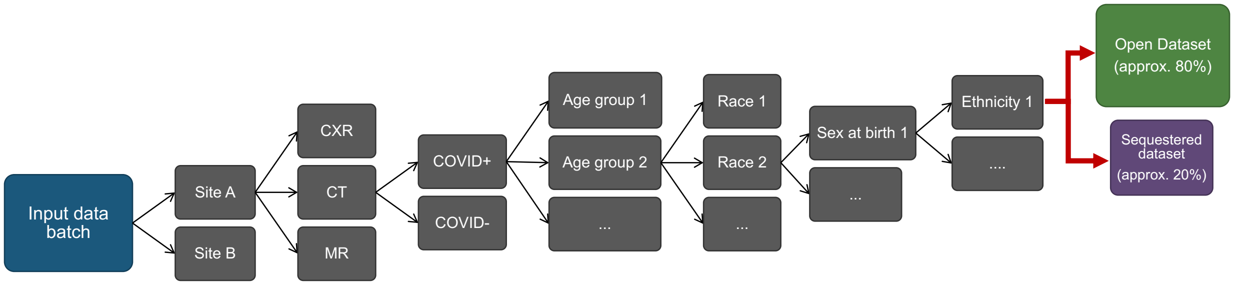


Figure 5.2. Diagram of demographic factors used to stratify data into sequestered and open databases. The input data batch is sequentially split into all possible variations of each category until an individual stratum, containing a unique combination of variables, is achieved. This individual stratum is then randomly separated into the open and sequestered commons with proportions of approximately 80:20.

Since, for any given patient, imaging studies from multiple modalities might be available, such as computed tomography (CT) or radiographs, the sequestering of patient data by modality was accomplished by first identifying the most prevalent image modalities in an intake data batch. If a patient entry contained more than one modality, the modality most prevalent in the intake data batch was established to be that patient's "primary" modality, and images from any less prevalent modality were assigned along with the most prevalent modality. Modalities present in this analysis included computed radiograph (CR), CT, digital radiography (DX), and magnetic resonance (MR) images. Patient age data were grouped into categories matching the age group categories provided by the Center for Disease Control COVID-19 database [114]. Patient sex at birth, race, and ethnicity were grouped in agreement with the categories defined by the NIH [115,116].

Demonstrating sequestration algorithm use in an example database

To demonstrate use of the previously described sequestration algorithm and variability of the produced results, we applied the algorithm to an example input dataset for 2000 independent trials. In each trial, a different random seed was used to initiate the splitting. 5000 patients were randomly selected from the public data commons (data.MIDRC.org) to serve as the example of an input dataset. This number of patients approximates a typical data submission from a single clinical site to MIDRC from 2021-2022. As such, the example input dataset was not separated by contributing clinical site. The developed sequestration algorithm was applied to the remaining demographic variables (age, race, sex at birth, ethnicity, COVID-19 status, and image modality) to achieve a similar distribution of variables between the input dataset and the two subsets, i.e., the open and the sequestered data commons. The variables of age, race, sex at birth, ethnicity, COVID-19 status, and image modality contained 9, 7, 4, 3, 3, and 4 categories respectively, resulting in a total of 9,072 strata. The mean and standard deviation of each demographic category's prevalence were calculated over all independent trials and compared to the prevalence in the input dataset and between the two subsets.

Comparison of sequestration algorithm to naïve random sampling

To extend this evaluation of the sequestration algorithm, the performance of the sequestration algorithm was compared to “naïve” separation of the dataset with an overall 80:20 random drawing. In this ‘naïve’ separation the assignment to either the open or sequestered dataset is made randomly without considering demographic variables. This was applied to the input dataset for 2000 independent trials and the resulting distribution of cases across demographic categories were compared. Balance of demographic distributions were first compared by creating histograms of the scaled difference from expectation in each category, calculated according to Equation (5.1).

$$Scaled\ difference\ from\ expectation = \frac{|(f)N_T - N_{Open}|}{(f)N_T} \tag{5.1}$$

Here, f represents the fraction placed in the open commons (here equal to 0.8), and N_T represents the total number of patients in a given category from the input dataset. Similarly, N_{Open} represents the number of patients in a given category in the open dataset, which would be expected to be 80% of N_T in size if the dataset split was exact. The value of this metric from all trials was plotted in histograms for a given bin within a category, e.g., in the race categories, all Asian patients. Distributions from our developed stratified sampling algorithm and from naïve random sampling were compared using the one-tailed Mann-Whitney U test, with an alternative hypothesis indicating the median of the distribution of the scaled difference from expectation, Equation (5.1), for stratified sampling was less than that for the naïve random sampling. P-values less than 0.05 were considered significant. Results were adjusted for multiple comparisons using the Holm-Bonferroni multiple comparisons correction [103]. In this correction, the calculated p-value must be less than the chosen significance threshold divided by the rank of said p-value, after ordering from least to greatest where the smallest p-value would have a rank of the total number of comparisons. In this analysis, 29 comparisons were evaluated.

Evaluation of sequestration methodology through joint distribution balance

While the balance of demographic variables was evaluated individually to verify the developed sequestration methodology in individual batches, verification of joint distribution balance over all data must also be confirmed. To ensure the developed method works as intended in a larger cohort over multiple dimensions, we also investigated the balance in the joint distributions of patient demographic characteristics between the open and sequestered commons.

This study included 54,185 patients whose de-identified imaging studies and metadata had been submitted to and curated by MIDRC as of August 31st, 2022, including patient data ingested, but not yet published on data.midrc.org. Imaging studies were analytically separated by patient into open and sequestered commons using the previously described sequestration method, aiming to jointly balance distributions of demographic characteristics. Variables stratified over included acquisition site, imaging modality, COVID-19 status, age, race, sex assigned at birth, and ethnicity, as outlined in Figure 5.2. Patient age data were grouped into categories matching the age group categories provided by the Center for Disease Control COVID-19 database [114]. Patient sex at birth, race, and ethnicity were grouped in agreement with the categories defined by the NIH [115,116]. The resulting public and sequestered commons included 41,556 (77%) and 12,629 (23%) patients, respectively.

To compare the balance of the joint demographic distributions between the commons, patient characteristics were re-separated into bins representing unique combinations of the selected variables. To limit the dimensionality of our analysis, we selected four of the seven demographic variables: COVID-19 status, age, race, and sex at assigned birth. Patients with COVID-19 status, age, or sex at assigned birth fields that were not reported were omitted from the analysis (*n* = 24, 0.04% of all patients). Patients with reported race as "Multi-Race/Ethnicity" (*n* = 15, 0.03% of all patients) were grouped with the category of reported race as "Other." The remaining selected demographic variables contained 224 unique combinations, resulting in 224 total bins (two COVID-19 statues × two sex at birth × seven race × eight age categories). The resulting multi-dimensional distributions were plotted on a three-dimensional heatmap for qualitative analysis. Within the heatmap, each unique combination of demographic variables was represented by a point, and the color of each point represented the population in each bin.

To quantitatively compare each joint distribution bin to a theoretical "perfect" balance of demographic variables, the absolute and percent differences from an exact 77:23 split were calculated for all demographic combinations evaluated, according to Equations (5.2) and (5.3), respectively.

$$(Absolute\ difference\ from\ expectation)_i = \left|(f)\left(N_{i,all}\right) - N_{i,public}\right| \tag{5.2}$$

$$(Percent\ difference\ from\ expectation)_i = \frac{\left|(f)\left(N_{i,all}\right) - N_{i,public}\right|}{(f)\left(N_{i,all}\right)} \tag{5.3}$$

In Equations (5.2) and (5.3), *i* represents the individual bin *(i* = 1, … 224) for which each difference was calculated, *f* represents the fraction placed in the open commons, $N_{i,all}$ is the total number of patients across both commons, and $N_{i,public}$ is the total number of patients in the public commons. Bins containing fewer than 10 patients (n = 69 bins) were not included in the calculation of percent difference from expectation, as the small denominator would induce large fluctuations in the value.

5.2.2 Task-based sampling

To describe the developed method of task-based sampling, we first begin with describing the associated processes of data use within MIDRC. The open data commons (approximately 80% of all images and metadata) is published publicly and serves as a large pool from which representative samples can be drawn to develop and test users' algorithms. Once a user has finalized their algorithm and is preparing for regulatory evaluation, they may submit the algorithm to MIDRC for formal performance evaluation using data from the sequestered commons (approximately 20% of all images and metadata). An overview of this process is shown in Figure 5.3. The sequestered commons will serve as a large pool from which task-specific samples can be drawn to evaluate the performance of user-developed algorithms. It is important to note that algorithms will be tested on subsets of a cohort matched to a user's task, to enable estimates of performance variation and an opportunity to have their algorithm re-evaluated on new data.

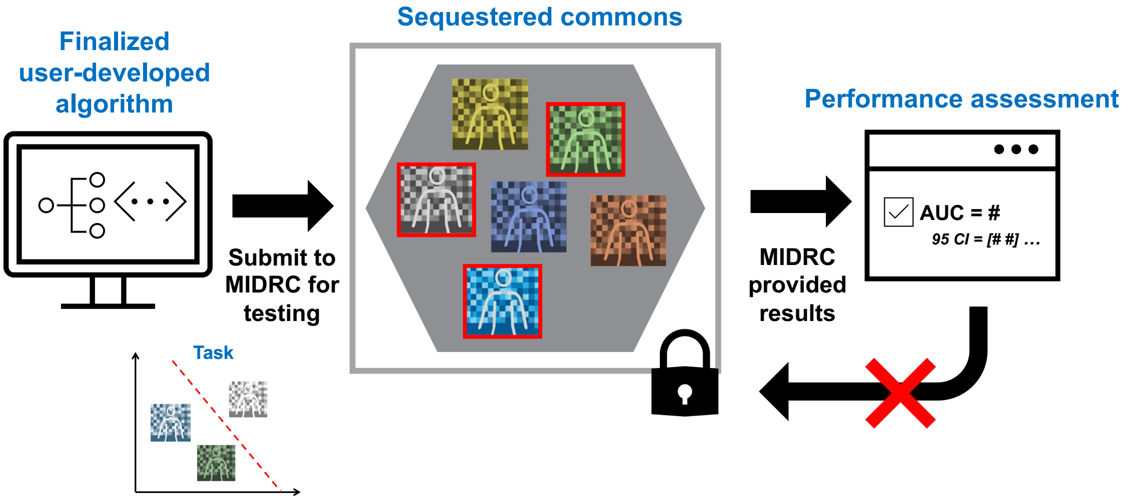


Figure 5.3. Use of sequestered data in MIDRC. A subset of the sequestered commons, which is restricted from public access, matched to the user's task is selected for algorithm performance assessment. MIDRC provides the user summary results without letting the user know what cases were used for testing, to maintain integrity of the sequestered data commons.

Task-based sampling optimization algorithm

In this workflow, the developed algorithm optimizes the initial cohort to be matched to the target population, while keeping as many patients as achievable. An overview of the described process is shown in Figure 5.4. To select a sample for testing of a research claim on a specified demographic, the user must first specify the initial cohort, the target population, and a threshold value used to stop the optimization.

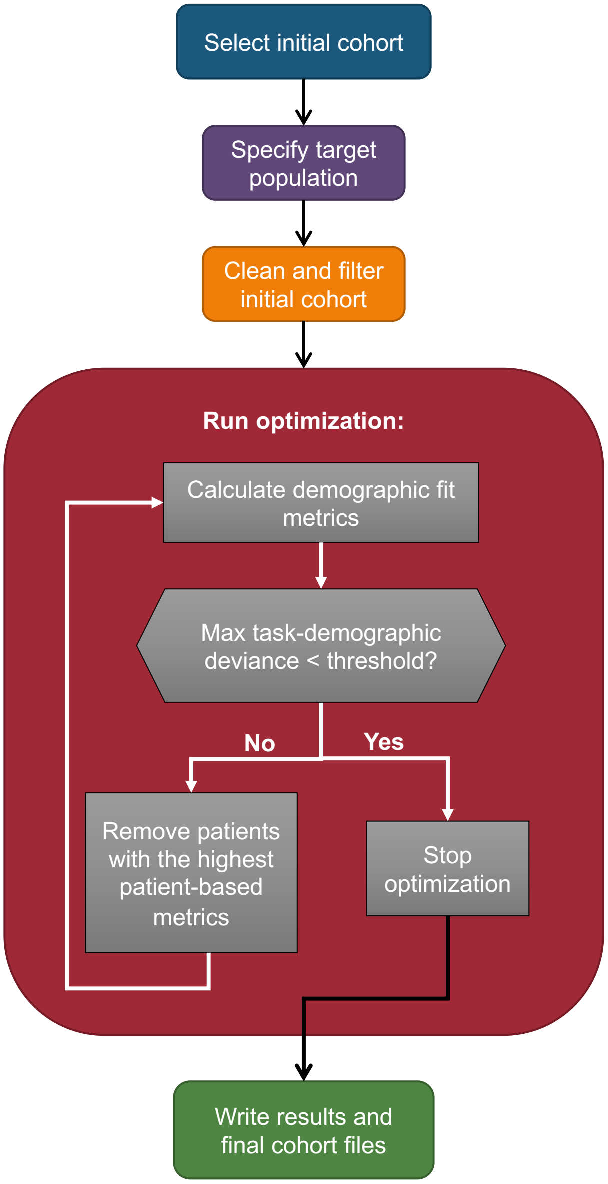

Select initial cohort
Specify target population
Clean and filter initial cohort
Run optimization:
Calculate demographic fit metrics
Max task-demographic deviance < threshold?
No
Yes
Remove patients with the highest patient-based metrics
Stop optimization
Write results and final cohort files

Figure 5.4. Schematic of the developed task-based sampling process and optimization algorithm flowchart.

The initial cohort is the subset to perform optimization on and draw from it a subset matched to the associated target population. Prior to beginning the optimization algorithm, the initial cohort should be filtered, to ensure all key variables are labeled uniformly, and only unique patients remain. Leaving duplicate patients in the optimization will lead to misleading results since demographic characteristics will be represented twice.

The target demographic distribution describes the population characteristics the user is seeking to match. For integration with the developed optimization algorithm and the data model within data.MIDRC.org, a standardized list for the demographic categories of age, race, sex at birth, ethnicity, and COVID-19 status was established, as shown in Table 5.1. This list outlines the 26 default subcategories for the four demographic categories listed above. A user must specify percentages of each subcategory to their target demographic distribution. For an example generated for this report, a target population approximately matched to the CDC case distribution, with a 50:50 COVID-19 status, was defined, as shown Table 5.1.

Table 5.1. Example target demographic specification. List includes the 26 default subcategories used in the optimization algorithm and the data model at data.MIDRC.org.

Demographic category	**Demographic subcategory**	**Example target distribution**
Age group	[0, 18)	0.00
	[18, 30)	0.26
	[30, 40)	0.21
	[40, 50)	0.17
	[50, 65)	0.22
	[65, 75)	0.08
	[75, 85)	0.04
	[85, 140)	0.02
	Not Reported	0.00
Race	American Indian or Alaska Native	0.01
	Asian	0.04
	Black or African American	0.12
	Native Hawaiian or other Pacific Islander	0.00
	Not Reported	0.00

Table 5.1. (Continued) Example target demographic specification. List includes the 26 default subcategories used in the optimization algorithm and the data model at data.MIDRC.org.

Race	Other	0.04
	White	0.79
Sex at birth	Female	0.53
	Male	0.47
	Other	0.00
	Not Reported	0.00
Ethnicity	Hispanic or Latino	0.25
	Not Hispanic or Latino	0.75
	Not Reported	0.00
COVID-19 Positive	No	0.50
	Not Reported	0.00
	Yes	0.50

The threshold value is defined as the maximum percent deviation allowable in any demographic subcategory, between the current sample and the target prevalence. The threshold is set by the user prior to initiating the optimization and is used to stop the optimization. This value, the maximum task-demographic deviance metric, is shown in Equation (5.4). In Equation (5.4), i represents each individual demographic subcategory, as defined in Table 5.1.

$$\text{Maximum task-demographic deviance metric} =$$
$$Max(|\text{Current prevalence}_i - \text{Target prevalence}_i\,|) \quad \text{for all } i \tag{5.4}$$

One the optimization process has begun, in each iteration, the distributions of all demographic subcategories are compared to the target prevalence distribution and each patient is assigned a patient demographic-fit deviance metric, calculated according to Equation (5.5).

$$\text{Patient demographic-fit deviance metric} =$$
$$\textstyle\sum_{i=1}^{n}(\text{Current prevalence}_i - \text{Target prevalence}_i) \ \text{ for all } i \text{ specific to each patient} \tag{5.5}$$

The patient demographic-fit deviance metric is equal to the sum of the deviances of each demographic subcategory of that patient to the target prevalence, where a negative value would indicate more of that given category is needed to reach the target prevalence, and a positive value indicates less of a given category is needed to reach the target demographic. The sum of the deviances jointly balances how the demographic profile of each patient matches the target prevalence distribution. After calculation of both metrics, if the maximum task-demographic deviance metric is greater than the specified threshold, patients with the highest patient demographic-fit deviance metrics are removed from the sample at a fixed rate and the algorithm begins a new iteration. The rate at which patients are removed from the sample is the greater of 1% of the current sample or 5 patients, by default, but may be modified to accommodate different circumstances. Once the calculated maximum task-demographic deviance metric is less than the specified threshold, the process is complete, and the selected sample may be saved for future use.

