THE UNIVERSITY OF CHICAGO

VARIABILITY AND HARMONIZATION OF THE RADIOMICS WORKFLOW FOR IMPROVED CLINICAL IMPLEMENTATION

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
JOSEPH JAMES FOY

CHICAGO, ILLINOIS
JUNE 2020

Dedicated to Andrew

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to offer my deepest appreciation to my advisors, Dr. Sam Armato and Dr. Hania Al-Hallaq. I am truly so grateful to have received such gracious and unconditional guidance from two of the most supportive people I've ever met. Words cannot express how much I appreciate having been able to work with them over the years. I can only hope that future students will be fortunate enough to have a relationship with their advisor(s) that is also characterized by trust, encouragement, and mutual respect. From the bottom of my heart, thank you so much. I would also like to thank my committee members, Dr. Maryellen Giger and Dr. Steffen Sammet, for their advice and recommendations throughout the course of my graduate career. The invaluable medical knowledge Dr. Sammet offered ensured that my research was clinically applicable and relevant. Dr. Giger provided me with such a broad and extensive understanding of all things machine learning and computer-aided diagnosis, and her knowledge, expertise, and attention to detail is truly irreplaceable.

Several additional members of the University of Chicago faculty, staff, and student body, past and present, had a pivotal role in my success as a PhD student. I am grateful to Roger Engelmann and Adam Starkey for sharing their extensive knowledge pertaining to computer software development and coding, often writing additional programs to help expedite data collection and analysis. Thank you to Dr. Eyjolfur Gudmundsson for walking me through the intricacies of the Armato lab and graciously translating your vast knowledge of computer science to me. I would also like to thank Dr. Alex Cunliffe for building the foundation of my thesis research and for providing me with all the resources I needed to successfully follow through on my aims. Thank you to Dr. Nicholas Gruszauskas and Susan Fruth for their help in the Human Imaging Research Office collecting valuable imaging data and offering their knowledge on clinical research policies. I am truly indebted to Kristen Wroblewski from the Biostatistics clinic who met with me routinely and answered several emails to confirm the soundness of my statistical approaches and data analysis. She truly has an uncanny ability to grasp abstract concepts from vastly different fields while offering her expert advice and

guidance. Dr. Karen Drukker was also glad to offer her knowledge of various modeling and statistical approaches from across the hall. I'm grateful to Dr. Ingrid Reiser for her input and expertise pertaining to a number of imaging concepts. Of course, I would like to thank the GPMP student body, particularly my classmates Dr. Kayla Robinson and Talon Chandler, for providing I corridor with a light-hearted and welcoming atmosphere. I feel truly blessed to have worked alongside such approachable and pleasant colleagues. I would also like to thank my masters advisor at the University of Michigan, Dr. Martha Matuszak, who initially inspired me to pursue a career in medical physics. She has always been an unwavering powerhouse in the field and a shining light for me to aspire to. Thank you for believing in me.

I would like to offer my greatest and most sincere gratitude to my family. Thank you to my parents, Karl and Julie Foy, for always supporting and uplifting me. They have fostered in me the sense of self-worth, dedication, and confidence that I have relied on over the course of my college career. Thank you to my brother, Mark, for his undying kindness. You always made your love for your family known, which I've always appreciated. Thank you to Gabriel for supporting and comforting me in the hard times when I may not have completely believed in myself. Finally, thank you so much to my lovely fiance, Andrew, for standing by my side all the way, offering words of encouragement and inspiring me to be a better man. I can't wait to see what our future holds. Thank you.

ABSTRACT

The interest in texture analysis and radiomics has increased greatly in recent years resulting in an increase in the number researchers incorporating these machine-learning methods into their investigations. More recently, however, investigators have illustrated the lack of reproducibility in radiomics research preventing the translation of radiomics-based classification models into clinical practice. Therefore, it is important to understand the degree of variability in radiomics research due to differences in each component of the radiomics workflow and whether this variability persists when applied to a clinical task. Additionally, methods that could allow for the harmonization of radiomics research across institutions may allow for this research to be more easily reproduced, validated, and implemented into clinical practice.

This dissertation investigates the dependency of radiomic features on various components of the radiomics workflow. This work begins by quantifying the effect of CT image acquisition and reconstruction parameters on the resultant radiomic features and assessing potential methods to mitigate these effects. To accomplish this goal, radiomic features were extracted from CT scans depicting a cadaveric liver when scans were acquired and reconstructed with 17 different imaging parameters. Feature values were compared between one reference scan and the remaining 16 modified scans. We found that reducing the field of view or using coronal slices instead of axial slices resulted in the greatest number of features reflecting significant differences (67.6% and 35.9%, respectively), while slight changes in tube voltage, pitch, or slice interval resulted in the smallest number of features reflecting significant differences (0.7% each). To mitigate the differences in feature values between scans, five harmonization methods were implemented: histogram normalization, pixel size resampling, Butterworth filtering, resampling and filtering combined, and ComBat harmonization. While histogram normalization maintained or reduced the number of features reflecting significant differences for each scan, ComBat harmonization reduced the number of features reflecting significant differences to zero for all imaging parameters.

In addition to the variability of image acquisition and reconstruction parameters, the dependence of radiomic features on the feature calculation process was also investigated. Five radiomic software packages (A1, A2, IBEX, MaZda, and Pyradiomics) were used to calculate 12 features common among the five packages using databases of mammograms, head and neck CT scans, and breast MRI scans as input. For each database, 11 out of 12 features reflected significant differences among packages for the mammography and head and neck CT databases, while 9 out of 12 features reflected significant differences for the breast MRI database. When assessing the agreement in feature values among packages using the intraclass correlation coefficient (ICC), 5, 4, and 5 out of 12 features reflected excellent agreement for the mammography, head and neck CT, and breast MRI databases, respectively. These discrepancies in feature values were found to be the result of differences in image importation and preprocessing, algorithm implementation, and feature-specific parameters.

The effect of the variability in radiomics software was also quantified by extracting feature values with various software packages from CT scans of patients undergoing radiation therapy (RT) for esophageal cancer. Due to therapy, 19% of patients developed radiation pneumonitis (RP). The changes in eight feature values between pre- and post-RT CT scans were calculated using the three software packages: A1, IBEX, and Pyradiomics. The changes in feature values were used in logistic regression to classify patients with RP, and the AUC value for each feature was compared across packages. Based on analysis of variance (ANOVA), features associated with RP development differed among the three packages for 2 out of 8 features. When assessing classification ability based on AUC value, first-order features reflected greater agreement in classification ability but began to deviate for higher-order features.