Example cohort task-based sampling

To demonstrate the task-based sampling process in simulation, metadata from 5539 imaging studies (4193 patients) were selected from the public data commons (data.MIDRC.org) to serve as an example of an initial cohort. This cohort was selected on March 11, 2023, as imaging studies with selected LOINC properties of Method (Modality) = 'CT' and System (Body Region) = 'Chest'. The target demographic distribution was defined in agreement with the modified CDC case demographic as listed in Table 5.1.

Patients with listed age as less than 18 or not reported were removed from the sample (n = 5 studies), and multiple entries from the same patient were removed to have only one study per patient remain (n = 1318 studies). The developed task-based sampling algorithm was applied to the remaining 4016 patients to select a sample matched to the target demographic distribution with a maximum task-demographic deviance metric thresholds of 10% and 5%. To summarize the performance, we calculated the difference in each demographic subcategory from the final cohort generated with each maximum task-demographic deviance metric threshold to the target distribution.

5.3 Results

5.3.1 Sequestration by stratified sampling

Demonstrating sequestration algorithm use in an example database

Results obtained from splitting the input dataset using our stratified sampling method over 2000 trials are shown in Table 5.2. The mean and standard deviation of each demographic category's prevalence was calculated over all trials and compared to the prevalence in the input dataset and between the two subsets. For all demographic subcategories, the prevalence in the input dataset was matched in both the open and sequestered subsets within one standard deviation. For the category of image modality, the prevalence represents the percentage of patients with a given image modality available, and since many patients have images from multiple modalities, these percentages will not add to 100%.

Table 5.2. Distribution of all balanced variables in the input, open (approximately 80%), and sequestered (approximately 20%) datasets following splitting via stratified sampling. Prevalence values are written as the mean percent (standard deviation) over 2000 independent trials. The label of "Not Reported" was added to variables with blank entries.

Demographic Subcategory	Input Dataset Count	Input Dataset Prevalence	Open Subset Prevalence	Sequestered Subset Prevalence
Age Group				
[0, 18)	74	1.5%	1.5% (0.1%)	1.4% (0.3%)
[18, 30)	393	7.9%	7.9% (0.1%)	7.9% (0.3%)
[30, 40)	529	10.6%	10.6% (0.1%)	10.3% (0.4%)
[40, 50)	687	13.7%	13.7% (0.1%)	13.9% (0.4%)
[50, 65)	1434	28.7%	28.6% (0.1%)	29.1% (0.4%)
[65, 75)	909	18.2%	18.2% (0.1%)	18.1% (0.4%)
[75, 85)	597	11.9%	11.9% (0.1%)	11.9% (0.3%)
[85, 140)	284	5.7%	5.7% (0.1%)	5.5% (0.3%)
Not Reported	93	1.9%	1.8% (0.1%)	1.9% (0.2%)

Table 5.2. (Continued) Distribution of all balanced variables in the input, open (approximately 80%), and sequestered (approximately 20%) datasets following splitting via stratified sampling. Prevalence values are written as the mean percent (standard deviation) over 2000 independent trials. The label of “Not Reported” was added to variables with blank entries.

Race				
American Indian or Alaska Native	17	0.3%	0.3% (0.0%)	0.3% (0.2%)
Asian	294	5.9%	5.9% (0.1%)	5.9% (0.4%)
Black or African American	1386	27.7%	27.8% (0.1%)	27.6% (0.4%)
Native Hawaiian or other Pacific Islander	15	0.3%	0.3% (0.0%)	0.3% (0.2%)
White	2568	51.4%	51.2% (0.1%)	51.8% (0.6%)
Not Reported	554	11.1%	11.1% (0.1%)	10.8% (0.5%)
Other	166	3.3%	3.3% (0.1%)	3.3% (0.4%)
Sex at Birth				
Female	2533	50.7%	50.6% (0.1%)	50.8% (0.5%)
Male	2464	49.3%	49.3% (0.1%)	49.1% (0.5%)
Other	0	0.0%	0.0% (0.0%)	0.0% (0.0%)
Not Reported	3	0.1%	0.1% (0.0%)	0.1% (0.1%)
Ethnicity				
Hispanic or Latino	499	10.0%	10.0% (0.1%)	9.7% (0.6%)
Not Hispanic or Latino	4443	88.9%	88.8% (0.2%)	89.2% (0.6%)
Not Reported	58	1.2%	1.2% (0.1%)	1.2% (0.3%)
COVID-19 Status				
No	2602	52.0%	52.0% (0.1%)	52.2% (0.5%)
Not Reported	1	0.0%	0.0% (0.0%)	0.0% (0.0%)
Yes	2397	47.9%	48.0% (0.1%)	47.8% (0.5%)
Image Modality				
CR	2049	41.0%	40.9% (0.2%)	41.3% (0.8%)
CT	910	18.2%	18.3% (0.2%)	18.0% (0.8%)
DX	2596	51.9%	52.0% (0.1%)	51.9% (0.5%)
MR	27	0.5%	0.5% (0.0%)	0.5% (0.2%)

Comparison of sequestration algorithm to naïve random sampling

Histograms of the scaled difference from expectation over 2000 independent trials for the categories of age and race in the “Open” subset (approx. 80%) are shown in Figure 5.5. For most categories analyzed, sequestration by stratified sampling provided lower scaled differences from expectation, in general, than from the naïve randomization, as indicated by the narrower distributions for stratified sampling than naïve randomization. However, for some categories with low prevalence, such as race of American Indian or Native American, stratified sampling showed similar variation from expectation as the naïve randomization.

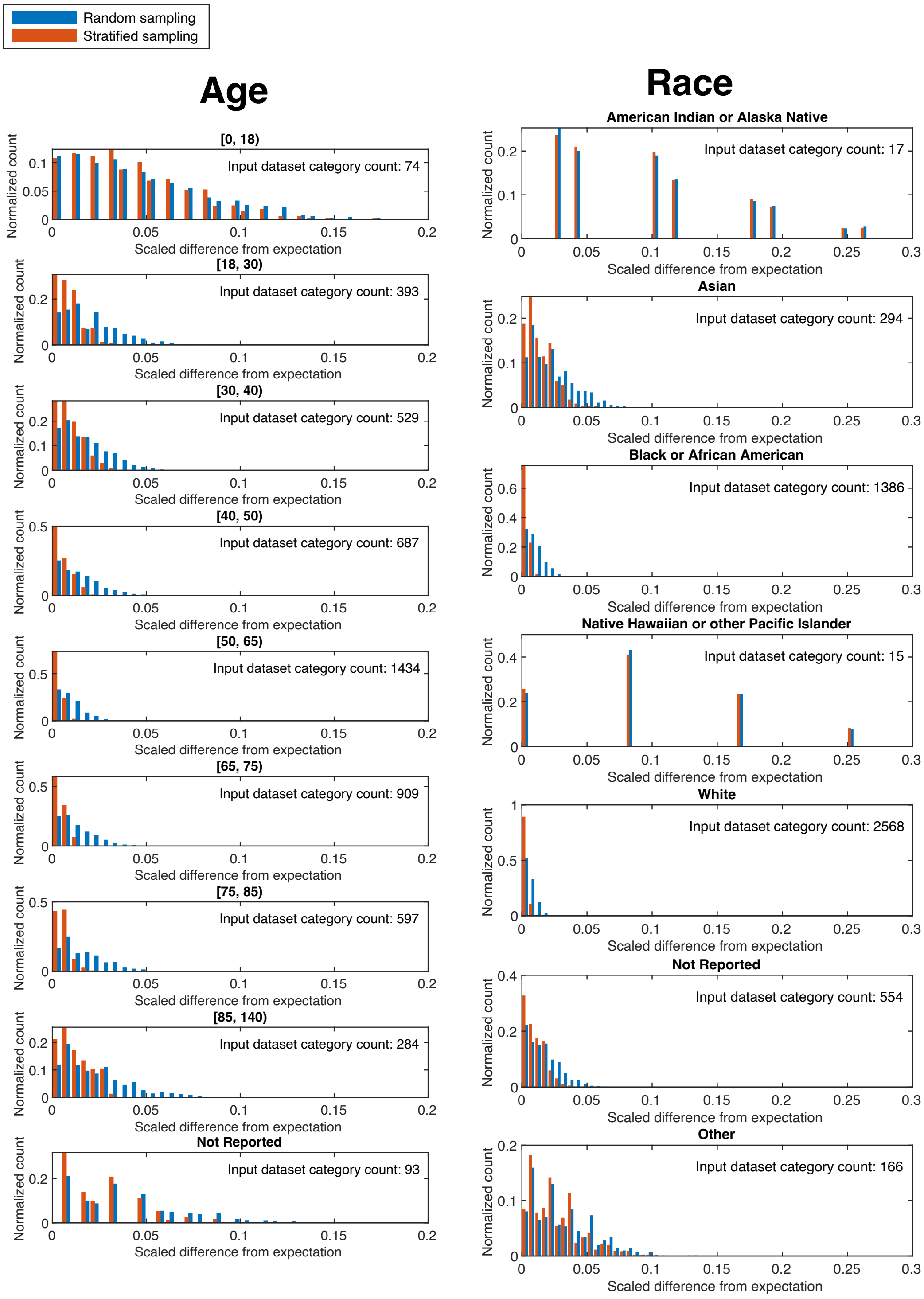


Figure 5.5 Histograms of the scaled difference from expectation for the categories of age and race in the “Open” dataset after 2000 independent trials of dataset splitting using naïve random sampling (blue) and our stratified sampling method (orange).

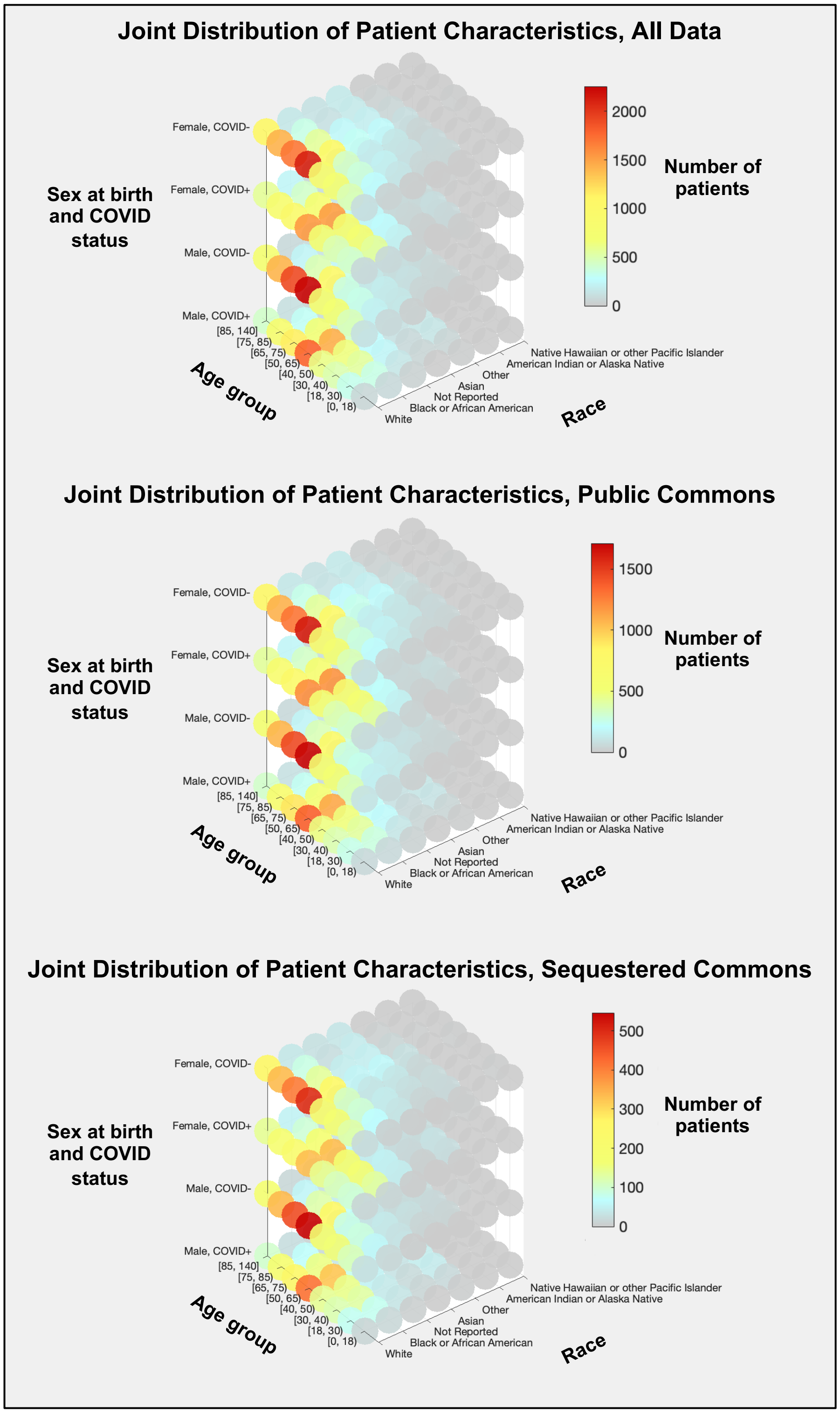
Results of the statistical comparison of the histograms of the scaled difference from expectation for all demographic subcategories categories in the “Open” subset using the one-tailed Mann-Whitney U test are shown in Table 5.3. Similar results were found for the “Sequestered” subset, despite having smaller relative sample size, but are omitted for brevity. Statistical results support the qualitative summary that sequestration by stratified sampling provided lower differences from expectation, in general, than did the naïve randomization, with the exception of demographic subcategories with very low prevalence. *P*-values less than 0.05 were considered significant. Correction for multiple comparisons using Holm-Bonferroni did not change significance between the differences from expectation for any subcategory except the category of having an Ethnicity of Not Hispanic or Latino. As a result, for most demographic categories, using our developed method of stratified sampling provided significantly “more balanced” distributions on average than naïve random sampling.

Table 5.3. Results of the one-tailed Mann-Whitney U test, comparing distributions of the scaled difference from expectation in the "Open" dataset after 2000 independent trials of dataset splitting using naïve random sampling and using our stratified sampling method. P-values shown in bold were considered significant ($p < 0.05$) and p-values with two asterisks were considered significant after Holm-Bonferroni multiple comparison correction.

Demographic Subcategory	Input dataset count	Mann-Whitney U test result
Age Group		
[0, 18)	74	***p < 0.01*****
[18, 30)	393	***p < 0.01*****
[30, 40)	529	***p < 0.01*****
[40, 50)	687	***p < 0.01*****
[50, 65)	1434	***p < 0.01*****
[65, 75)	909	***p < 0.01*****
[75, 85)	597	***p < 0.01*****
[85, 140)	284	***p < 0.01*****
Not Reported	93	***p < 0.01*****
Race		
American Indian or Alaska Native	17	*p = 0.70*
Asian	294	***p < 0.01*****
Black or African American	1386	***p < 0.01*****
Native Hawaiian or other Pacific Islander	15	*p = 0.29*
White	2568	***p < 0.01*****
Not Reported	554	***p < 0.01*****
Other	166	***p < 0.01*****
Sex at Birth		
Female	2533	***p < 0.01*****
Male	2464	***p < 0.01*****
Other	0	*N/A*
Not Reported	3	*p = 0.52*
Ethnicity		
Hispanic or Latino	499	***p < 0.01*****
Not Hispanic or Latino	4443	***p = 0.01***
Not Reported	58	*p = 0.57*
COVID-19 Status		
No	2602	***p < 0.01*****
Not Reported	1	*p = 0.29*
Yes	2397	***p < 0.01*****
Image Modality		
CR	2049	***p < 0.01*****
CT	910	***p < 0.01*****
DX	2596	***p < 0.01*****
MR	27	***p < 0.01*****

Evaluation of sequestration methodology through joint distribution balance

Joint distributions of patient characteristics in both the public and sequestered imaging commons were found to closely match each other as well as that of all available data. Qualitative results showing three-dimensional heatmaps of all joint distribution categories analyzed in all available data, the public commons, and the sequestered commons, are plotted in Figure 5.6. From these plots, it can be noted that the joint distributions of patients within both the public and sequestered commons follow a very similar pattern to the joint distributions of patients from all available data, and to each other.



Joint Distribution of Patient Characteristics, All Data

Joint Distribution of Patient Characteristics, Public Commons

Joint Distribution of Patient Characteristics, Sequestered Commons

Figure 5.6. Joint distributions of patient demographic characteristics of 54,185 patients collected from August 2021 to August 2022 in all available data, the public commons, and the sequestered commons. Color scales have been adjusted to reflect the maximum number of patients in one bin for each set.

Figure 5.7 shows histograms of (a) the total number of patients in each of the 224 unique patient-characteristic bins from all available data, (b) the absolute difference from expectation in the public commons, Equation (5.2), and (c) the percent difference from expectation in the public commons in bins containing 10 or more patients (n = 155 of 224 total bins), Equation (5.3). Since the absolute difference from expectation is inherently mirrored in the public and sequestered commons, as patients are only ever in one of the two commons, the histogram of the absolute difference from expectation for the sequestered commons would be identical to Figure 5.7 (b). Absolute differences in the patient population of each bin from an exact split indicated 54.0% of bins obtained differences of 5 patients or less, and 75.9% of bins obtained differences of 15 patients or less, with a median difference of 3.6 patients from the total data for both public and sequestered commons. Percent differences in the patient population of each bin from an exact 77:23 split in the open commons indicated 54.5% of bins obtained a percent difference of 5.0% or less, with a median percent difference of 4.6% for the public commons.

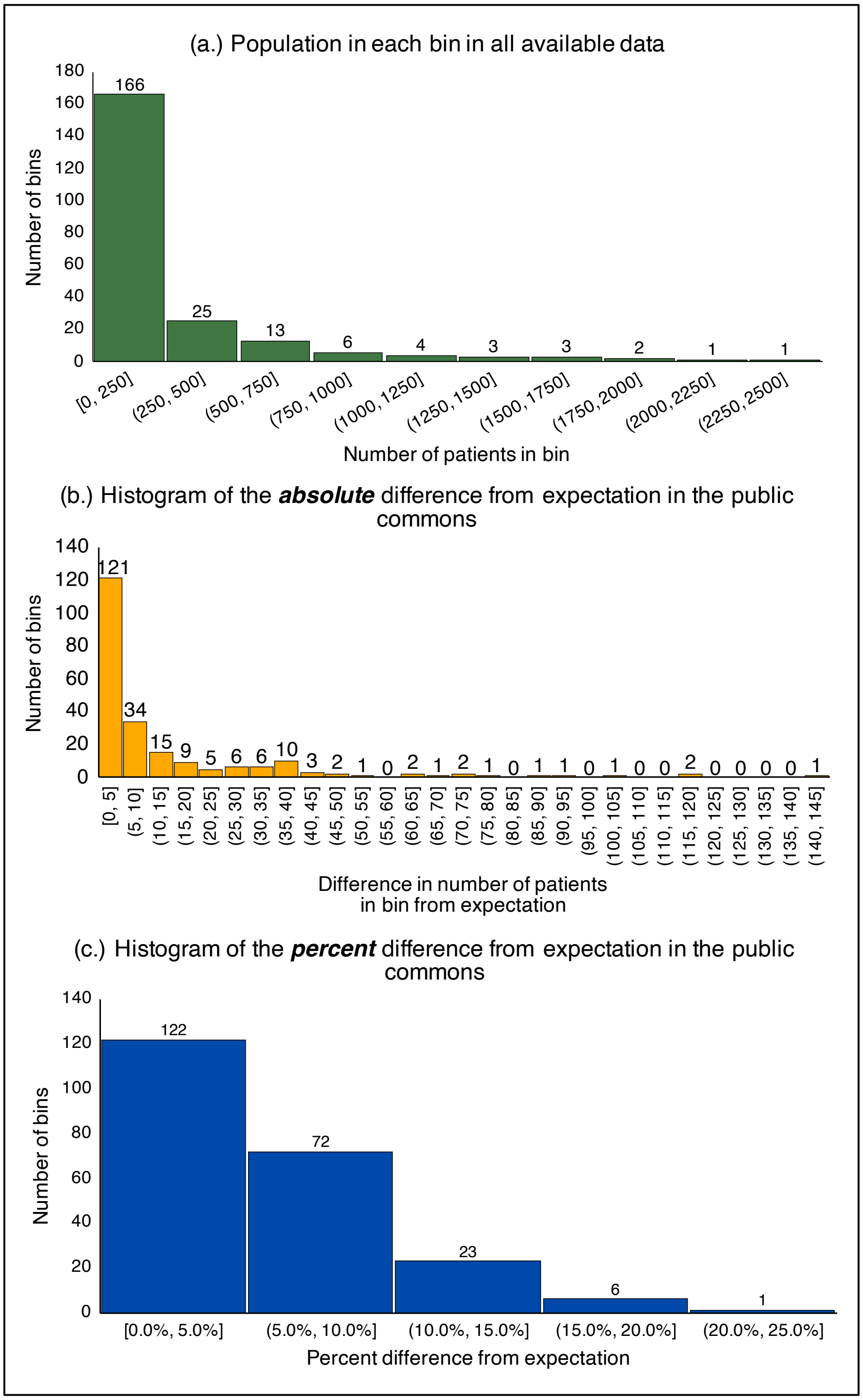


Figure 5.7. Histogram of (a) the total population in each bin from all available data, (b) the absolute difference from expectation in the public commons (Eq. 1), and (c) the percent difference from expectation in the public commons (Eq. 2).

As can be seen from Figure 5.6 and Figure 5.7, there exists a large variation in the population of individual demographic category bins across the 224 total bins. Over 50% of the bins contained 250 patients or less, while the highest populated bin contained 2256 patients (representing White, 50-64 year old, female, patients with no positive COVID-19 test). This highly skewed data distribution is an important aspect in the subsequent analysis of absolute and percentage differences. Since the majority of bins are quite sparsely populated the absolute difference from expectation, Equation (5.2), may give a more interpretable metric of "achieved balance" for sparse bins. For example, a difference of less than 5 patients in a bin of 50 patients may be simply deemed reasonable balance. However, since the absolute difference from expectation, Equation (5.2), is not scaled by the individual bin population, its overall magnitude tends to be correlated with the bin population. A difference, for example, of 10 patients from a bin of 20 total patients or 10 patients from a bin of 2000 total patients would be interpreted very differently in terms of "achieved balance". Conversely, the percent difference from expectation, Equation (5.3), is scaled by the individual bin population; but for very sparsely populated bins, a small denominator may inflate the percentage, reducing the interpreted "achieved balance", while the absolute difference may only be off by a few patients. Given these two conflicting circumstances, and the highly skewed nature of the data distributions, both metrics are included for evaluation. However, a threshold of at least 10 patients in each bin was implemented for the percent difference from expectation, Equation (5.3), to avoid over-inflation from very sparsely populated bins.

5.3.2 Task-based sampling

Example cohort task-based sampling

Results obtained from splitting the initial cohort using our task-based sampling method with maximum task-demographic deviance thresholds of 10% and 5% are shown in Table 5.4 and

Table 5.5, respectively. Using the threshold of 10%, the resulting final cohort contained 870 unique patients. Using the threshold of 5%, the resulting final cohort contained 542 patients. The average absolute difference across all demographic subcategories was 2.1% and 1.0% for thresholds of 10% and 5%, respectively.

Table 5.4. Task-based sample results for 10% threshold of the maximum task-demographic deviance metric. The target demographic prevalence is defined in alignment with Table 5.1. Differences in prevalence from the target greater than 10% are shown in italics, and the maximum value in each difference column is shown in bold and italics.

		Initial cohort, n = 4016			**Final cohort: 10% threshold, n = 870**		
Demographic subcategory	**Target prevalence**	**Initial cohort count**	**Initial cohort prevalence**	**Difference from target**	**Final cohort count**	**Final cohort prevalence**	**Difference from target**
Age							
[0, 18)	**0.0%**	0	0.0%	0.0%	0	0.0%	0.0%
[18, 30)	**26.0%**	146	3.6%	*22.4%*	140	16.1%	***9.9%***
[30, 40)	**20.5%**	247	6.2%	*14.3%*	190	21.8%	-1.3%
[40, 50)	**17.4%**	380	9.5%	7.9%	167	19.2%	-1.8%
[50, 65)	**22.1%**	1259	31.3%	-9.2%	210	24.1%	-2.0%
[65, 75)	**8.1%**	1071	26.7%	*-18.6%*	84	9.7%	-1.6%
[75, 85)	**3.9%**	698	17.4%	*-13.5%*	46	5.3%	-1.4%
[85, 140)	**2.0%**	215	5.4%	-3.4%	33	3.8%	-1.8%
Not Reported	**0.0%**	0	0.0%	0.0%	0	0.0%	0.0%
Race							
American Indian or Alaska Native	**1.1%**	10	0.2%	0.9%	2	0.2%	0.9%
Asian	**3.8%**	320	8.0%	-4.2%	16	1.8%	2.0%
Black or African American	**12.3%**	714	17.8%	-5.5%	87	10.0%	2.3%
Native Hawaiian or other Pacific Islander	**0.3%**	9	0.2%	0.1%	2	0.2%	0.1%
Not Reported	**0.0%**	701	17.5%	*-17.5%*	80	9.2%	-9.2%
Other	**4.0%**	62	1.5%	2.5%	23	2.6%	1.4%
White	**78.5%**	2200	54.8%	***23.7%***	660	75.9%	2.6%

Table 5.4. (Continued) Task-based sample results for 10% threshold of the maximum task-demographic deviance metric. The target demographic prevalence is defined in alignment with Table 5.1. Differences in prevalence from the target greater than 10% are shown in italics, and the maximum value in each difference column is shown in bold and italics.

Sex at birth							
Female	**53.1%**	1955	48.7%	4.4%	464	53.3%	-0.2%
Male	**46.9%**	2060	51.3%	-4.4%	406	46.7%	0.2%
Other	**0.1%**	0	0.0%	0.1%	0	0.0%	0.1%
Not Reported	**0.1%**	1	0.0%	0.1%	0	0.0%	0.1%
Ethnicity							
Hispanic or Latino	**25.0%**	309	7.7%	*17.3%*	155	17.8%	7.2%
Not Hispanic or Latino	**75.0%**	3468	86.4%	*-11.4%*	695	79.9%	-4.9%
Not Reported	**0.0%**	239	6.0%	-6.0%	20	2.3%	-2.3%
COVID-19 positive							
No	**50.0%**	2517	62.7%	*-12.7%*	438	50.3%	-0.3%
Not Reported	**0.0%**	0	0.0%	0.0%	0	0.0%	0.0%
Yes	**50.0%**	1499	37.3%	*12.7%*	432	49.7%	0.3%

Table 5.5. Task-based sample results for 5% threshold of the maximum task-demographic deviance metric. The target demographic prevalence is defined in alignment with Table 5.1. Differences in prevalence from the target greater than 10% are shown in italics, and the maximum value in each difference column is shown in bold and italics.

		Initial cohort, n = 4016			**Final cohort: 5% threshold, n = 542**		
Demographic subcategory	**Target prevalence**	**Initial cohort count**	**Initial cohort prevalence**	**Difference from target**	**Final cohort count**	**Final cohort prevalence**	**Difference from target**
Age							
[0, 18)	**0.0%**	0	0.0%	0.0%	0	0.0%	0.0%
[18, 30)	**26.0%**	146	3.6%	*22.4%*	127	23.4%	2.6%
[30, 40)	**20.5%**	247	6.2%	*14.3%*	110	20.3%	0.2%
[40, 50)	**17.4%**	380	9.5%	7.9%	101	18.6%	-1.2%
[50, 65)	**22.1%**	1259	31.3%	-9.2%	121	22.3%	-0.2%
[65, 75)	**8.1%**	1071	26.7%	*-18.6%*	49	9.0%	-0.9%
[75, 85)	**3.9%**	698	17.4%	*-13.5%*	22	4.1%	-0.2%

Table 5.5. (Continued) Task-based sample results for 5% threshold of the maximum task-demographic deviance metric. The target demographic prevalence is defined in alignment with Table 5.1. Differences in prevalence from the target greater than 10% are shown in italics, and the maximum value in each difference column is shown in bold and italics.

Age (continued)							
[85, 140)	**2.0%**	215	5.4%	-3.4%	12	2.2%	-0.2%
Not Reported	**0.0%**	0	0.0%	0.0%	0	0.0%	0.0%
Race							
American Indian or Alaska Native	**1.1%**	10	0.2%	0.9%	2	0.4%	0.7%
Asian	**3.8%**	320	8.0%	-4.2%	14	2.6%	1.2%
Black or African American	**12.3%**	714	17.8%	-5.5%	60	11.1%	1.2%
Native Hawaiian or other Pacific Islander	**0.3%**	9	0.2%	0.1%	2	0.4%	-0.1%
Not Reported	**0.0%**	701	17.5%	*-17.5%*	25	4.6%	***-4.6%***
Other	**4.0%**	62	1.5%	2.5%	17	3.1%	0.9%
White	**78.5%**	2200	54.8%	***23.7%***	422	77.9%	0.6%
Sex at birth							
Female	**53.1%**	1955	48.7%	4.4%	291	53.7%	-0.6%
Male	**46.9%**	2060	51.3%	-4.4%	251	46.3%	0.6%
Other	**0.1%**	0	0.0%	0.1%	0	0.0%	0.1%
Not Reported	**0.1%**	1	0.0%	0.1%	0	0.0%	0.1%
Ethnicity							
Hispanic or Latino	**25.0%**	309	7.7%	*17.3%*	113	20.8%	4.2%
Not Hispanic or Latino	**75.0%**	3468	86.4%	*-11.4%*	421	77.7%	-2.7%
Not Reported	**0.0%**	239	6.0%	-6.0%	8	1.5%	-1.5%
COVID-19 positive							
No	**50.0%**	2517	62.7%	*-12.7%*	273	50.4%	-0.4%
Not Reported	**0.0%**	0	0.0%	0.0%	0	0.0%	0.0%
Yes	**50.0%**	1499	37.3%	*12.7%*	269	49.6%	0.4%

To describe the process of the optimization algorithm as it selects patients for the final cohort, the maximum task-demographic deviance metric was plotted as a function of the optimization iteration for both thresholds, as shown in Figure 5.8 and Figure 5.9. These plots may be interpreted similar to an optimization loss function. As expected, Figure 5.8 represents a subset of Figure 5.9, as the optimization follows the same process to reach a threshold of 10% before continuing on to the final value of 5% in Figure 5.9. As such, we can establish the optimization algorithm follows a deterministic path given the same initial cohort and set of input parameters. Discontinuities in the plot slope indicate a shift from one category dominating the maximum task-demographic deviation to another category, as subjects are removed from the sample and overall demographic distributions shift. Jagged fluctuations in the maximum task-demographic deviance metric, as seen toward the end of both figures, indicate rapid shifts in the current sample's demographic distribution with each iteration of patients that are removed. Small fluctuations such as those present in Figure 5.8 and Figure 5.9, are to be expected for a relatively small sample size. However, if this metric plots erratically or unexpectedly increases, the demographic distribution of the sample is shifting too fast for the optimization. The number of patients removed in each iteration may need to be decreased, or the optimization may be initiated with a higher threshold.

In each iteration of drawing from the available data to reach the specified population, every patient is assigned a demographic-fit deviance metric, in which higher values indicate the demographic characteristics of that patient are over-represented in the sample compared to the target population (i.e., having higher prevalence than the target population). The patient demographic-fit deviance metric histograms are shown in Figure 5.10 for all patients in the current sample at four selected timepoints in the optimization: (a) near initiation at 10 iterations, (b) at 50 iterations, (c) at the final iteration for the threshold of 10%, and (d) at the final iteration for the threshold of 5%. As the histogram narrows, the drawn sample becomes closer to the target population. However, the histogram may also take on a bi-modal shape around zero, such as is shown in Figure 5.10 (c), as a few demographic categories are strongly overrepresented, while others are underrepresented, skewing the "balance".

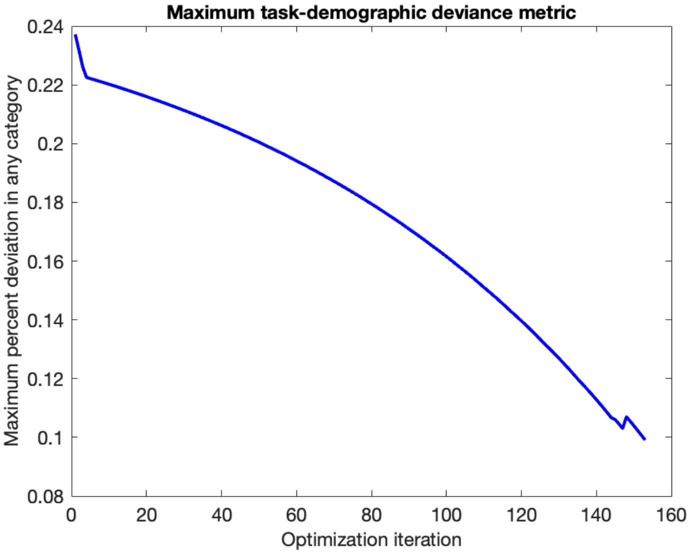


Figure 5.8. Maximum task-demographic deviance metric plotted for algorithm initiated with initial cohort as defined in Table 5.4 and 10% stopping threshold, minimum percent deviation equal to 0.1.