Finally, the potential of mitigating the differences in software packages was assessed using four modeling methods: one using the differences in feature values between pre- and post-RT CT scans as was performed in the previous chapter (M_{Avg} models) and three using ComBat harmonization implemented at different components of the feature calculation and model-

ing workflow ($M_{ComBat1}$, $M_{ComBat2}$, $M_{ComBat3}$ models). Using each of the four methods, patients were again classified based on RP status when features were calculated using each of the three radiomics software packages (A1, IBEX, and Pyradiomics). Repeated measures ANOVA assessed differences in feature values among packages, and the agreement in AUC values among packages was quantified with the ICC. M_{Avg} models resulted in 5 out of 8 features reflecting significant differences, while the three ComBat-based models reduced the number of feature reflecting significant differences (0 - 2 features). Using the differences in feature values resulted in moderate agreement in AUC values (ICC: 0.727), while the three ComBat-based methods resulted in decreased agreement (ICC: 0.637 - 0.677). When features were normalized prior to ComBat harmonization, ICC values increased, but did not greatly improve agreement over the M_{Avg} Models (ICC: 0.733). While ComBat harmonization has shown great potential in mitigating the effects of image acquisition and reconstruction parameters, removing the differences in radiomic software packages to align classification performance may be outside the scope of what ComBat is designed to accommodate.

This dissertation demonstrated the potential variability that can be introduced into radiomics research due to differences in each component of the radiomics workflow. Fully understanding the implications of altering the image acquisition and reconstruction of medical images or changing the feature calculation process will aid investigators in determining methods appropriate for their radiomics research.

CHAPTER 1

INTRODUCTION

1.1 Texture Analysis

Medical imaging has played a vital role in clinical practice, especially for the detection and diagnosis of various diseases and assessing a patient's response to treatment. Typically, radiologists make these assessments through the qualitative assessment of patient images; however, visual interpretations of a patient's condition can vary among groups of radiologists and also within the same radiologist at different times (i.e., radiologist fatigue) [1,2]. Additionally, early disease detection and accurate prognosis estimation could be improved with more standardized, quantitative methods as clinical practice becomes more precise and personalized.

Previous investigations have attempted to measure the association between various quantitative clinical metrics and a patient's condition. When evaluating pulmonary function using the standard uptake values (SUVs) derived from [F-18]-fluorodeoxyglucose positron emission tomography (FDG-PET) scans, Rakheja et al. [3] found significant relationship between several SUV metrics and neoplastic lesion malignancy. A study by Ma et al. [4] related regional doses from chest radiation therapy (RT) to changes in lung density in computed tomography (CT) scans. Even earlier studies starting in the 1960s were capable of quantifying various gray-level and shape-based metrics to characterize different tissues or detect lesions in radiographs and mammograms [5-13].

Through the refinement of each of these automated detection and diagnosis schemes, more comprehensive and accurate computer-aided diagnosis (CAD) and quantitative imaging methods have been developed to aid radiologist in making medical decisions. These studies have developed a number of features and feature categories each designed to quantify various intensity and textural characteristics attributed to the image in question. The subsequent analysis of these features, also known as texture analysis, has allowed researchers to quantify

the mathematical interrelationships of pixels with similar or dissimilar contrast.

1.1.1 Gray-Level Histogram Features

Gray-level histogram features (also called first-order or gray-level intensity features) quantify the distribution of pixel intensities by constructing a histogram of the pixel values. This histogram is then characterized by various mathematical properties such as the maximum and minimum pixel values, the average and median pixel values, as well as the pixel values that partition the pixel value distribution at different quantiles. The shape of the distribution can be quantified by the skewness, kurtosis and standard deviation, which reflect the concentration of pixels that are assigned different values. These features can capture valuable information reflecting the general intensity of different tissues; however, these do not capture image "texture" because they do not illustrate the spatial relationships of pixels.

1.1.2 Gray-Level Co-Occurrence Matrix

While initial studies typically focused on a limited number of intensity- and shape-based features, subsequent studies quantified the spatial relationship of pixels within an image (i.e., texture) to characterize different tissues more comprehensively. Intensity-based features quantify an image based on the relative distribution of pixel values; however, they do not reflect how pixels with different intensities relate to one another spatially. In response to this, Haralick et al. [14] compiled a number of texture features that quantify the spatial variability of pixels in any image, not only medical images. These features are derived from the gray-tone spatial-dependence matrix, commonly referred to as the gray level co-occurrence matrix (GLCM), which quantifies how often a pixel with a value a is at some distance from another pixel with a value b at a specified angle (Figure 1.1). A number of GLCM features are calculated using the resultant matrix such as the contrast or homogeneity, which characterize image gradients or patterns of differing intensity and direction.

Since the development of GLCM features, additional textural feature categories have

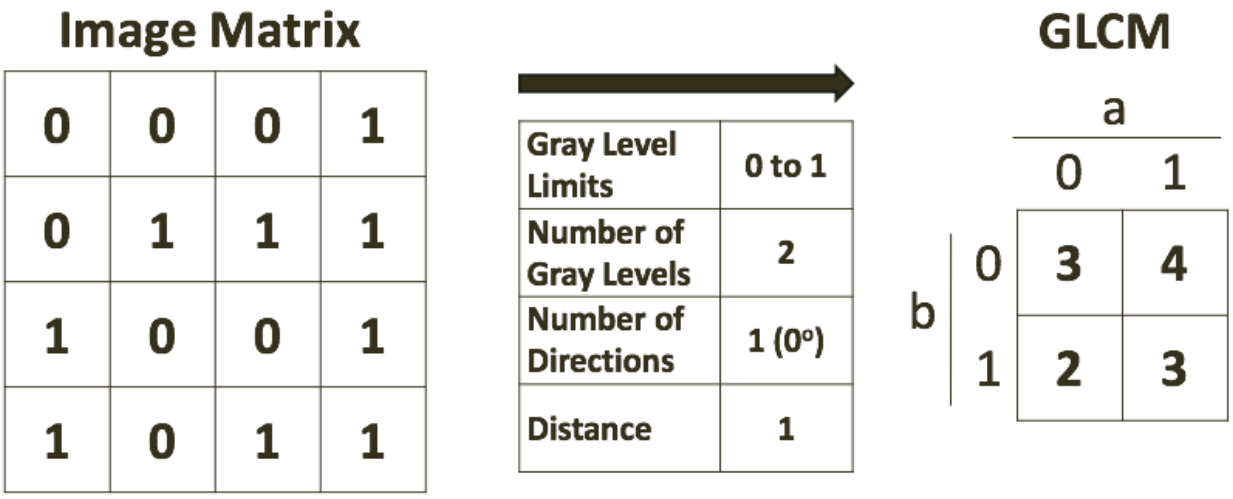


Figure 1.1: The gray-level co-occurrence matrix constructed from a binary image matrix. The construction of the GLCM is dependent on a number of parameters including the gray level limits, the number of gray levels, and the direction and distance between the two pixels in question.

been introduced including gray-level run-length matrix (GLRLM) and neighboring gray tone difference matrix (NGTDM) features. The combination of these high-dimensional texture features with shape-based features, intensity features, and other patient-specific measurements can subsequently be mined to understand trends that span populations [15].