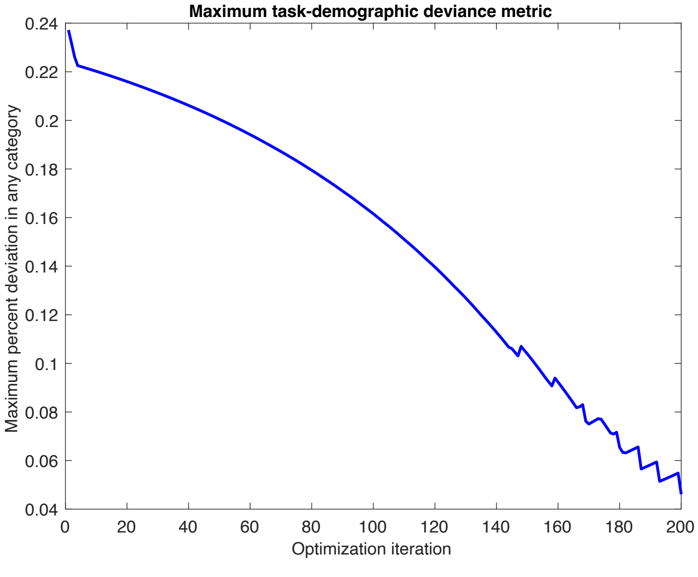
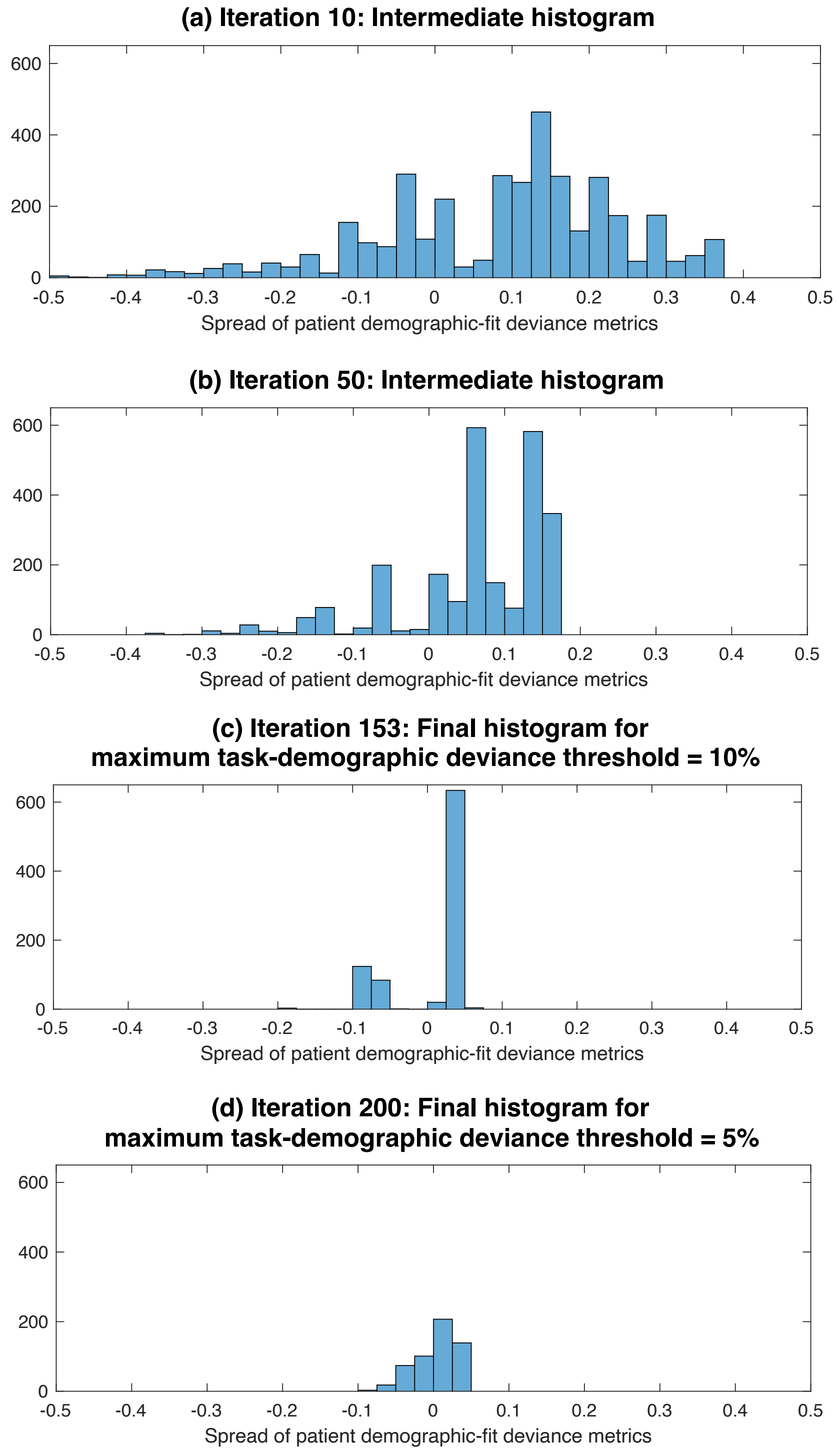


Figure 5.9. Maximum task-demographic deviance metric plotted for algorithm initiated with initial cohort as defined in Table 5.5 and 5% stopping threshold, minimum percent deviation equal to 0.05.


(a) Iteration 10: Intermediate histogram

(b) Iteration 50: Intermediate histogram

(c) Iteration 153: Final histogram for maximum task-demographic deviance threshold = 10%

(d) Iteration 200: Final histogram for maximum task-demographic deviance threshold = 5%

Figure 5.10. Patient demographic-fit deviance metric histograms for patients in the current sample at four selected timepoints in the optimization: (a) near initiation at 10 iterations, (b) at 50 iterations, (c) at the final iteration for the threshold of 10%, and (d) at the final iteration for the threshold of 5%. A narrower histogram indicates a closer match to the target distribution.

5.4 Discussion

5.4.1 Sequestration by stratified sampling

We demonstrated, using our proposed method of multi-dimensional stratified sampling, that splitting an input dataset of 5000 COVID-19 patients into an 80% open dataset and a 20% sequestered dataset based on the variables of age, race, sex at birth, ethnicity, COVID-19 status, and image modality resulted in subsets that exhibited distributions very similar to those of the input dataset and each other. The high degree of similarity in the distributions indicates that the sequestration algorithm operated as expected. Moreover, distributions of the differences from the expected values for the developed stratified sampling algorithm and naïve randomization indicated that the stratified sampling algorithm provided, in general, more balanced distributions of variables in subsets of patients versus those obtained from the naïve randomization.

Assessment of machine learning algorithm performance is often achieved through methods such as k-fold cross validation or bootstrapping [117]. These methods sample a limited dataset many times to test the algorithm on a variety of sample characteristics. Additionally, stratified randomization is an existing process used in separating training and testing datasets, but generally only allows for stratification across a single variable. However, balancing of multiple variables across public and sequestered datasets, from which cases cannot be made known, or used and replaced, is a task not typically considered in machine intelligence applications. Using the presented process, which sequentially steps through each branch until a single multi-variable stratum is obtained, balance across all possible combinations of the selected variables can be controlled. Similar processes are used in the construction of case and control populations in clinical trials, but these processes are typically conducted once, after collection of the entire population. Our process is implemented on each incoming MIDRC data batch, which are received on an ongoing basis. To the knowledge of the authors, this is the first application to machine intelligence datasets [112,113].

Results of the evaluation of sequestration methodology through joint distribution balance showed an in-depth investigation of the demographic balance, beyond simple comparison of individual demographic distributions, evaluating the joint distribution of four key demographic variables from over 54,000 unique patients collected over one year. Results demonstrated that the joint distributions for data submitted to MIDRC from August 2021 to August 2022 in the public and sequestered subsets reasonably match the joint distributions from all available data. This high degree of balance indicates that the multi-dimensional stratified sampling algorithm, used to separate the data into the respective subsets, is operating as intended and both data commons are representative of the data available.

While the high degree of similarity in the distributions of variables across both subsets is promising, indicating that the proposed sampling method worked as intended, the ultimate goal in constructing a sequestered dataset for algorithm evaluation does not aim for perfect symmetry relative to the data going to the open dataset. This is to avoid matching a test set to a training set, since that could allow one to approximately "train to the test." Sequestration will provide an ongoing method to monitor and maintain a high level of similarity in the variable distributions, but perturbations in the demographics will also be purposely implemented to assure algorithm generalizability. When an algorithm is tested using data from the sequestered dataset, test samples will be drawn from the sequestered set according to the distributions related to the task (e.g., clinical question, clinical claim, intended population), that is, the sequestered set in its entirety will not be used in the test. Furthermore, from the algorithm testing using sequestered data, only summary performance information will be reported back and not case-specific results.

5.4.2 Task-based sampling

Using the developed method of task-based sampling, the process and outcomes of drawing subsets matched to an approximate CDC distribution from an example cohort of over 4000 patients were described. Results indicated the developed optimization algorithm operates as expected, selecting subsets matched to the target demographic distribution within the specified thresholds of the maximum task-demographic deviance metric. Explanatory figures of the maximum task-demographic deviance metric and the patient demographic-fit histograms provide supportive feedback on the optimization process, such that the optimization parameters may be modified to ensure an acceptable result.

This research outlines the proposed framework for large-scale sampling of task-based populations to create test sets from a large data commons. Using the described task-based sampling optimization algorithm will provide researchers with the ability to curate a dataset from public data repositories for AI development or for internal MIDRC performance assessment to nearly any demographic distribution, through a relatively simple and efficient workflow. It will also allow MIDRC to perform formalized evaluation of algorithms using data from the sequestered commons that are matched to a user's task with a reduced burden of data curation and patient selection. If task-based sampling is being used for cohort building in the open data commons, the initial cohort can be specified by the metadata downloadable from the MIDRC data explorer. Alternatively, if used for algorithm performance assessment on sequestered data, the initial cohort will be selected internally.

A large, demographically diverse sequestered commons which is restricted from public access and can be repeatedly sampled for task-specific algorithm testing may provide a new gold standard for performance verification and could allow for expedited regulatory clearance of algorithms if accepted by regulating bodies. However, for performance assessment from the sequestered commons, it is important to note that algorithms will only be tested on subsets of task-based cohorts matched to a user's task. This will enable estimates of performance variation and allow users the opportunity to have their algorithm re-evaluated on new data, if elected. Users will not be allowed to have their algorithm tested on sequestered data an unregulated number of times, or without due process, to prevent the possibility of users training to the test distribution.

The methodology of task-based sampling described here was also implemented for selection of test and validation cohorts for the first two MIDRC Grand Challenges. The first challenge sought to classify chest radiographs of patients with COVID-19 from those without COVID-19. From an initial cohort of 4,639 patients processed, but not yet published on the public-facing data portal, 409 patients matched with a threshold maximum deviance metric of 10% to an approximate CDC case distribution were selected, and 334 patient were used in the challenge after adjustments for image quality. The 334 patients were separated between test and validation using the sequestration methodology described in Section 5.2.1 to maintain approximately even distributions of demographic variables in validation and test sets. The second Grand Challenge will seek to evaluate COVID-19 severity on chest radiographs. From an initial cohort of 6,202 patients processed, but not yet published on the public-facing data portal, 1,502 patients were selected with a threshold maximum deviance metric of 10% to an entirely COVID-19 positive, approximate CDC case distribution. The number of patients in the final test and validation cohorts for this challenge will be made public when the challenge is hosted in Summer 2023. Variance in the proportion of patients selected for each of these challenges from the patients initially available were likely due to the initial distributions of patients available from which to draw.

5.4.3 Limitations and future work

For all developed methodologies and measurements of evaluation described, the size and initial demographic distribution of the input dataset may present a limitation. An input dataset that does not contain sufficient patients may result in poor performance of the described methods. However, the "sufficient" number of patients is dependent on the specified goals of each task. The ability of stratified sampling to achieve a much higher degree of balance than simple randomization is highly dependent on the incoming dataset size and prevalence of a given demographic subcategory. This is noted to be a limitation for clinical trials that use similar methods as well [112]. Incoming dataset size and distributions are also a limitation in the evaluation of the joint demographic distributions, as sparsely populated bins may also be more difficult to quantify "achieved balance." For task-based sampling, the ability of the developed optimization program to achieve the specified level of similarity to a target demographic distribution is likewise limited by the availability of patients that match the target distribution in the initial cohort.

An important component of this work is the validation of the demographic balance between patients placed in the public and sequestered commons. Recent work by Gichoya et al. has illustrated the potential influence of patient demographic profiles on machine learning network performance and generalizability due to unintended bias [106]. MIDRC places a strong emphasis on ensuring the data publicly available on data.midrc.org and data held in the sequestered commons are diverse and representative of the intended population. Ensuring appropriate representation of demographic characteristics in the MIDRC data commons will increase user and regulatory body confidence that performance estimates from algorithm testing will generalize to real-world use. While ensuring that the distributions of demographic characteristics remain balanced across the two data commons may aid in a reduction of potential algorithm bias, algorithm bias may still arise and must be monitored. It is also acknowledged that certain labels of race, ethnicity, or sex assigned at birth may not adequately describe all populations or provide a clear correlate to genetic ancestry. Analysis of demographic balance will be modified to accommodate additional descriptors as they become available through future data contributions and on the data model available at data.MIDRC.org.

5.5 Conclusions

In summary, this research presents a novel sequestration method using multi-dimensional stratified sampling that effectively separates incoming data batches into "open" and "sequestered" commons and assessed the performance of this method in terms of similarities between the two commons. In addition, we proposed a task-based distribution sampling method to draw from a data commons a sample matched to a specified demographic distribution. With the continuous growth of both commons, performance of the developed methodology with respect to changing population distributions will continue to be monitored. While these methods were primarily developed for use in the MIDRC, we believe they can be applied to other databases and research fields as well. Overall, our findings demonstrate the potential of these methods to improve data management and analysis across various domains.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

This work presents major contributions to the field of mammography-based breast cancer risk assessment through characterization of AI-based features across the parenchymal field and evaluation of a temporal classifier for predicting future occurrence of cancer. Also, contributions to the field of data informatics for AI applications were made through development, evaluation, and implementation of methodology for sequestration and task-based sampling of imaging studies for a COVID-19 use-case.

The use of AI methods in mammography may improve screening sensitivity in the general population through identification of a potential cancer field effect in the breast parenchyma. Chapter 3 demonstrated the characterization of relationships among computer-extracted features in women with cancer, specifically in mammograms and radiographs of mastectomy specimens, including tumor, near-tumor, and far from the tumor parenchymal patterns. This study investigated a mammography-based field effect and the radiomic features of specimen radiographs and their relation to mammographic features. Results found radiomic features in the subcategories of intensity-based, edge frequency, and Fourier-based features from ROIs closer to the tumor tended to show more similarity to the tumor than features from ROIs far from the tumor within mammograms. In corresponding specimen radiographs, intensity-based, edge frequency, and GLCM radiomic features followed a similar trend. Integration of mammogram and specimen radiograph radiomic features showed strong relationships of intensity-based features across the parenchymal field in in- and ex-vivo imaging.

To better understand the potential clinical utility of mammography-based field effect for risk assessment, it is necessary to examine characteristics of the breast tissue in sequential imaging before cancer development to identify any pre-diagnosis changes in imaging features. Chapter 4 evaluated two methods for predicting future cancer risk: a single time point SVM classifier and a temporal LSTM classifier that integrated multiple time points. Results showed that the single time point SVM classifier, which used only the first prior image, did not perform significantly better than random guessing. However, the temporal sequence LSTM classifiers showed statistically significant performance better than guessing and were comparable to other risk-based models in the literature. Further, the similarity in classifier performance across both lateralities further supported the detectability of a cancer field effect in mammography imaging.

This first portion of research focused on the application of breast cancer risk assessment, and a few limitations were identified. While the characterization of the potential cancer field focused on the general location in relation to the tumor, future studies should aim to better define the physical extent of the cancer field for each tumor subtype. Additionally, the relative size of the cohort used in Chapter 3 was limited due to the connection with corresponding specimen radiographs, which are not typically available for many patients. Chapter 4 focused on the performance of commonly used SVM and LSTM classifiers, but many alternative single time point and temporal networks exist that could be investigated in the future. For both chapters investigating breast cancer risk, the data used was only from a single institution and utilized only the cranio-caudal mammogram. Future work could expand similar analysis to other institutions and other mammographic views and tomosynthesis derived images. Lastly, alternative neural networks for generating deep-learning features, other than VGG19, could be investigated.

To ensure that AI models are reliable and applicable to a broader population, the data used to evaluate them must be independent from the data used to train them, and both sets of data must be representative of the population. A possible solution to this common challenge is to create a centralized database with representative data from multiple institutions that can be used to assess algorithm performance. In Chapter 5, a novel sequestration method was presented that uses multi-dimensional stratified sampling to effectively separate incoming data batches into two subsets. The performance of this method was evaluated in terms of similarities between resulting subsets of an example database. Additionally, a novel task-based distribution sampling method was discussed, which draws from a data commons to create a sample that matches a specified demographic distribution. This method was demonstrated on an example cohort of patients.

For the developed methodology and evaluations described in Chapter 5, a few key limitations were noted. The size and initial demographic distribution of the input dataset to both sequestration and task-based sampling methods may present a limitation, as an input dataset that does not contain sufficient patients may result in poor performance of the described methods. The size of a sufficient database is dependent on the goals of each task. For stratified sampling to achieve a higher degree of balance than simple randomization, each unique combination of demographic variables must be populated enough to be separated in proportions of 80:20. For task-based sampling, the ability of the optimization algorithm to achieve a specified level of similarity to a target demographic distribution is inherently limited by the availability of patients that match the target distribution in the initial cohort.

Future directions for this work may include evaluating alternative metrics of balance in the optimization algorithm developed for task-based sampling. The current algorithm utilizes a maximum deviance from the target distribution, but average deviance or squared deviance from the target may prove to be advantageous in certain applications. Lastly, the methodology developed in this work is intended to be made publicly available and used for selecting data to evaluate algorithm performance before submission for regulatory clearance. To ensure consistent performance, the algorithms and workflow processes will need to be regularly evaluated for compatibility with current software functionality.

This work has the potential improve patient care through the use of AI in breast cancer risk assessment and COVID-19. To characterize a mammographically detectable breast cancer field effect with the goal of improving risk stratification in general public, radiomics and deep learning features were evaluated across the parenchymal field, and a risk model utilizing temporal mammography data prior to diagnosis was developed. To create robust tools for algorithm performance assessment and acceleration of clinical translation, a methodology for the initiation of a sequestered data commons and task-based sampling methods of a public data commons were described. Overall, this work demonstrated advancements in AI-based evaluation, classification, and methodology with applications in breast cancer risk assessment and COVID-19.

REFERENCES

1. Winsberg, F.; Elkin, M.; Macy, J.; Bordaz, V.; Weymouth, W. Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis. *Radiology* **1967**, *89*, 211–215, doi:10.1148/89.2.211.

2. Chan, H.P.; Doi, K.; Galhotra, S.; Vyborny, C.J.; MacMahon, H.; Jokich, P.M. Image Feature Analysis and Computer-Aided Diagnosis in Digital Radiography. I. Automated Detection of Microcalcifications in Mammography. *Med Phys* **1987**, *14*, 538–548, doi:10.1118/1.596065.

3. Giger, M.L.; Doi, K.; MacMahon, H. Image Feature Analysis and Computer-Aided Diagnosis in Digital Radiography. 3. Automated Detection of Nodules in Peripheral Lung Fields. *Med Phys* **1988**, *15*, 158–166, doi:10.1118/1.596247.

4. Chan, H.P.; Doi, K.; Vyborny, C.J.; Schmidt, R.A.; Metz, C.E.; Lam, K.L.; Ogura, T.; Wu, Y.Z.; MacMahon, H. Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms. The Potential of Computer-Aided Diagnosis. *Invest Radiol* **1990**, *25*, 1102–1110, doi:10.1097/00004424-199010000-00006.

5. Giger, M.L.; Doi, K.; MacMahon, H.; Metz, C.E.; Yin, F.F. Pulmonary Nodules: Computer-Aided Detection in Digital Chest Images. *Radiographics* **1990**, *10*, 41–51, doi:10.1148/radiographics.10.1.2296696.

6. Zhang, W.; Doi, K.; Giger, M.L.; Wu, Y.; Nishikawa, R.M.; Schmidt, R.A. Computerized Detection of Clustered Microcalcifications in Digital Mammograms Using a Shift-Invariant Artificial Neural Network. *Med Phys* **1994**, *21*, 517–524, doi:10.1118/1.597177.

7. Rao, V.M.; Levin, D.C.; Parker, L.; Cavanaugh, B.; Frangos, A.J.; Sunshine, J.H. How Widely Is Computer-Aided Detection Used in Screening and Diagnostic Mammography? *J Am Coll Radiol* **2010**, *7*, 802–805, doi:10.1016/j.jacr.2010.05.019.

8. Giger, M.L.; Vyborny, C.J.; Schmidt, R.A. Computerized Characterization of Mammographic Masses: Analysis of Spiculation. *Cancer Letters* **1994**, *77*, 201–211, doi:10.1016/0304-3835(94)90103-1.

9. Huo, Z.; Giger, M.L.; Vyborny, C.J.; Bick, U.; Lu, P.; Wolverton, D.E.; Schmidt, R.A. Analysis of Spiculation in the Computerized Classification of Mammographic Masses. *Med Phys* **1995**, *22*, 1569–1579, doi:10.1118/1.597626.

10. Jiang, Y.; Nishikawa, R.M.; Wolverton, D.E.; Metz, C.E.; Giger, M.L.; Schmidt, R.A.; Vyborny, C.J.; Doi, K. Malignant and Benign Clustered Microcalcifications: Automated Feature Analysis and Classification. *Radiology* **1996**, *198*, 671–678, doi:10.1148/radiology.198.3.8628853.

11. Gilhuijs, K.G.; Giger, M.L.; Bick, U. Computerized Analysis of Breast Lesions in Three Dimensions Using Dynamic Magnetic-Resonance Imaging. *Med Phys* **1998**, *25*, 1647–1654, doi:10.1118/1.598345.

12. Chen, W.; Giger, M.L.; Bick, U.; Newstead, G.M. Automatic Identification and Classification of Characteristic Kinetic Curves of Breast Lesions on DCE-MRI. *Med Phys* **2006**, *33*, 2878–2887, doi:10.1118/1.2210568.

13. Chen, W.; Giger, M.L.; Li, H.; Bick, U.; Newstead, G.M. Volumetric Texture Analysis of Breast Lesions on Contrast-Enhanced Magnetic Resonance Images. *Magn Reson Med* **2007**, *58*, 562–571, doi:10.1002/mrm.21347.

14. Giger, M.L.; Doi, K.; MacMahon, H.; Nishikawa, R.M.; Hoffmann, K.R.; Vyborny, C.J.; Schmidt, R.A.; Jia, H.; Abe, K.; Chen, X. An "Intelligent" Workstation for Computer-Aided Diagnosis. *RadioGraphics* **1993**, *13*, 647–656, doi:10.1148/radiographics.13.3.8316671.

15. Horsch, K.; Giger, M.L.; Vyborny, C.J.; Lan, L.; Mendelson, E.B.; Hendrick, R.E. Classification of Breast Lesions with Multimodality Computer-Aided Diagnosis: Observer Study Results on an Independent Clinical Data Set. *Radiology* **2006**, *240*, 357–368, doi:10.1148/radiol.2401050208.

16. Shimauchi, A.; Giger, M.L.; Bhooshan, N.; Lan, L.; Pesce, L.L.; Lee, J.K.; Abe, H.; Newstead, G.M. Evaluation of Clinical Breast MR Imaging Performed with Prototype Computer-Aided Diagnosis Breast MR Imaging Workstation: Reader Study. *Radiology* **2011**, *258*, 696–704, doi:10.1148/radiol.10100409.

17. Huo, Z.; Giger, M.L.; Olopade, O.I.; Wolverton, D.E.; Weber, B.L.; Metz, C.E.; Zhong, W.; Cummings, S.A. Computerized Analysis of Digitized Mammograms of BRCA1 and BRCA2 Gene Mutation Carriers. *Radiology* **2002**, *225*, 519–526, doi:10.1148/radiol.2252010845.

18. Sahiner, B.; Chan, H.-P.; Petrick, N.; Wei, D.; Helvie, M.A.; Adler, D.D.; Goodsitt, M.M. Classification of Mass and Normal Breast Tissue: A Convolution Neural Network Classifier with Spatial Domain and Texture Images. *IEEE Transactions on Medical Imaging* **1996**, *15*, 598–610, doi:10.1109/42.538937.

19. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine Learning for Medical Imaging. *Radiographics* **2017**, *37*, 505–515, doi:10.1148/rg.2017160130.

20. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer Statistics, 2023. *CA: A Cancer Journal for Clinicians* **2023**, *73*, 17–48, doi:10.3322/caac.21763.

21. Lee, C.H.; Dershaw, D.D.; Kopans, D.; Evans, P.; Monsees, B.; Monticciolo, D.; Brenner, R.J.; Bassett, L.; Berg, W.; Feig, S.; et al. Breast Cancer Screening With Imaging: Recommendations From the Society of Breast Imaging and the ACR on the Use of Mammography, Breast MRI, Breast Ultrasound, and Other Technologies for the Detection of Clinically Occult Breast Cancer. *Journal of the American College of Radiology* **2010**, *7*, 18–27, doi:10.1016/j.jacr.2009.09.022.

22. World Health Organization *WHO Position Paper on Mammography Screening*; World Health Organization: Geneva, 2014; ISBN 978-92-4-150793-6.

23. Oeffinger, K.C.; Fontham, E.T.H.; Etzioni, R.; Herzig, A.; Michaelson, J.S.; Shih, Y.-C.T.; Walter, L.C.; Church, T.R.; Flowers, C.R.; LaMonte, S.J.; et al. Breast Cancer Screening for Women at Average Risk. *JAMA* **2015**, *314*, 1599, doi:10.1001/jama.2015.12783.

24. Sprague, B.L.; Gangnon, R.E.; Burt, V.; Trentham-Dietz, A.; Hampton, J.M.; Wellman, R.D.; Kerlikowske, K.; Miglioretti, D.L. Prevalence of Mammographically Dense Breasts in the United States. *JNCI: Journal of the National Cancer Institute* **2014**, *106*, doi:10.1093/jnci/dju255.

25. Giger, M.L. Machine Learning in Medical Imaging. *Journal of the American College of Radiology* **2018**, *15*, 512–520, doi:10.1016/j.jacr.2017.12.028.

26. Li, H.; Giger, M.L. Breast Cancer. In *Radiomics and Radiogenomics*; Chapman and Hall/CRC, 2019; pp. 229–249 ISBN 978-1-351-20827-7.

27. Li, H.; Mendel, K.R.; Lan, L.; Sheth, D.; Giger, M.L. Digital Mammography in Breast Cancer: Additive Value of Radiomics of Breast Parenchyma. *Radiology* **2019**, *291*, 15–20, doi:10.1148/radiol.2019181113.

28. Gastounioti, A.; Conant, E.F.; Kontos, D. Beyond Breast Density: A Review on the Advancing Role of Parenchymal Texture Analysis in Breast Cancer Risk Assessment. *Breast Cancer Research* **2016**, *18*, 91, doi:10.1186/s13058-016-0755-8.

29. Tan, M.; Mariapun, S.; Yip, C.H.; Ng, K.H.; Teo, S.-H. A Novel Method of Determining Breast Cancer Risk Using Parenchymal Textural Analysis of Mammography Images on an Asian Cohort. *Physics in Medicine & Biology* **2019**, *64*, 035016, doi:10.1088/1361-6560/aafabd.

30. Li, H.; Giger, M.L.; Huynh, B.Q.; Antropova, N.O. Deep Learning in Breast Cancer Risk Assessment: Evaluation of Convolutional Neural Networks on a Clinical Dataset of Full-Field Digital Mammograms. *Journal of Medical Imaging* **2017**, *4*, 1, doi:10.1117/1.jmi.4.4.041304.

31. Gastounioti, A.; Desai, S.; Ahluwalia, V.S.; Conant, E.F.; Kontos, D. Artificial Intelligence in Mammographic Phenotyping of Breast Cancer Risk: A Narrative Review. *Breast Cancer Research* **2022**, *24*, 1–12, doi:10.1186/S13058-022-01509-Z/FIGURES/4.

32. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in Health and Medicine. *Nat Med* **2022**, *28*, 31–38, doi:10.1038/s41591-021-01614-0.

33. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial Intelligence in Radiology. *Nat Rev Cancer* **2018**, *18*, 500–510, doi:10.1038/s41568-018-0016-5.

34. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Kai Li; Li Fei-Fei ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; IEEE, June 1 2009; pp. 248–255.

35. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in ML-Based Science 2022.

36. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; 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* **2021**, *3*, 199–217, doi:10.1038/s42256-021-00307-0.

37. Wynants, L.; Van Calster, B.; Collins, G.S.; Riley, R.D.; Heinze, G.; Schuit, E.; Albu, E.; Arshi, B.; Bellou, V.; Bonten, M.M.J.; et al. Prediction Models for Diagnosis and Prognosis of COVID-19: Systematic Review and Critical Appraisal. *BMJ* **2020**, m1328, doi:10.1136/bmj.m1328.

38. Alsharif, W.; Qurashi, A. Effectiveness of COVID-19 Diagnosis and Management Tools: A Review. *Radiography* **2021**, *27*, 682–687, doi:10.1016/j.radi.2020.09.010.

39. Islam, N.; Ebrahimzadeh, S.; Salameh, J.-P.; Kazi, S.; Fabiano, N.; Treanor, L.; Absi, M.; Hallgrimson, Z.; Leeflang, M.M.; Hooft, L.; et al. Thoracic Imaging Tests for the Diagnosis of COVID-19. *Cochrane Database Syst Rev* **2021**, *3*, CD013639, doi:10.1002/14651858.CD013639.pub4.

40. Mbunge, E.; Akinnuwesi, B.; Fashoto, S.G.; Metfula, A.S.; Mashwama, P. A Critical Review of Emerging Technologies for Tackling COVID-19 Pandemic. *Human Behavior and Emerging Technologies* **2021**, *3*, 25–39, doi:10.1002/hbe2.237.

41. Mckinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International Evaluation of an AI System for Breast Cancer Screening. *Nature* **2020**, *577*, 89, doi:10.1038/s41586-019-1799-6.

42. El Naqa, I.; Haider, M.A.; Giger, M.L.; Ten Haken, R.K. Artificial Intelligence: Reshaping the Practice of Radiological Sciences in the 21st Century. *British Journal of Radiology* **2020**, *93*, doi:10.1259/bjr.20190855.

43. Sahiner, B.; Pezeshk, A.; Hadjiiski, L.M.; Wang, X.; Drukker, K.; Cha, K.H.; Summers, R.M.; Giger, M.L. Deep Learning in Medical Imaging and Radiation Therapy. *Medical Physics* **2019**, *46*, e1–e36, doi:10.1002/mp.13264.

44. Ross, B.; Gambhir, S.S. *Molecular Imaging: Principles and Practice*; Academic Press, 2016; Vol. 4; ISBN 978-0-12-816386-3.

45. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016;

46. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37–52, doi:10.1016/0169-7439(87)80084-9.

47. Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, doi:10.1007/s10479-011-0841-3.

48. Bi, W.L.; Hosny, A.; Schabath, M.B.; Giger, M.L.; Birkbak, N.J.; Mehrtash, A.; Allison, T.; Arnaout, O.; Abbosh, C.; Dunn, I.F.; et al. Artificial Intelligence in Cancer Imaging: Clinical Challenges and Applications. *CA: A Cancer Journal for Clinicians* **2019**, *0*, 1–31, doi:10.3322/caac.21552.

49. Sheth, D.; Giger, M.L. Artificial Intelligence in the Interpretation of Breast Cancer on MRI. *Journal of Magnetic Resonance Imaging* **2020**, *51*, 1310–1324, doi:10.1002/jmri.26878.

50. Masud, R.; Al-Rei, M.; Lokker, C. Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review. *JMIR Med Inform* **2019**, *7*, e12660, doi:10.2196/12660.

51. Recht, M.; Bryan, R.N. Artificial Intelligence: Threat or Boon to Radiologists? *J Am Coll Radiol* **2017**, *14*, 1476–1480, doi:10.1016/j.jacr.2017.07.007.

52. Wolfe, J.N. Breast Patterns as an Index of Risk for Developing Breast Cancer. *AJR Am J Roentgenol* **1976**, *126*, 1130–1137, doi:10.2214/ajr.126.6.1130.

53. Warner, E.; Lockwood, G.; Tritchler, D.; Boyd, N.F. The Risk of Breast Cancer Associated with Mammographic Parenchymal Patterns: A Meta-Analysis of the Published Literature to Examine the Effect of Method of Classification. *Cancer Detect Prev* **1992**, *16*, 67–72.

54. Giger, M.L.; Chan, H.P.; Boone, J. Anniversary Paper: History and Status of CAD and Quantitative Image Analysis: The Role of Medical Physics and AAPM. *Medical Physics* **2008**, *35*, 5799–5820, doi:10.1118/1.3013555.

55. Huo, Z.; Giger, M.L.; Wolverton, D.E.; Zhong, W.; Cumming, S.; Olopade, O.I. Computerized Analysis of Mammographic Parenchymal Patterns for Breast Cancer Risk Assessment: Feature Selection. *Medical Physics* **2000**, *27*, 4–12, doi:10.1118/1.598851.

56. Li, H.; Giger, M.L.; Olopade, O.I.; Lan, L. Fractal Analysis of Mammographic Parenchymal Patterns in Breast Cancer Risk Assessment. *Academic Radiology* **2007**, *14*, 513–521, doi:10.1016/j.acra.2007.02.003.

57. Li, H.; Giger, M.L.; Olopade, O.I.; Chinander, M.R. Power Spectral Analysis of Mammographic Parenchymal Patterns for Breast Cancer Risk Assessment. *Journal of Digital Imaging* **2008**, *21*, 145–152, doi:10.1007/s10278-007-9093-9.

58. Foy, J.J.; Robinson, K.R.; Li, H.; Giger, M.L.; Al-Hallaq, H.; Armato, S.G. Variation in Algorithm Implementation across Radiomics Software. *JMI* **2018**, *5*, 044505, doi:10.1117/1.JMI.5.4.044505.