1.1.3 Fractal Analysis

Fractal analysis and fractal features characterize the self-similarity of an image at different scales with the aims of quantifying the detail of a region [16]. This self-similarity is often characterized by the fractal dimension, which compares the number of self-similar pieces contained within the image with the magnification factor required to scale the self-similar pieces to obtain the original image. For example, a square can be broken up into 4 self-similar pieces with a magnification factor of two. The fractal dimension is subsequently quantified using the equation:

$$Dimension = \frac{log(\#Self - SimilarPieces)}{log(MagnificationFactor)} = \frac{log(4)}{log(2)} = 2 \tag{1.1}$$

A square therefore has a fractal dimension of 2, a cube a fractal dimension of 3, and so on. The fractal dimension quantifies how "complicated" a particular image is and how many points generally fall within that image set. For gray-scale images such as those used in medical imaging, the number of self-similar pieces and the magnification factor are more complicated to compute, and pixel values are often conceptualized as the "heights" of the pixels in a third dimension normal to the image plane. The heights of these columns and how self-similar they are can then be characterized by a number of methods including the blanket method, Brownian motion method, and the box-counting method [17-20].

1.1.4 Fourier Analysis

Fourier analysis uses a rotationally invariant Fourier transform to decompose an image region into a sum of sine and cosine waves of varying amplitude, frequency, and phase reflecting the relative fluctuations of pixel values in that region. The transformed image reflects the spatial frequency of pixel values in the x and y spatial directions. Images that contain fine detail and closely spaced pixels of different values will have a larger amount of high-frequency data contained in the corresponding Fourier transform, while low-contrast "smooth" textural patterns will be reflected in the low-frequency portions of the Fourier transform. The various components of the Fourier transform can be characterized to describe the original image such as the root-mean square (RMS) and first moment of the noise power spectrum [21,22].

1.1.5 Laws' Filter Features

Laws' filter features are used to describe the region microstructure through the convolution of the image regions with various filters, which subsequently emphasize patterns that

align with the filters. One-dimensional filter vectors including spot, wave, ripple, edge, and level filters can be combined using an outer product of two vectors to create a two-dimensional filter.

$$\text{Level: } L5 = [1\ 4\ 6\ 4\ 1]$$

$$\text{Edge: } E5 = [-1\ -2\ 0\ 2\ 1]$$

$$\text{Spot: } S5 = [-1\ 0\ 2\ 0\ -1]$$

$$\text{Wave: } W5 = [-1\ 2\ 0\ -2\ 1]$$

$$\text{Ripple: } R5 = [1\ -4\ 6\ -4\ 1]$$

From these convolved images, a number of first-order features can be calculated with the aims of characterizing the emphasized structures [23]. In this thesis, a total of 142 features were computed. A complete list of these features along with their mathematical definitions can be found in the appendix.

1.2 Radiomics

The mining of high-dimensional image data to support a clinical decision, also known as radiomics, has allowed investigators to aid radiologists in performing a number of clinical tasks. Based on the radiomic features extracted from previous patients with known disease status or prognosis, radiomic features extracted from new patients can be combined with a classifier to answer a specific clinical question; for example, is the tissue depicted in a given image healthy or cancerous? Is the tumor in the image malignant or benign?

Many investigators have shown that these radiomics-based models can accurately detect different diseases, segment different tissues, and predict patient prognosis [2,13,24,25]. Wibmer et al. [26] analyzed magnetic resonance imaging (MRI) scans and reported that several GLCM features were capable of differentiating between cancerous and non-cancerous prostates. When applied to a segmentation task, Sepate et al. [27] used a fuzzy region growing algorithm to automatically segment lesion candidates in mammograms and combined

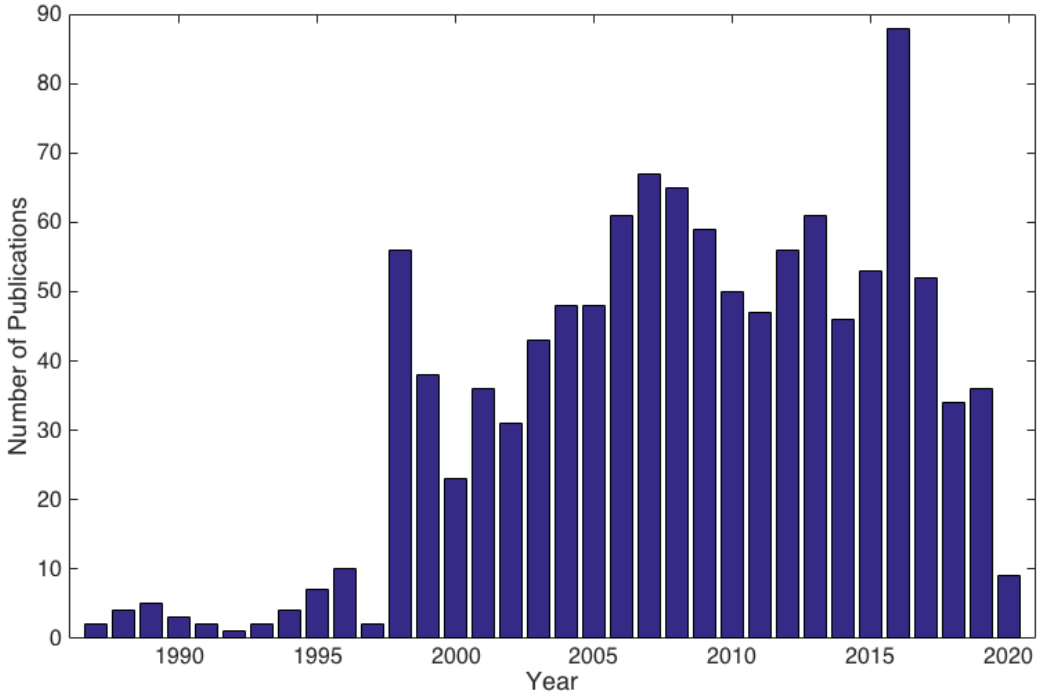


Figure 1.2: The number of publications incorporating computer-aided diagnosis/detection (CAD), texture analysis, or radiomics in Medical Physics.

textural and geometric features to accurately classify candidates as benign or malignant.

Because of the promising results these radiomics-based studies have reported when performing a variety of clinical tasks, many investigators have become interested in incorporating radiomics into their work. A review article published in 2008 reported that the number of manuscripts published in *Medical Physics* that incorporated CAD have increased by over ten times over the course of almost two decades, and since then, the interest in radiomics has grown even more [13]. When including texture analysis and radiomics, the number of publications have increased by nearly two orders of magnitude from 1987 to 2020 (Figure 1.2).