59. McNitt-Gray, M.; Napel, S.; Jaggi, A.; Mattonen, S.A.; Hadjiiski, L.; Muzi, M.; Goldgof, D.; Balagurunathan, Y.; Pierce, L.A.; Kinahan, P.E.; et al. Standardization in Quantitative Imaging: A Multicenter Comparison of Radiomic Features from Different Software Packages on Digital Reference Objects and Patient Data Sets. *Tomography* **2020**, *6*, 118–128, doi:10.18383/j.tom.2019.00031.

60. Choy, G.; Khalilzadeh, O.; Michalski, M.; Do, S.; Samir, A.E.; Pianykh, O.S.; Geis, J.R.; Pandharipande, P.V.; Brink, J.A.; Dreyer, K.J. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology* **2018**, *288*, 318–328, doi:10.1148/radiol.2018171820.

61. Huynh, B.Q.; Li, H.; Giger, M.L. Digital Mammographic Tumor Classification Using Transfer Learning from Deep Convolutional Neural Networks. *Journal of Medical Imaging* **2016**, *3*, 034501, doi:10.1117/1.JMI.3.3.034501.

62. Antropova, N.; Huynh, B.Q.; Giger, M.L. A Deep Feature Fusion Methodology for Breast Cancer Diagnosis Demonstrated on Three Imaging Modality Datasets. *Medical Physics* **2017**, *44*, 5162–5171, doi:10.1002/MP.12453.

63. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **2018**, *3*, 861, doi:10.21105/joss.00861.

64. Jamieson, A.R.; Giger, M.L.; Drukker, K.; Pesce, L.L. Enhancement of Breast CADx with Unlabeled Data. *Medical Physics* **2010**, *37*, 4155–4172, doi:10.1118/1.3455704.

65. Jamieson, A.R.; Giger, M.L.; Drukker, K.; Li, H.; Yuan, Y.; Bhooshan, N. Exploring Nonlinear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t -SNE. *Medical Physics* **2010**, *37*, 339–351, doi:10.1118/1.3267037.

66. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation* **2019**, *31*, 1235–1270, doi:10.1162/neco_a_01199.

67. Zaremba, W.; Sutskever, I.; Vinyals, O.; Brain, G. Recurrent Neural Network Regularization. **2014**, doi:10.48550/arxiv.1409.2329.

68. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780, doi:10.1162/neco.1997.9.8.1735.

69. Giger, M.L. Future Perspectives: CAD to Quantitative Image Biomarkers, Phenotypes, and Imaging Genomics. In *Computer-Aided Detection and Diagnosis in Medical Imaging*; CRC Press, 2015 ISBN 978-0-429-19362-0.

70. Whitney, H.M.; Giger, M.L. Artificial Intelligence in Medical Imaging. *Quantitative Imaging in Medicine: Background and Basics* **2021**, *2094*, doi:10.1088/1742-6596/2094/3/032008.

71. Fuhrman, J.D.; Gorre, N.; Hu, Q.; Li, H.; El Naqa, I.; Giger, M.L. A Review of Explainable and Interpretable AI with Applications in COVID-19 Imaging. *Med Phys* **2022**, *49*, 1–14, doi:10.1002/mp.15359.

72. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization.

73. Crosby, J.; Chen, S.; Li, F.; MacMahon, H.; Giger, M. Network Output Visualization to Uncover Limitations of Deep Learning Detection of Pneumothorax. In Proceedings of the Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment; SPIE, March 16 2020; Vol. 11316, pp. 125–128.

74. Whitney, H.M.; Li, H.; Ji, Y.; Liu, P.; Giger, M.L. Harmonization of Radiomic Features of Breast Lesions across International DCE-MRI Datasets. *Journal of Medical Imaging* **2020**, *7*, 1, doi:10.1117/1.jmi.7.1.012707.

75. Whitney, H.M.; Drukker, K.; Abe, H.; Giger, M.L. Case-Based Repeatability and Operating Point Variability of AI: Breast Lesion Classification Based on Deep Transfer Learning. In Proceedings of the Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment; SPIE, April 4 2022; Vol. 12035, pp. 223–227.

76. Amstutz, P.; Drukker, K.; Li, H.; Abe, H.; Giger, M.L.; Whitney, H.M. Case-Based Diagnostic Classification Repeatability Using Radiomic Features Extracted from Full-Field Digital Mammography Images of Breast Lesions. In Proceedings of the Medical Imaging 2021: Computer-Aided Diagnosis; SPIE, February 15 2021; Vol. 11597, pp. 217–222.

77. El Naqa, I.; Li, H.; Fuhrman, J.; Hu, Q.; Gorre, N.; Chen, W.; Giger, M.L. Lessons Learned in Transitioning to AI in the Medical Imaging of COVID-19. *J Med Imaging (Bellingham)* **2021**, *8*, 010902, doi:10.1117/1.JMI.8.S1.010902.

78. Medical Imaging and Data Resource Center (MIDRC) Available online: https://www.midrc.org (accessed on 7 February 2023).

79. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging* **2013**, *26*, 1045–1057, doi:10.1007/s10278-013-9622-7.

80. Mainiero, M.B.; Moy, L.; Baron, P.; Didwania, A.D.; diFlorio, R.M.; Green, E.D.; Heller, S.L.; Holbrook, A.I.; Lee, S.-J.; Lewin, A.A.; et al. ACR Appropriateness Criteria Breast Cancer Screening. *Journal of the American College of Radiology* **2017**, *14*, S383–S390, doi:10.1016/j.jacr.2017.08.044.

81. Bista, R.K.; Wang, P.; Bhargava, R.; Uttam, S.; Hartman, D.J.; Brand, R.E.; Liu, Y. Nuclear Nano-Morphology Markers of Histologically Normal Cells Detect the "Field Effect" of Breast Cancer. *Breast Cancer Research and Treatment* **2012**, *135*, 115–124, doi:10.1007/S10549-012-2125-2/FIGURES/6.

82. Braakhuis, B.J.M.; Tabor, M.P.; Kummer, J.A.; Leemans, C.R.; Brakenhoff, R.H. A Genetic Explanation of Slaughter's Concept of Field Cancerization: Evidence and Clinical Implications. *Cancer research* **2003**, *63*, 1727–1730.

83. Heaphy, C.M.; Griffith, J.K.; Bisoffi, M. Mammary Field Cancerization: Molecular Evidence and Clinical Importance. *Breast Cancer Research and Treatment* **2009**, *118*, 229–239, doi:10.1007/s10549-009-0504-0.

84. Sen, P.K. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association* **1968**, *63*, 1379, doi:10.2307/2285891.

85. Sheskin, D.J. *Parametric and Nonparametric Statistical Procedures*; 2nd ed.; CRC Press, 2000; ISBN 1-58488-133-X.

86. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, *57*, 289–300.

87. Reiner, A.; Yekutieli, D.; Benjamini, Y. Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures. *Bioinformatics* **2003**, *19*, 368–375, doi:10.1093/bioinformatics/btf877.

88. Sickles, E.; D'Orsi, C.; Bassett, L.; Appleton, C.M.; Berg, W.; Burnside, E.S.; Feig, S.; Gavenonis, S.; Newell, M.S.; Trinh, M. Breast Imaging Reporting and Data System - Mammography. *American College of Radiology* **2013**, doi:10.1016/S0033-8389(01)00017-3.

89. Li, H.; Giger, M.L.; Huo, Z.; Olopade, O.I.; Lan, L.; Weber, B.L.; Bonta, I. Computerized Analysis of Mammographic Parenchymal Patterns for Assessing Breast Cancer Risk: Effect of ROI Size and Location. *Medical Physics* **2004**, *31*, 549–555, doi:10.1118/1.1644514.

90. Yankaskas, B.C.; May, R.C.; Matuszewski, J.; Bowling, J.M.; Jarman, M.P.; Schroeder, B.F. Effect of Observing Change from Comparison Mammograms on Performance of Screening Mammography in a Large Community-Based Population. *Radiology* **2011**, *261*, 762–770, doi:10.1148/radiol.11110653.

91. Hayward, J.H.; Ray, K.M.; Wisner, D.J.; Kornak, J.; Lin, W.; Joe, B.N.; Sickles, E.A. Improving Screening Mammography Outcomes Through Comparison With Multiple Prior Mammograms. *American Journal of Roentgenology* **2016**, *207*, 918–924, doi:10.2214/AJR.15.15917.

92. Gail, M.H.; Brinton, L.A.; Byar, D.P.; Corle, D.K.; Green, S.B.; Schairer, C.; Mulvihill, J.J. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *JNCI: Journal of the National Cancer Institute* **1989**, *81*, 1879–1886, doi:10.1093/jnci/81.24.1879.

93. Tyrer, J.; Duffy, S.W.; Cuzick, J. A Breast Cancer Prediction Model Incorporating Familial and Personal Risk Factors. *Stat Med* **2004**, *23*, 1111–1130, doi:10.1002/sim.1668.

94. Baughan, N.; Douglas, L.; Giger, M.L. Past, Present, and Future of Machine Learning and Artificial Intelligence for Breast Cancer Screening. *Journal of Breast Imaging* **2022**, *2022*, 1–9, doi:10.1093/jbi/wbac052.

95. Tan, M.; Zheng, B.; Leader, J.K.; Gur, D. Association Between Changes in Mammographic Image Features and Risk for Near-Term Breast Cancer Development. *IEEE Transactions on Medical Imaging* **2016**, *35*, 1719–1728, doi:10.1109/TMI.2016.2527619.

96. Antropova, N.; Huynh, B.; Li, H.; Giger, M.L. Breast Lesion Classification Based on Dynamic Contrast-Enhanced Magnetic Resonance Images Sequences with Long Short-Term Memory Networks. *Journal of Medical Imaging* **2018**, *6*, 1, doi:10.1117/1.JMI.6.1.011002.

97. Gierach, G.L.; Li, H.; Loud, J.T.; Greene, M.H.; Chow, C.K.; Lan, L.; Prindiville, S.A.; Eng-Wong, J.; Soballe, P.W.; Giambartolomei, C.; et al. Relationships between Computer-Extracted Mammographic Texture Pattern Features and BRCA1/2 Mutation Status: A Cross-Sectional Study. *Breast Cancer Research* **2014**, *16*, doi:10.1186/s13058-014-0424-8.

98. Chollet, F. Keras 2015.

99. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

100. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* **1999**, *9*, 293–300, doi:10.1023/A:1018628609742.

101. Lopez-del Rio, A.; Martin, M.; Perera-Lluna, A.; Saidi, R. Effect of Sequence Padding on the Performance of Deep Learning Models in Archaeal Protein Functional Prediction. *Sci Rep* **2020**, *10*, 14634, doi:10.1038/s41598-020-71450-8.

102. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization 2017.

103. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **1979**, *6*, 65–70.

104. Yala, A.; Lehman, C.; Schuster, T.; Portnoi, T.; Barzilay, R. A Deep Learning Mammography-Based Model for Improved Breast Cancer Risk Prediction. *Radiology* **2019**, *292*, 60–66, doi:10.1148/radiol.2019182716.

105. Dembrower, K.; Liu, Y.; Azizpour, H.; Eklund, M.; Smith, K.; Lindholm, P.; Strand, F. Comparison of a Deep Learning Risk Score and Standard Mammographic Density Score for Breast Cancer Risk Prediction. *Radiology* **2020**, *294*, 265–272, doi:10.1148/radiol.2019190872.

106. Gichoya, J.W.; Banerjee, I.; Bhimireddy, A.R.; Burns, J.L.; Celi, L.A.; Chen, L.-C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; et al. AI Recognition of Patient Race in Medical Imaging: A Modelling Study. *The Lancet Digital Health* **2022**, *4*, e406–e414, doi:10.1016/S2589-7500(22)00063-2/ATTACHMENT/77F04D9F-1929-404F-B593-094817C7C39B/MMC1.PDF.

107. Johnson, K.S.; Conant, E.F.; Soo, M.S. Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *Journal of Breast Imaging* **2021**, *3*, 12–24, doi:10.1093/jbi/wbaa110.

108. Cucinotta, D.; Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Biomed* **2020**, *91*, 157–160, doi:10.23750/abm.v91i1.9397.

109. Hu, Q.; Drukker, K.; Giger, M.L. Role of Standard and Soft Tissue Chest Radiography Images in Deep-Learning-Based Early Diagnosis of COVID-19. *J Med Imaging (Bellingham)* **2021**, *8*, 014503, doi:10.1117/1.JMI.8.S1.014503.

110. Shi, F.; Wang, J.; Shi, J.; Wu, Z.; Wang, Q.; Tang, Z.; He, K.; Shi, Y.; Shen, D. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. *IEEE Rev Biomed Eng* **2021**, *14*, 4–15, doi:10.1109/RBME.2020.2987975.

111. Larrazabal, A.J.; Nieto, N.; Peterson, V.; Milone, D.H.; Ferrante, E. Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis. *Proceedings of the National Academy of Sciences* **2020**, *117*, 12592–12594, doi:10.1073/PNAS.1919012117.

112. Lin, Y.; Zhu, M.; Su, Z. The Pursuit of Balance: An Overview of Covariate-Adaptive Randomization Techniques in Clinical Trials. *Contemporary Clinical Trials* **2015**, *45*, 21–25, doi:10.1016/j.cct.2015.07.011.

113. Sverdlov, O. *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*; 2016; Vol. 84; ISBN 978-1-4822-3989-8.

114. Centers for Disease Control and Prevention COVID-19 Death Data and Resources: Weekly Updates by Select Demographic and Geographic Characteristics Available online: https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm#SexAndAge.

115. NIH Consideration of Sex as a Biological Variable in NIH-Funded Research.

116. NIH Racial and Ethnic Categories and Definitions for NIH Diversity Programs and for Other Reporting Purposes. **2015**.

117. Budka, M.; Gabrys, B. Density-Preserving Sampling: Robust and Efficient Alternative to Cross-Validation for Error Estimation. *IEEE Transactions on Neural Networks and Learning Systems* **2013**, *24*, 22–34, doi:10.1109/TNNLS.2012.2222925.

118. Ahn, S.; Park, S.H.; Lee, K.H. How to Demonstrate Similarity by Using Noninferiority and Equivalence Statistical Testing in Radiology Research. *Radiology* **2013**, *267*, 328–338, doi:10.1148/radiol.12120725.

APPENDIX A: PRELIMINARY USE OF THE KOLMOGOROV-SMIRNOV TEST IN ASSESSING FEATURE STATISTICAL SIMILARITY ACROSS THE PARENCHYMAL FIELD

In Chapter 3, analysis of radiomic and deep learning features to characterize a potential cancer field effect focused on the correlation of features between ROI regions and between modalities. However, features were also investigated for statistical similarity across ROI regions using the Kolmogorov-Smirnov (KS) test.

A.1 Methods

The KS test was used to compare human-engineered radiomic and deep learning features across the tumor, near, and far regions of the breast, as defined in Figure 3.1. The KS test statistic is defined as the maximum vertical distance between the cumulative distribution functions of the two feature distributions [85]. A KS test statistic equal to zero would represent perfectly equivalent distributions, and larger KS test statistics represent distributions that are more different. If the KS test statistic is greater than the critical value, as defined by Sheskin (2000) for the size of the two subsets, the distributions are considered to be significantly different at the applicable significance threshold [85]. The KS test is known to be robust, as it is not dependent on sample size and does not assume any distribution shape. For the purposes of this study, the KS test was used to test for equivalence, as described by Ahn et al. [118]. After computing a KS test statistic and critical value for each feature comparison, a 95% confidence interval (CI) on the KS test statistic was constructed through bootstrapping, using 2000 bootstrap iterations. Using the calculated KS test statistic and

95% CI, statistical equivalence or noninferiority between the two ROI feature distributions was assessed, using the critical value as the cutoff point, delta.

Only the mammogram features were used in this preliminary evaluation of statistical similarity using the KS test. To compare the shape of feature distributions between ROI regions only, distributions were shifted to align the means of each distribution. To compare both feature distribution shape and absolute feature values, no shifting of distribution means was applied.

A.2 Results

KS statistical equivalence test results for evaluating the shape of feature distributions between ROI regions only (means aligned) are shown in Figure A.1 and Figure A.2 for radiomic and deep learning features, respectively. Results indicated that most feature distributions in all ROI regions were found to be equivalent, with an equivalence threshold equal to the critical value at the p = 0.05 level. Table A.1 shows the percentages of feature comparisons found to be statistically equivalent. 81.8% of human-engineered radiomic features and 90.5% of deep learning extracted feature distributions across all regions reached statistical equivalence. A slightly higher percentage of non-tumor to non-tumor ROI comparisons were found to be equivalent than tumor to non-tumor ROI comparisons for both human-engineered radiomic features and deep learning extracted features. This result seems reasonable, given that tumors have been found to be more dense and coarser in texture than parenchymal tissue, while indicating a high degree of similarity across all tumor and non-tumor tissue feature distributions [88,89].