The increased interest in radiomics research is in part due to the transparency of radiomics-based models. Compared to other contemporary machine learning methods, such as convolutional neural networks (CNNs), radiomics models are relatively easy to conceptu-

alize and use feature selection methods and classification models that are well understood. These methods have been used in publications that span several fields, not only medicine, and are quite popular because they are considered generalizable, digestible, and intuitive. Radiomic features often have physical interpretations: for example, the first-order features variance and entropy are related to the spiculation and heterogeneity of tissues, and mean and median are related to the general intensity of pixel values in a given region. In theory, the combination of these quantitative descriptive features can be subsequently analyzed to fully characterize different tissues. In comparison, CNNs learn how much weight to apply to different filters to maximize the classification ability based on the training dataset provided to the CNN; however, these filters usually do not have physical interpretations. Additionally, CNNs typically require much larger databases of medical images for training compared to radiomics-based models before they achieve clinically acceptable performance [25,28].

1.3 Variability in Radiomics Research

While radiomics research has shown promise when applied to a variety of clinical tasks, its translation into clinical practice has been limited because radiomics research is difficult to reproduce and validate. The radiomics workflow is highly complex and requires several complex steps (Figure 1.3) [29]. All radiomics studies begin with the medical image acquisition, where a number of image acquisition and reconstruction parameters affect the appearance of the resultant image and consequently affect the radiomic feature values calculated with these images. Several studies have quantified the variability in radiomics studies because of differences in image acquisition parameters [30-34]. Mackin et al. [35] studied the effect of x-ray tube current on radiomic features when features were extracted from CT scans of a non-anatomic phantom. They reported that tube current can greatly affect radiomic features, but not for tissues or materials similar to lung tumors for clinically relevant tube currents around 120 kVp. Mendel et al. [36] also quantified the dependency of radiomic features on digital mammography system manufacturers and found a large degree of variability

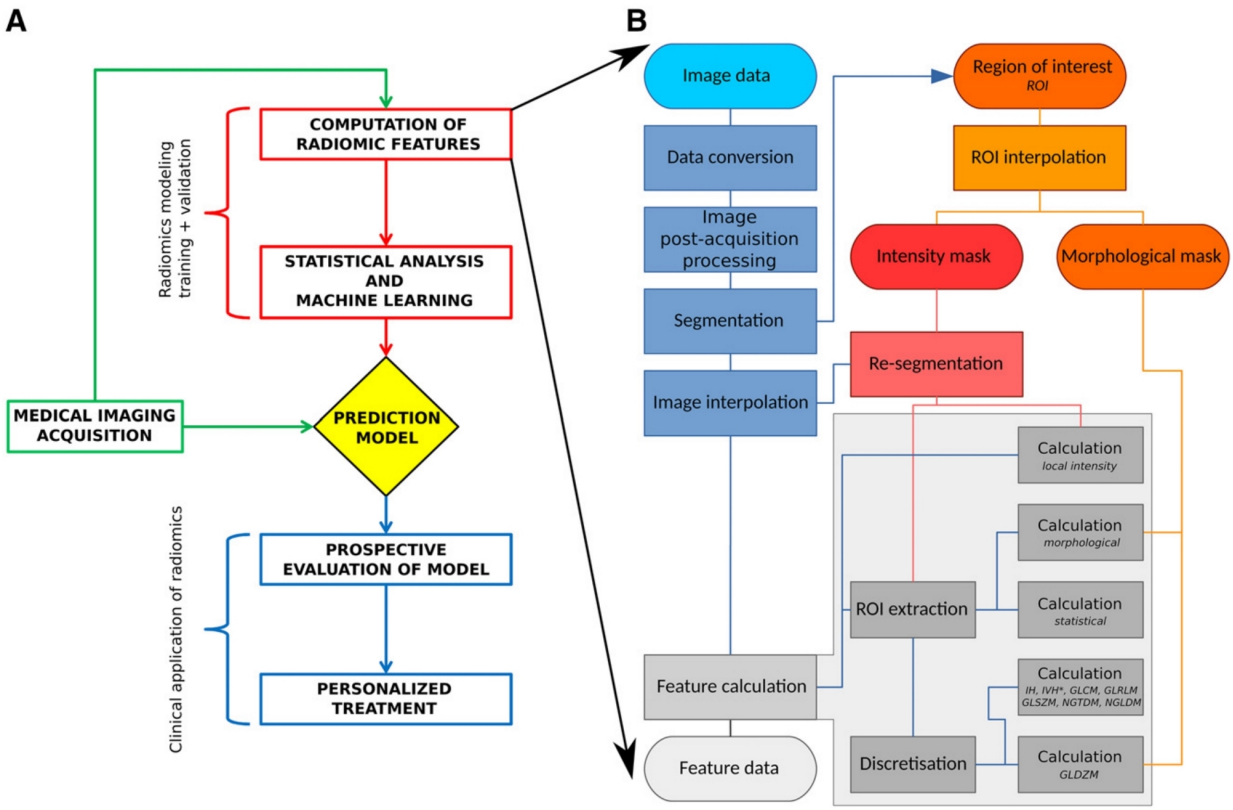


Figure 1.3: Illustration of the radiomics workflow with the aims of producing a predictive model for improved personalized medicine. (Adapted from Valliéres et al. J. Nucl. Med. 59(2) 2018).

in many features due to differences in unit manufacturer. To limit the dose to individual patients, these studies have focused primarily on images of phantoms or images from an ensemble of patients. Phantom images, however, are not a perfect surrogate for human tissue, and a large degree of uncertainty is introduced when assessing trends across patient images because of the inherent variability in patient anatomy and positioning. Additionally, allowances typically have to made when curating patient databases because images are often acquired with differing parameters to best suit clinical practice. Retrospective studies that use patient images, therefore, can only limit the differences in patient images by ensuring imaging parameters are as similar as possible across patients.

Once the images are acquired, the radiomic features are calculated and the statistical analysis and machine learning are performed. In Figure 1.3.b, the complexity of the radiomic

feature calculation process can be seen, where variability can be introduced in any of the numerous steps [29,37].

When the potential of radiomics-based detection and diagnosis schemes was first recognized, many institutions developed their own radiomics software packages to be used in their labs or institutions because of the limited access to this software at the time. However, due to the complexity of each of these packages, the continued development of radiomics packages resulted in discrepancies among them. Furthermore, because of the digestibility and continued success of radiomics research, additional investigators attempted to incorporate radiomics and quantitative imaging into their work. However, radiomics software is often designed with a particular purpose and to process images from a particular imaging modality or tissue type. The increased desire to tap into radiomics research resulted in many of these packages being used with a "one-size-fits-all" approach without considering the underlying mechanisms embedded into the algorithms that could potentially result in variations among radiomics software.

The complexity of the radiomics workflow also results in inadequate reporting of research methods, and each step shown in Figure 1.3 can vary due to differences in the computational feature definitions and feature parameters. The breadth of this work makes it nearly impossible for researchers to report a completely comprehensive outline of their methods, making it difficult for other researchers to reproduce and validate their results. This lack of validation prevents the implementation of radiomics research into clinical practice.

1.4 Standardization of Radiomics Research

Many initiatives have recognized the need for greater standardization of radiomics research with the aims of achieving improved reproducibility and translation of radiomics research into clinical practice [38-41]. In response to the growing demand for more translatable radiomics research, additional initiatives have attempted to offer recommendations to standardize research across institutions. The International Biomarker Standardization

Initiative (IBSI) is an international effort composed of 55 researchers from 19 institutions. Each researcher used their respective radiomics software to calculate the union of features across all researchers. Features were extracted using a small 3-dimensional digital phantom in addition to CT scans of lung cancer patients. Each researcher iteratively modified their software until feature values agreed among institutions, and features were considered "standardized" if at least half of the packages achieved the same feature values. Through this process, 99.4% of features agreed when computed with the digital phantom, and 96.4% agreed when computed with CT scans [42-44]. Despite their efforts, complete harmonization was not achieved among a relatively limited cohort of researchers even when agreement in feature values among half of the researchers was considered "standardized." Such a lenient threshold may not be sufficient to state that agreement among features was truly achieved. Additionally, given the limited size and range in pixel values, the digital phantom used in the analysis is insufficient and is not reflective of human tissue. Because radiomics packages are often designed to analyze images of a particular imaging modality or tissue type, additional initiatives should incorporate more diverse imaging databases, stricter standardization thresholds, and a larger number of participating institutions.

In addition to the IBSI, the Quantitative Imaging Network, the American Association of Physicists in Medicine (AAPM), the European Society of Radiology (ESR), and the Quantitative Imaging Biomarkers Alliance (QIBA) from the Radiological Society of North America (RSNA) have created initiatives aimed at harmonizing radiomics research. A number of tools and resources have been established as well to further guide radiomics research towards greater reproducibility such as the Responsible Research and Innovation website (www.rri-tools.eu) as well as the Cancer Imaging Archive (www.cancerimagingarchive.net). These resources advocate for the "FAIR guiding principles," which state that "all research objects should be findable, accessible, interoperable, and reusable" to facilitate the translation of radiomics research into clinical practice [29,45]. Valliéres et al. [29] provided a relatively comprehensive list of recommendations on the various aspects of radiomics research that

should be reported to make this work more reproducible including the details of the image preprocessing (image conversion, processing, and ROI segmentation) as well as pixel interpolation methods (voxel dimensions, image interpolation method, etc.) (Table 1.1).

Additional studies have investigated the potential of harmonizing radiomic features through image preprocessing or by harmonizing radiomic features after they have been calculated from the original, unmodified images. Mackin et al. [46] combined voxel size resampling and Butterworth filtering to mitigate the effects of differing voxel sizes on radiomic feature values from CT scans of a non-anatomic phantom and found moderately improved agreement in feature values after image processing. Orlhac et al. [47] investigated the potential of an empirical Bayes harmonization method, which was originally developed to mitigate differences in data acquisition methods in genomics studies. Features were calculated from breast tumors in PET scans when scans were acquired with different system manufacturers. After harmonization, the agreement in textural and standard uptake values between manufacturers was greatly improved.

1.5 Purpose and Clinical Significance

Initial radiomics studies were capable of automatically detecting lung nodules and calcifications in mammograms; however, these studies were plagued by a large number of false positives for each patient [13]. As radiomics research developed and became more refined, subsequent radiomics schemes were found to aid radiologist by greatly reducing the number of false negatives [24,48,49]. For example, when screening for breast cancer in mammograms, previous studies had reported false-negative rates between 10% and 30% [50,51], but even studies conducted in the early 1990s illustrated the potential benefit of integrating radiomics into clinical practice. Chan et al. [52] reported a breast calcification detection method in 1990 that used spatial filters and signal-to-noise ratio (SNR) enhancement combined with power spectra discrimination to greatly reduce the false-positive rate. This study reported a statistically significant improvement in a radiologist's ability to detect microcalcifications

Table 1.1: Reporting guidelines for more responsible and reproducible radiomics research proposed by Valliéres et al. (Adapted from Valliéres et al. J. Nucl. Med. 59(2) 2018))

Category	*Guideline*
General	
Image acquisition	Acquisition protocols and scanner parameters such as equipment vendor, reconstruction algorithms and filters, field of view and acquisition matrix dimensions, MRI sequence parameters, PET dose, CT x-ray energy (kVp), and exposure (mAs).dimensions, MRI sequence parameters, PET acquisition time and injected acquisition time and injected dose, CT x-ray energy (kVp), and exposure (mAs).
Volumetric analysis	Specification of whether imaging volumes were analyzed as separate images (2-dimensional) or as fully-connected volumes (3-dimensional).
Workflow structure	Sequence of processing steps leading to extraction of features.
Software	Software type and version of code used for computation of features.
Image preprocessing	
Conversion	How data were converted from input images (e.g., conversion of PET activity counts to SUV and calculation of ADC maps from raw diffusion weighted MRI signal).
Processing	Image-processing steps after acquisition (e.g., noise filtering, intensity nonuniformity correction in MRI, and partial-volume effect corrections).
ROI segmentation	How ROIs were delineated in images (e.g., software or algorithms used, number of persons and their level of expertise [specialty, experience], method of reaching consensus, and mode [automatic or semiautomatic]).
Interpolation	
Voxel dimensions	Original and interpolated voxel dimensions.
Image interpolation method	Method used for interpolating voxel values (e.g., linear, cubic, or spline) and for aligning original and interpolated grids.
Intensity rounding	Rounding procedures for non-integer interpolated gray levels (if applicable) (e.g., rounding of Hounsfield units in CT images after interpolation).
ROI interpolation method	Methods used for interpolating ROI masks and for aligning original and interpolated grids.
ROI partial volume	Minimum partial-volume fraction required to include an interpolated mask voxel in the interpolated ROI (if applicable) (e.g., minimum partial-volume fraction of 0.5 when using linear interpolation).
ROI resegmentation	
Inclusion/exclusion criteria	Criteria for inclusion or exclusion of voxels from the ROI intensity mask (if applicable) (e.g., exclusion of voxels with Hounsfield unit values outside predefined range inside the ROI intensity mask on CT images).
Image discretization	
Discretization method	Method used for discretizing image intensities before feature extraction (e.g., fixed bin number, fixed bin width, and histogram equalization).
Discretization parameters	Parameters for image discretization (e.g., number of bins, bin width, and minimal value of discretization range).
Feature calculation	
Feature set	Description and formulas of all calculated features.
Feature parameters	Settings for calculation of features (e.g., voxel connectivity, with or without merging by slice, and with or without merging directional texture matrices).
Calibration	
Image-processing steps	Specification of which image-processing steps match benchmarks of the IBSI.
Feature calculation	Specification of which feature calculations match benchmarks of the IBSI.

in mammograms when aided by a computer compared to when the same observer study was performed without the aid of a computer. Since these earlier studies, radiomics-based methods have become even more sophisticated, enabling radiologists to detect smaller calcifications, predict patient prognosis, and streamline clinical practice [2,13].