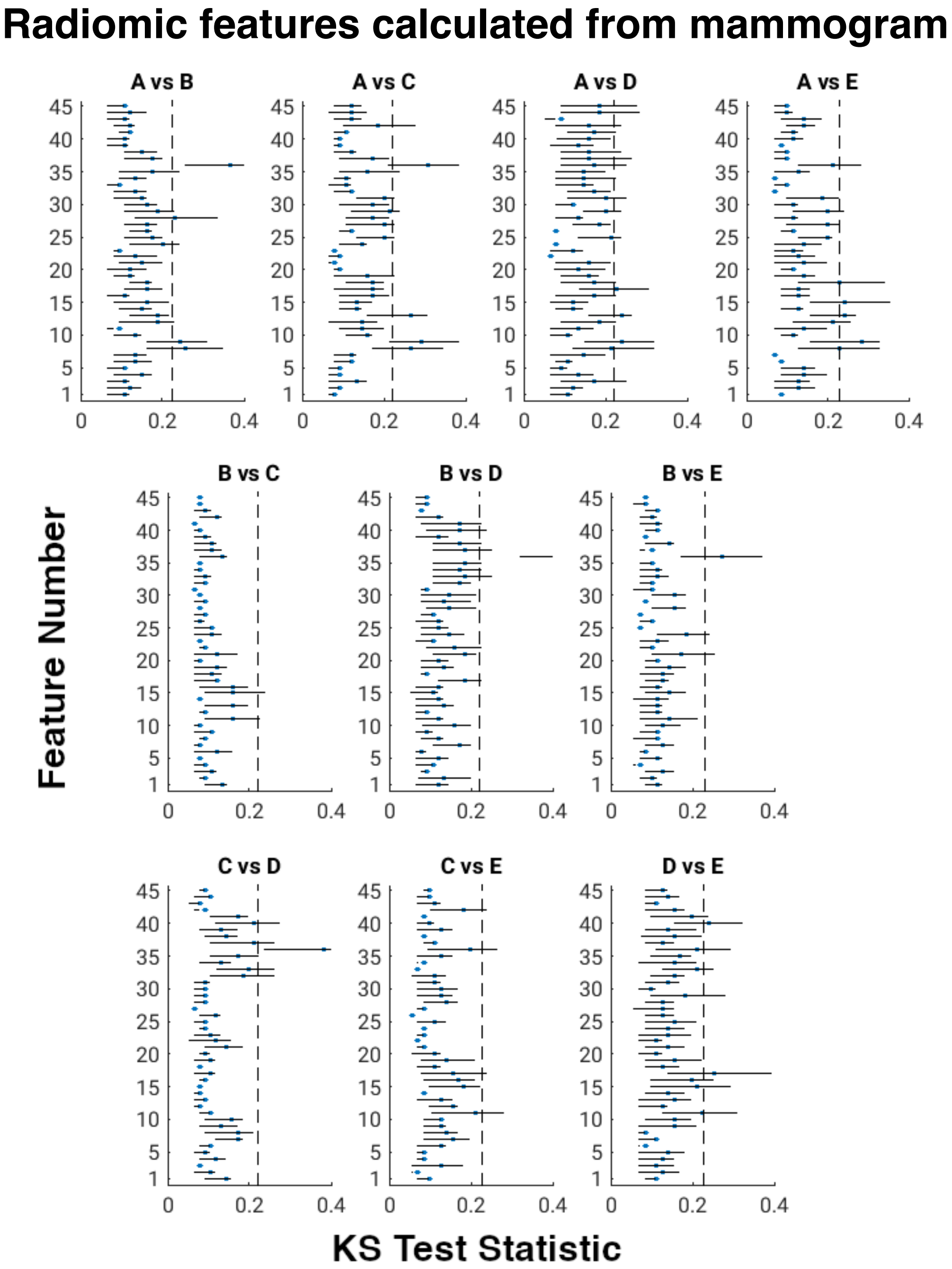


Figure A.1. KS equivalence test results for comparisons of mammogram ROI regions using radiomic features (A – tumor, B – near to the tumor, C and D – far from the tumor, and E – contralateral breast). KS test statistic values (means aligned) are shown with a 95% CI calculated from bootstrapping with 2000 iterations. Dotted line indicates the critical value threshold, below which a KS test statistic would indicate distributions that are statistically equivalent at the p = 0.05 significance threshold.

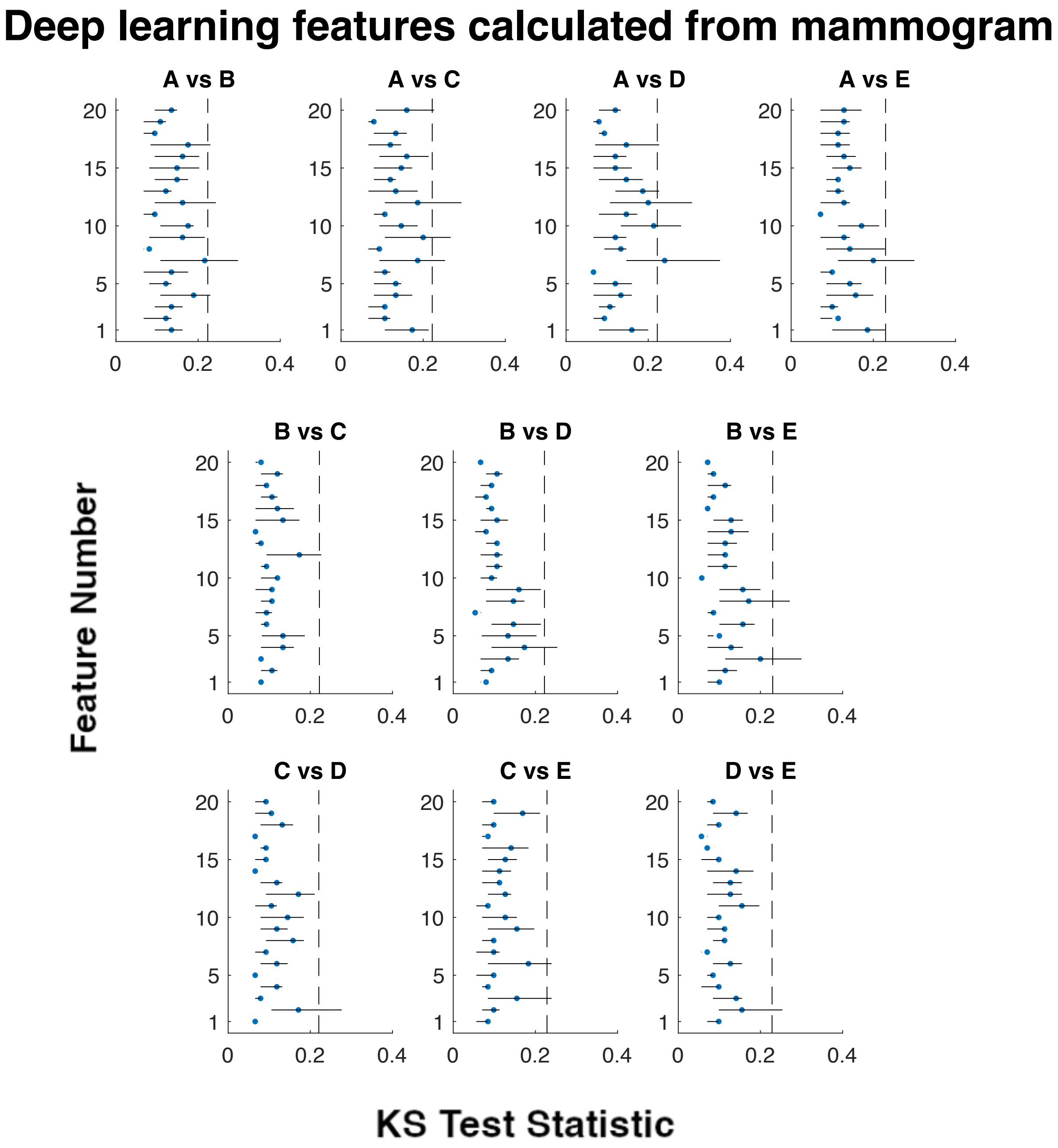


Figure A.2. KS equivalence test results for comparisons of mammogram ROI regions using deep learning features (A – tumor, B – near to the tumor, C and D – far from the tumor, and E – contralateral breast). KS test statistic values (means aligned) are shown with a 95% CI calculated from bootstrapping with 2000 iterations. Dotted line indicates the critical value threshold, below which a KS test statistic would indicate distributions that are statistically equivalent at the p = 0.05 significance threshold.

Table A.1. Percentage (%) of features that reached statistical equivalence for each ROI region comparison. For each, the KS test statistic and 95% CI were less than the equivalence margin, calculated at the p = 0.05 critical value. Regions A, B, C, D, and E are defined in Figure 3.1.

	A vs. B	**A vs. C**	**A vs. D**	**A vs. E**	**B vs. C**	**B vs. D**	**B vs. E**	**C vs. D**	**C vs. E**	**D vs. E**	**All**
Radiomic features	80.0	75.6	57.8	80.0	95.6	75.6	93.3	88.9	91.1	80.0	81.8
Deep learning features	80.0	85.0	80.0	95.0	95.0	95.0	90.0	95.0	95.0	95.0	90.5

Results shown in Figure A.1, Figure A.2, and Table A.1 indicate strong similarity in feature distribution shape between ROI regions. However, similarity in the absolute magnitude of feature distributions in addition to their shape provides important information and should be evaluated as well. To enable comparison of KS test statistics in ROI regions with and without alignment of distribution means, the test statistics were represented in color scale plots. Within the plots, values closer to zero that represented more similar distributions were shown in blue while values closer to one that represented more different distributions were shown in red; values in the middle (closer to 0.5) were represented in white. Colors were scaled linearly with numerical values. The color scale plots for radiomics features with and without alignment of the distribution means are shown in Figure A.3 and Figure A.4, respectively, while the corresponding plots for deep learning features are shown in Figure A.5 and Figure A.6, respectively.

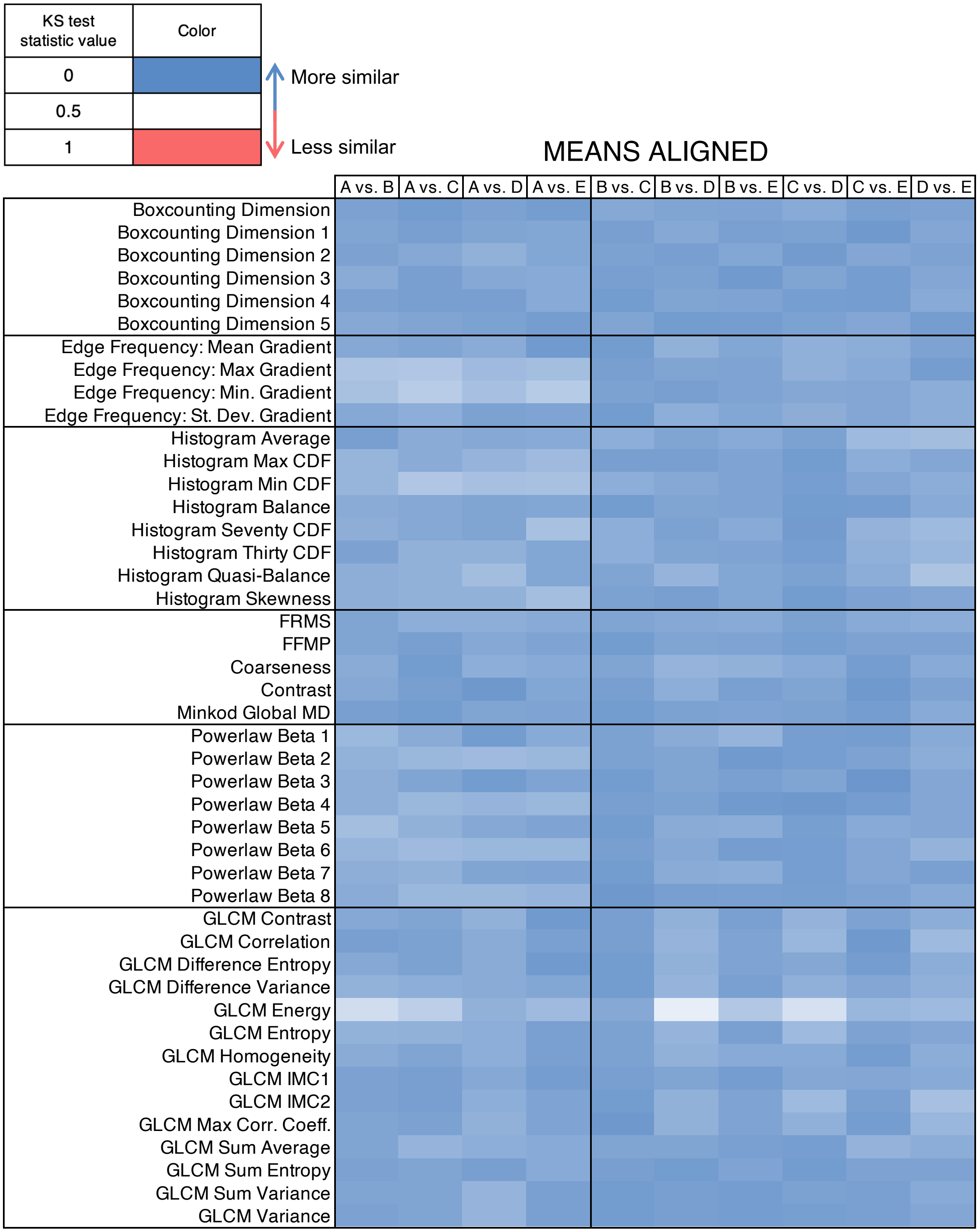


Figure A.3. KS test statistic color scale plot for radiomic features after alignment of feature distribution means. The color of each cell represents the direction and strength of each correlation, as noted in the legend.

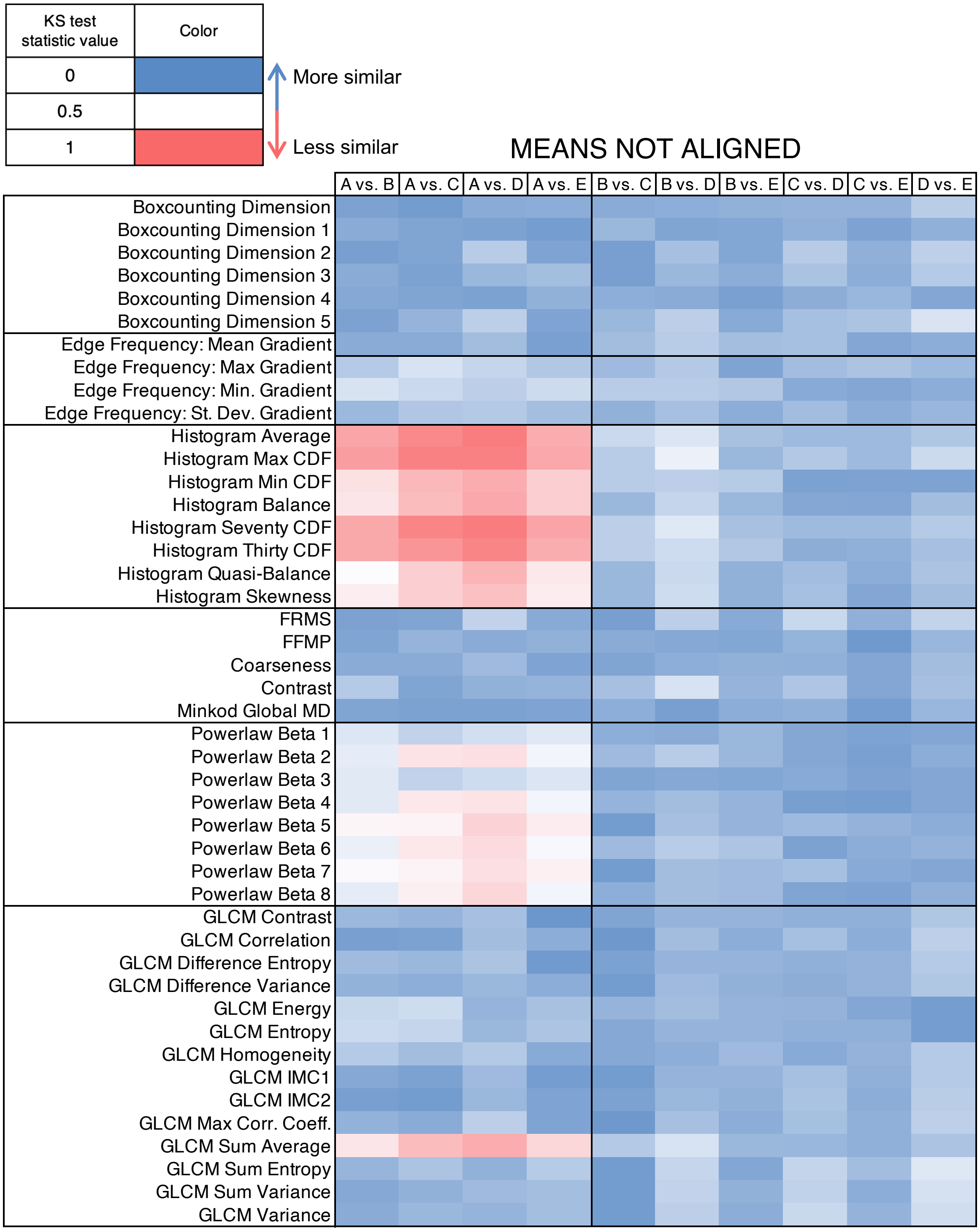


Figure A.4. KS test statistic color scale plot for radiomic features without alignment of distribution means. The color of each cell represents the direction and strength of each correlation, as noted in the legend.