Despite these promising results, clinical implementation has been hindered by the limited validation of radiomics models. As previously mentioned, the radiomics workflow is highly complex and requires several steps, making it difficult for subsequent studies to reproduce and validate the results from previous radiomics studies. Consequently, the clinical implementation of these classification schemes has been limited. The purpose of this dissertation is to quantify the variability in the radiomics workflow, specifically during the acquisition and reconstruction of medical images as well as during the feature calculation process. Understanding the sources of this variability can highlight areas of radiomics research that are more prone to variability and identify the aspects of this research that should be explicitly reported in publications. Radiomic features that are invariant to the components of the radiomics workflow such as various image acquisition or reconstruction parameters or feature-specific parameters (e.g., GLCM parameters) should be highlighted. Additionally, this dissertation aims to further illustrate the need for greater harmonization of radiomics research across institutions.

This dissertation also aims to evaluate the effect of variability in radiomics research when applied to a clinical task, namely, classifying whether esophageal cancer patients developed radiation pneumonitis (RP). Previous studies have emphasized the need for early detection of RP [53]; however, if the classification of patients with RP varies when different software packages are used to calculate radiomic features, the translation of this highly beneficial detection method into clinical practice will be hindered. The implications of this variability will also be analyzed when the statistical analysis and machine learning steps are altered (Figure 1.3).

The final purpose of this dissertation is to investigate the potential of different modeling

and harmonization techniques that may be implemented to align the classification ability when different software packages are used to calculate radiomic features. Retrospective analysis of medical images and radiomic features would allow for investigators to continue to use clinically acquired images and to also use their own radiomics software while allowing for the translation of their radiomics research to outside institutions.

1.6 Dissertation Outline

The aim of this dissertation was to quantify the variability of radiomic feature values due to the inconsistencies in each aspect of the radiomics workflow. Furthermore, this dissertation aimed to quantify the effect of this variability and to also offer recommendations on how to limit this variability using the workflow illustrated in Figure 1.4. To accomplish this goal, the variability in image acquisition and reconstruction parameters was quantified, and methods of mitigating these differences were developed. In Chapter 2, a number of CT scans depicting a cadaveric liver were obtained, each acquired and reconstructed with different imaging parameters. Each scan had only one parameter changed from the corresponding reference scan to isolate each variable. Radiomic features from various feature categories were calculated from the segmented liver in each scan. Features were compared across scans, and the relative number of features reflecting significant differences between reference and modified scans was calculated. Differences in field of view (FOV), imaging orientation, and imaging system manufacturer resulted in the greatest number of features reflecting significant differences, while differences in pitch, tube voltage, and slice interval resulted in the smallest. While several features reflected significant differences, the relative differences in the feature values themselves remained relatively small between reference and modified scans: only changes in FOV and image reconstruction plane resulted in a median relative change in feature values greater than 5%. An additional aim of this chapter was to use methods previously reported in the literature to limit the variability in feature values due to differences in image acquisition and reconstruction parameters: histogram normalization,

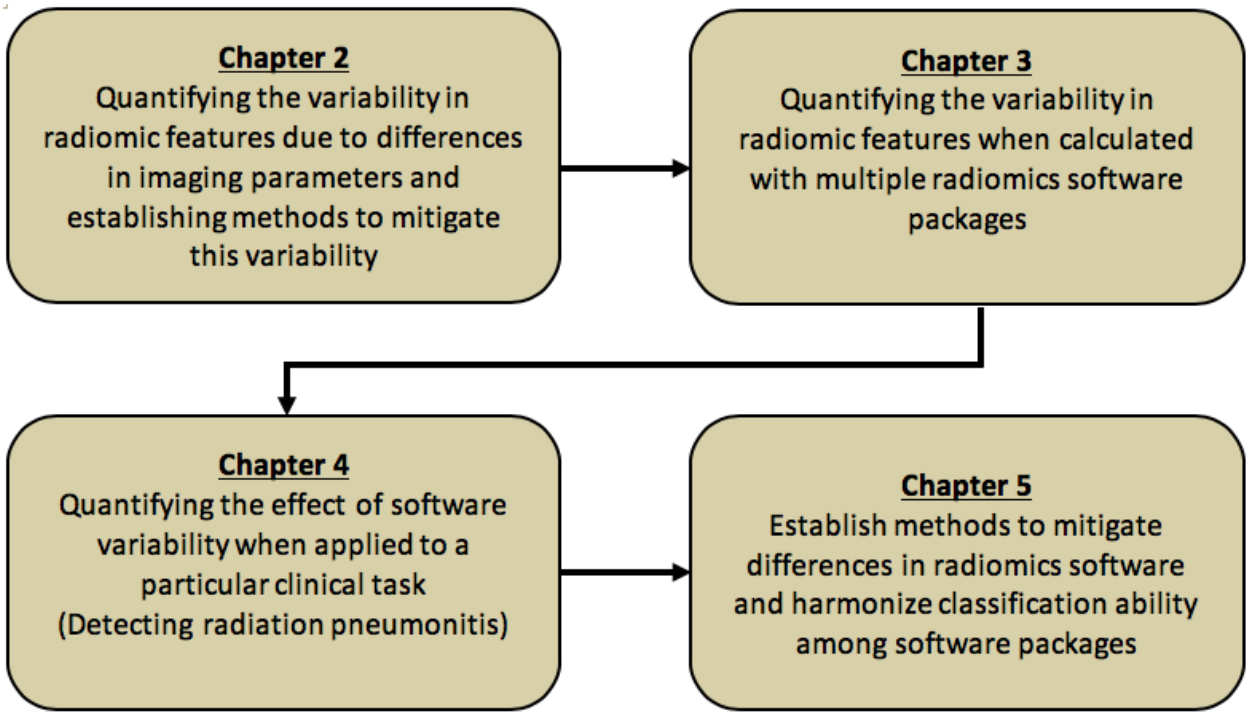


Figure 1.4: Dissertation Outline

voxel size resampling, Butterworth filtering, resampling and filtering combined, and ComBat harmonization. Histogram normalization reduced or maintained the number of features reflecting significant differences between reference and modified scans for all image acquisition and reconstruction parameters. In comparison, ComBat harmonization reduced the number of significantly different features to zero for each imaging parameter.

In addition to the variability in the medical imaging acquisition, we also assessed the dependence of radiomic features on the feature calculation process. To accomplish this goal, we obtained a number of open-source and in-house radiomics software packages, each designed to process medical images from various imaging modalities and tissue types. In Chapter 3, each package was used to calculate common features using mammography, head and neck CT tumors, and breast MRI scans as input. The variability and agreement across features was quantified, and sources of variability were assessed. Among the five packages analyzed, nearly all features differed, and most showed poor agreement; however, first-order features were much more consistent than higher-order features because of the simplicity of

these features. Sources of variation consisted of differences in image preprocessing, algorithm implementation, naming conventions, and imaging parameters.

The effect of this variability in a clinical setting was then assessed. In Chapter 4, a database of CT scans was obtained from esophageal cancer patients treated with radiation therapy (RT). A number of radiomic features were extracted from their pre- and post-RT CT scans using three radiomics software packages. Changes in feature values between time points were used to classify patients with radiation pneumonitis for each package, and the correlation of each feature with pneumonitis development was evaluated. Among the eight features all packages had in common, five features were significantly correlated with RP, and one feature was not significantly correlated with pneumonitis for all three packages. The remaining two features differed among packages in whether that feature was correlated with RP. When assessing classification ability using the area under the receiver operating characteristic curve (AUC), one feature had an AUC value that differed among packages in whether that feature had an AUC value that was significantly greater than 0.5.

The final component of this dissertation used ComBat harmonization outlined in Chapter 2 to mitigate the differences in feature values when calculated with different radiomics software to obtain greater agreement in classification ability among packages. Therefore, Chapter 5 was devoted to developing four methods of combining feature values between the pre- and post-RT CT scans. The first model used the differences in feature values between pre- and post-RT CT scans as was used in Chapter 4. The remaining three models used ComBat harmonization to align the feature values among software packages with ComBat implemented at various stages of the feature calculation and modeling process. The resultant feature values from each of the four models were used in the same modeling architecture, and the agreement in classification ability among radiomics software was assessed for each feature. We found that using ComBat successfully removed the differences in most features such that they did not significantly differ among packages, and the agreement in the feature values themselves among packages typically increased; however the agreement in classifica-

tion ability among packages was reduced. When features were normalized prior to ComBat, agreement in AUC values increased and approached that of the models using only the differences in feature values. Therefore, we concluded that the task of mitigating the differences in radiomic software packages to align the classification ability among packages may be outside the scope of what ComBat harmonization is designed to accommodate. This is the first study to quantify the effects of various components of the radiomics workflow when applied to a particular clinical task, and also the first study to attempt to align the classification ability across radiomics software.

CHAPTER 2
HARMONIZATION OF RADIOMIC FEATURE VARIABILITY RESULTING FROM DIFFERENCES IN CT IMAGE ACQUISITION AND RECONSTRUCTION: ASSESSMENT IN A CADAVERIC LIVER

2.1 Introduction

Image acquisition parameters have been shown to affect the resultant radiomic feature values including differences in tube voltage, tube current, and scanner manufacturer [30-36]. To limit the dose to individual patients, these studies have focused on phantom images or clinical images acquired from multiple patients to assess the effect of image acquisition parameters on radiomic features. Phantoms, however, are not perfect surrogates for human tissue, and using images acquired from several patients introduces a large degree of variability due to inherent differences in patient anatomy and positioning. Additionally, allowances typically must be made when curating patient databases as images are often acquired and reconstructed with different parameters to best suit clinical practice. These retrospective investigations can only limit the compounded variability introduced by these differences by ensuring image acquisition parameters are as similar as possible across patient scans.

In addition to variations in image acquisition parameters, studies have reported that radiomic features are dependent on various image reconstruction parameters [54-57]. Because a single patient scan can be reconstructed in a variety of ways without exposing the patient to additional radiation, previous studies have been able to investigate the effects of reconstruction parameters on radiomic features extracted from individual patients. Although, previous studies have not investigated how these variations compare to variations due to differences in image acquisition parameters when radiomic features are extracted from human tissue. A number of initiatives have proposed various methods of harmonizing radiomics research

[42-44]. Loizou et al. [58] investigated six post-processing MRI harmonization methods and their effect on inter- and intra-scan intensity variations in radiomic features and found that histogram normalization resulted in the smallest differences in radiomic features between scanners. Mackin et al. [46] successfully used a combination of pixel size resampling and Butterworth filtering to mitigate the effects of differing pixel sizes on radiomic features. The purpose of this chapter was to quantify the dependence of several radiomic features on various image acquisition and reconstruction parameters using a cadaveric liver and to mitigate variation using harmonization methods previously proposed in the literature.

2.2 Methods and Materials

2.2.1 Imaging Data

A normal liver was extracted from a cadaver used for medical anatomy training. Records did not indicate any liver pathology prior to death, nor were any abnormalities observed upon visual inspection or in subsequent CT scans. The liver was suspended in a polystyrene container in anatomical position, and the container was filled with a 1% agar solution to mimic human tissue attenuation and scatter. The dimensions of the resultant phantom were designed to simulate a typical human abdomen cross-section $31.0 \times 25.4 \times 18.8$ cm^3.

The routine abdominal clinical protocol implemented at AdventHealth was used as the reference CT scan, and an additional 16 scans were acquired each with modified image acquisition and reconstruction parameters (Table 2.1).

For each modified scan, one parameter was altered with the remaining parameters held constant, and the modified scans were compared to the reference scan so that the effect of each parameter could be quantified. It should be noted that two protocols, 'ThinSlice_ConvKernel' and 'ThinSlice_ReducedFOV', had modified reconstruction kernels and field-of-view (FOV), respectively, while also reconstructed with 1mm slice thickness. Both of these scans were compared to the 'ThinSlice' reconstruction, due to the matched slice thicknesses, rather than

Table 2.1: Outline of the 16 modified scans and how each modified scan differs from the reference scan. Modified parameters are highlighted in red. All modified scans were compared to the "Reference" scan other than ThinSlice_ConvKernel and ThinSlice_ReducedFOV, which were compared with the ThinSlice scan to limit the effects of confounding variables. Tube current (mAs) was not a parameter investigated in this study but is shown here to illustrate how tube voltage and CTDIvol relate to tube current.

	Protocol Name	kV	CTDIvol [mGy]	Manufacturer	Pitch	iDose	Orientation	Slice Interval [mm]	Slice Thickness [mm]	dFOV [mm]	Kernel	mAs
Image Acquisition Parameters	Reference	120	13	Philips	0.797	3	Axial	5	5	400	B	198
	80kV	80	13	Philips	0.798	3	Axial	5	5	400	B	685
	100kV	100	13	Philips	0.798	3	Axial	5	5	400	B	330
	140kV	140	13	Philips	0.797	3	Axial	5	5	400	B	134
	CTDIvol5	120	5	Philips	0.797	3	Axial	5	5	400	B	77
	CTDIvol26	120	26	Philips	0.797	3	Axial	5	5	400	B	398
	GE	120	13	GE	0.984	ASIR = 0	Axial	5	5	400	Std	232.5
	PitchHigh	120	13	Philips	1.11	3	Axial	5	5	400	B	198
	PitchLow	120	13	Philips	0.203	3	Axial	5	5	400	B	198
Image Reconstruction Parameters	iDose 0	120	13	Philips	0.797	0	Axial	5	5	400	B	198
	iDose 6	120	13	Philips	0.797	6	Axial	5	5	400	B	198
	Coronal	120	13	Philips	0.797	3	Coronal	5	5	400	B	198
	Sagittal	120	13	Philips	0.797	3	Sagittal	5	5	400	B	198
	Slice Interval	120	13	Philips	0.797	3	Axial	2.5 (overlap)	5	400	B	198
	ThinSlice	120	13	Philips	0.797	3	Axial	1	1	400	B	198
	ThinSlice_ReducedFOV	120	13	Philips	0.797	3	Axial	1	1	Liver Only	B	198
	ThinSlice_ConvKernel	120	13	Philips	0.797	3	Axial	1	1	400	C	198