KS test statistic value	Color
0	🟦
0.5	⬜
1	🟥

↑ More similar

↓ Less similar

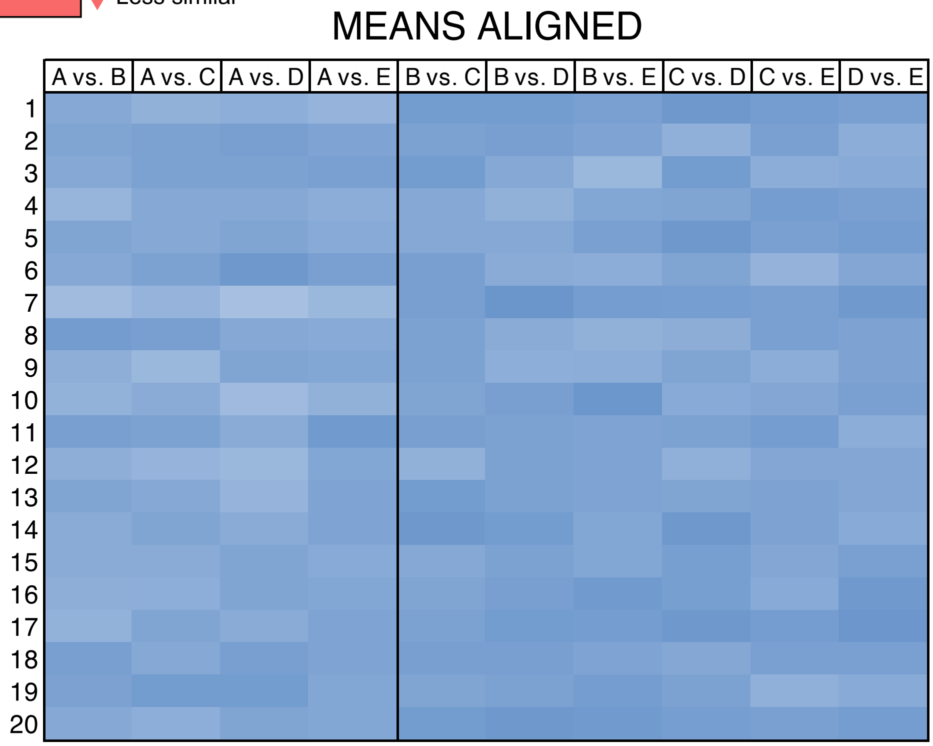


Figure A.5. KS test statistic color scale plot for deep learning features after alignment of feature distribution means. The color of each cell represents the direction and strength of each correlation, as noted in the legend.

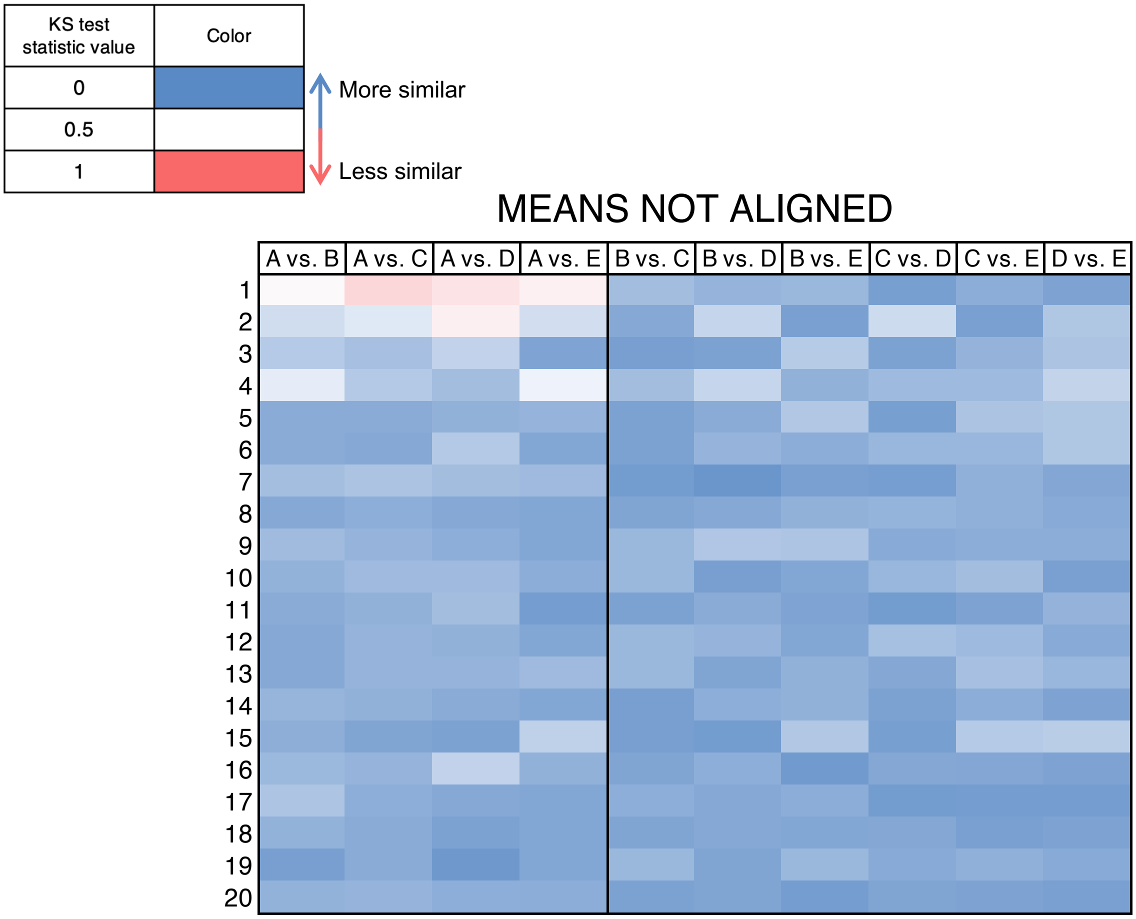


Figure A.6. KS test statistic color scale plot for deep learning features without alignment of distribution means. The color of each cell represents the direction and strength of each correlation, as noted in the legend.

A.3 Discussion

KS test results indicated that most feature distributions in all ROI regions were found to be equivalent, with an equivalence threshold equal to the critical value at the $p = 0.05$ level. 81.8% of radiomic and 90.5% of deep learning feature distributions across all regions showed this statistical equivalence. A slightly higher percentage of non-tumor to non-tumor ROI comparisons were found to be equivalent than tumor to non-tumor ROI comparisons for both radiomic and deep learning features. This result seems reasonable, given that tumors have been found to be more dense and coarser in texture than parenchymal tissue, while indicating a high degree of similarity across all tumor and non-tumor tissue feature distributions.

This broad similarity in feature distribution shapes across all ROI region comparisons may be an important component of a potential mammography-detectable field effect. However, it is also important to investigate other aspects of the data through alternative approaches, as is done in the subsequent color scale figures, which highlight the impact of alignment of the means of each distribution for each comparison.

Alignment of the distribution means only investigated the similarity of distribution shape and removed the effect of the absolute feature value change. As is shown in radiomic feature results, not aligning the distribution means particularly decreased similarity of tumor to non-tumor comparisons for histogram or intensity-based features and power law beta features, indicating differences in the absolute values of feature distributions. This is expected for intensity-based features, as their absolute values correlate to average gray values in the ROI, which are known to differ within a solid mass compared to "normal" parenchyma. For deep learning based features, not aligning the distribution means particularly increased the test statistic values for the first two principal components of tumor to non-tumor comparisons, again indicating differences in the absolute values of feature distributions. Although the underlying characteristics of these pseudo-features cannot be explained as simply as radiomic features, it is known that the first principal components will represent the most fundamental characteristics of the object [46]. Thus, it could be reasonable to infer that the first principal components may also quantify how light or dark the pixels from a given ROI are, describing the tissue intensity as well.

A.4 Conclusions

Given the distinctive subsets of features that were seen to change from “similar” to “non-similar” between Figure A.3 and Figure A.4, and between Figure A.5 and Figure A.6, the remainder of the analysis for mammogram and specimen radiographs was completed using absolute values of features without shifting of distributions. As a result of these distinctive groups of features indicating differences in the absolute values of feature distributions, these relationships were investigated further using the Kendall’s Tau-b rank correlation test and the Pearson correlation test as described in Chapter 3. However, it is important to note that the two results do not contradict one another but highlight different aspects of the feature relationships.

LIST OF PUBLICATIONS AND PRESENTATIONS

Peer-Reviewed Publications

N. Baughan, H. M. Whitney, K. Drukker, B. Sahiner, T. Hu, G. H. Kim, M. McNitt-Gray, K. J. Myers, M. L. Giger, "Sequestration of imaging studies in MIDRC: Stratified sampling to balance demographic characteristics of patients in a multi-institutional data commons." (Submitted)

H. M. Whitney, **N. Baughan**, K. J. Myers, K. Drukker, J. Gichoya, B. Bower, R. Sa, W. Chen, N. Gruszauskas, J. Kalpathy-Cramer, S. Koyejo, B. Sahiner, J. Zhang, and M. L. Giger, "Longitudinal assessment of demographic representativeness in the Medical Imaging and Data Resource Center Open Data Commons." (Under revision)

N. Baughan, H. Li, L. Lan, C. Chan, M. Embury, I. Yim, G. Whitman, R. El-Zein, I. Bedrosian, M. L. Giger, "Radiomic and deep learning characterization of breast parenchyma on FFDMs and specimen radiographs: A pilot study of a potential cancer field effect." (Under revision)

H. Li, K. Robinson, L. Lan, **N. Baughan**, C. Chan, M. Embury, G. Whitman, R. El-Zein, I. Bedrosian, M. L. Giger, "Temporal machine learning analysis of prior mammograms for breast cancer risk prediction," *Cancers,* 2023.

N. Baughan*, L. Douglas*, M. L. Giger, "Past, Present, and Future of Machine Learning and Artificial Intelligence for Breast Cancer Screening," *Journal of Breast Imaging,* 2022. doi:10.1093/jbi/wbac052 Invited. (*Authors contributed equally to this work)

Conference Proceedings

N. Baughan, H. M. Whitney, K. Drukker, B. Sahiner, T. Hu, G. H. Kim, M. McNitt-Gray, K. J. Myers, M. L. Giger, "Evaluation of demographic joint distributions describing 22,993 patients in the Medical Imaging and Data Resource Center (MIDRC) public and sequestered data commons," *Proceedings SPIE: Imaging Informatics for Healthcare, Research, and Applications* 12469: 2023.

K. Drukker, B. Sahiner, T. Hu, G. H. Kim, H. M. Whitney, **N. Baughan**, K. J. Myers, M. L. Giger, M. McNitt-Gray, "Assistance tools for the evaluation of machine learning algorithm performance: the decision tree based tools developed by the Medical Imaging and Data Resource Center (MIDRC) Technology Development Project (TDP) 3c effort," *Proceedings SPIE: Image Perception, Observer Performance, and Technology Assessment* 12467: 2023.

N. Baughan, H. M. Whitney, K. Drukker, B. Sahiner, T. Hu, G. H. Kim, M. McNitt-Gray, K. J. Myers, M. L. Giger, "Sequestration of Imaging Studies in MIDRC: a Multi-Institutional Data Commons," *Proceedings SPIE: Image Perception, Observer Performance, and Technology Assessment* 12035: 2022.

N. Baughan, L. Douglas, M. Ballard, E. S. Lee, A. Edwards, L. Lan, H. Li, M. L. Giger, "Association Between DCE MRI Background Parenchymal Enhancement and Mammographic Texture Features," *Proceedings SPIE: Computer-Aided Diagnosis* 12033: 2022.

E. L. Marshall, D. O. Velarde, **N. Baughan**, N. Reiser, C. Guo, J. P. Cruz, K. Feinstein, I. Reiser, "Task-Specific Evaluation of Clinical Pediatric Fluoroscopy Systems," *Proceedings SPIE: Image Perception, Observer Performance, and Technology Assessment* 12035: 2022.

N. Baughan, H. Li, L. Lan, C. Chan, M. Embury, G. Whitman, R. El-Zein, I. Bedrosian, M. L. Giger, "Parenchymal Field Effect Analysis for Breast Cancer Risk Assessment: Evaluation of FFDM Radiomic Similarity," *Proceedings SPIE: Computer-Aided Diagnosis* 11597: 2021.

J. P. Cruz-Bastida, **N. Baughan**, E. Marshall, K. Blunt, K. Feinstein, I. Reiser, "Towards the Objective Assessment of Fluoroscopy Systems: Development of a Framework to Aid the Design of Tasks and Metrics," *Proceedings SPIE: Image Perception, Observer Performance, and Technology* 11599: 2021.

Abstracts

H. M. Whitney, **N. Baughan**, K. Drukker, K. J. Myers, M. L. Giger, "Longitudinal assessment of multi-institutional data diversity in the Medical Imaging and Data Resource Center (MIDRC)," *RSNA Annual Meeting,* Chicago, IL, November 2022.

N. Baughan, H. Li, L. Lan, M. Embury, G. Whitman, R. El-Zein, I. Bedrosian, M. L. Giger, "Use of Radiomic Texture to Characterize Relationships Across the Parenchymal Field in Women with Breast Cancer on FFDMs and Specimen Radiographs," *AAPM Annual Meeting*, Washington D.C., July 2022.

N. Baughan, H. M. Whitney, K. Drukker, B. Sahiner, G. H. Kim, M. McNitt-Gray, K. J. Myers, M. L. Giger, "Task-Based Sampling of the MIDRC Sequestered Data Commons for Algorithm Performance Evaluation," *AAPM Annual Meeting*, Washington D.C., July 2022.

H. M. Whitney, **N. Baughan**, K. Drukker, K. J. Myers, M. L. Giger, "Evaluation of Diversity in the Medical Imaging and Data Resource Center (MIDRC) Open Data Commons," *AAPM Annual Meeting*, Washington D.C., July 2022.

K. Drukker, B. Sahiner, T. Hu, G. H. Kim, H. M. Whitney, **N. Baughan**, K. J. Myers, M. L. Giger, M. McNitt-Gray, "The Medical Imaging and Data Resource Center (MIDRC) Technology Development Project (TDP) 3c: Developing Tools to Assist in Task-Specific Performance Evaluation for Machine Learning Algorithms Employing MIDRC Data," *AAPM Annual Meeting*, Washington D.C., July 2022.

D. Olivera Velarde, E. Marshall, J. P. Cruz-Bastida, **N. Baughan**, C. Guo, K. Feinstein, I. Reiser, "Application of a Psychophysical Staircase Method for Determining Detection

Thresholds in Fluoroscopic Imaging," *AAPM Annual Meeting*, Washington D.C., July 2022.

N. Baughan, H. Li, L. Lan, C. Chan, M. Embury, G. Whitman, R. El-Zein, I. Bedrosian, M. L. Giger, "Parenchymal Field Effect Analysis for Breast Cancer Risk Assessment: Evaluation of FFDM Similarity Using Deep Learning Features," *RSNA Annual Meeting*, Chicago, IL, November 2021.

D. Olivera-Velarde, J. P. Cruz-Bastida, **N. Baughan**, E. Marshall, K. Blunt, K. Feinstein, I. Reiser, "Task-Based Assessment of Pediatric Fluoroscopy by Use of Vesico-Ureteral Reflux Grading," *AAPM Annual Meeting*, Virtual, July 2021.

N. Baughan, J. P. Cruz-Bastida, H. Al-Hallaq, I. Reiser, "Variations in Radiomics Features of a Multi-Texture Phantom Introduced by Deep Learning Iterative Reconstruction Algorithms," *AAPM Annual Meeting*, Virtual, July 2020.

J. P. Cruz-Bastida, N. Reiser, E. Pearson, E. Marshall, **N. Baughan**, J. George, H. Al-Hallaq, K. Feinstein, I. Reiser, "Development of Cost-Effective Phantoms for 2D X-Ray Imaging Applications Using Stackable Binary Images from an Inkjet Printer," *AAPM Annual Meeting*, Virtual, July 2020.

M. Shenouda, **N. Baughan**, J. P. Cruz-Bastida, E. Pearson, H. Al-Hallaq, "Does Radiomics Have the Potential to Assess KV-CBCT Image Performance Acquired from Phantom Data Used for Daily QA?" *AAPM Annual Meeting*, Virtual, July 2020.

Presentations

N. Baughan, J. Fuhrman, "MIDRC: Assessing ML-ready data and conducting benchmarking in AI research," *Practical Big Data Workshop*; Ann Arbor, MI, June 2022.

N. Baughan, H. M. Whitney, K. J. Myers, M. L. Giger, "Role of MIDRC sequestration and task-based distribution sampling in the independent evaluation of AI in Medical Imaging," *MIDRC Monthly Seminar*; Virtual, April 2022. (Invited.)