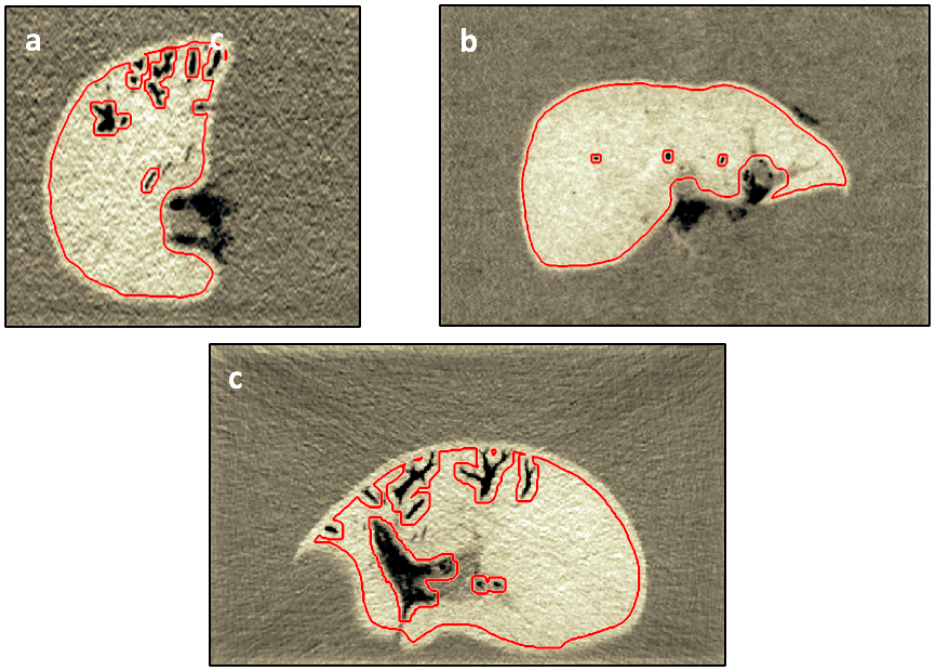


Figure 2.1: Example of a segmented sagittal (a), coronal (c), and axial section (c). (W: 100 and L: 0 in Hounsfield units (HU))

'Reference' to limit the effects of confounding variables. Additionally, the scan obtained using a GE unit instead of a Phillips unit was acquired and reconstructed using the standard site-specific imaging protocol, so not all parameters were the same between the GE scan and the reference Phillips scan. Some parameters, such as iDose level and convolution kernel, are manufacturer specific, so analogous parameters were used as recommended by a trained diagnostic medical physicist; however, while the unit manufacturer parameter combines a number of other imaging parameters (e.g., iDose level and convolution kernel), the variability among these parameters was limited.

For all 17 imaging protocols, the liver was segmented in each slice using a semi-automated segmentation method that also excluded air-filled vessels. The resultant contours were eroded by 4mm to ensure only liver tissue was analyzed (Figure 2.1).

2.2.2 Feature Calculation

A total of 142 two-dimensional radiomic features were calculated from each slice for all imaging protocols. The features calculated included first-order histogram features (n = 22), gray-level co-occurrence matrix (GLCM) features (n = 14), fractal features (n = 5), Fourier features (n = 17), and Laws' filter features (n = 84). Features were calculated using an in-house Matlab-based software with complete feature definitions provided in the appendix.

2.2.3 Statistical Analysis

For each feature, differences between each modified scan and the reference scan were assessed using unpaired two-tailed Student's t-tests with features extracted from each slice used as a distinct analyzable unit. The number of segmented slices for each scan ranged from 24 slices for scans reconstructed at 5mm to 126 slices for scans reconstructed at 1mm. Significance was assessed at the $\alpha = 0.5$ level after correcting for multiple comparisons using Bonferroni (142 features: $p < 0.0004$). Corrections were not made for each of the comparisons between modified and reference scans because the number of modified scans (n = 16) was much less than the number of features (n = 142) compared. Additionally, the Bonferroni correction is already quite conservative, and excessive significance correction has been shown to exaggerate the Type II error [59]. Features that reflected significant differences when a particular image acquisition or reconstruction parameter was changed were considered "sensitive" to these changes, while features that did not reflect significant differences when these parameters were changed were "robust" to these parameters.

To quantify the relative changes in feature values due to altered image acquisition and reconstruction parameters, each feature was averaged over all slices for each scan, resulting in 142 averaged feature values per scan. Then, the absolute relative difference between reference and modified scans was computed for each feature:

$$\Delta FV_{f,m} = \left| \frac{FV_{f,m} - FV_{f,r}}{FV_{f,r}} \right| \tag{2.1}$$

where $FV_{f,m}$ and $FV_{f,r}$ are the values of feature f for the modified and reference scans, respectively. The distributions of these relative differences were then compared across protocols to assess how much each parameter affected the resultant radiomic feature values.

2.2.4 Harmonization Methods

To investigate methods of mitigating the differences in radiomic features between reference and modified scans introduced by the differences in image acquisition and reconstruction parameters, various harmonization methods were enacted. These methods consisted of histogram normalization, voxel size resampling, Butterworth filtering, resampling and filtering combined, and ComBat harmonization. Excluding ComBat harmonization, each method harmonized the pixel values in the ROIs, and feature values were recalculated using the modified images. ComBat harmonization altered the feature values calculated using the original unmodified images.

Histogram Normalization

Loizou et al. [58] investigated various harmonization methods to control for inter- and intra-scanner variability and found that histogram normalization resulted in the greatest agreement in feature values when images were normalized prior to feature calculation. Histogram normalization modifies the distribution of pixel values within an image by scaling and translating these values to fit between a specified range of pixel values using the following equation:

$$f(x,y) = \frac{H_{Max} - H_{Min}}{g_{Max} - g_{Min}}(g(x,y) - g_{Min}) + H_{Min} \tag{2.2}$$

where $g(x,y)$ is the pixel value array of the original image, g_{Min} and g_{Max} are the minimum

and maximum pixel values of the original image, respectively, and H_{Min} and H_{Max} are the minimum and maximum pixel values of the normalized image, respectively. In this study, the pixel distributions within each segmented section were normalized such that H_{Min} was 1 and H_{Max} was 256 for all protocols.

Voxel Size Resampling

Previous studies have shown that radiomic features can vary greatly with differences in the physical size and number of voxels used to calculate these features; however, these differences can be mitigated by resampling the images to have the same physical voxel size [46,60]. To accomplish this, each scan was resampled using bilinear interpolation so that all scans have a uniform $1mm^2$ in-plane pixel resolution while the original slice thickness was maintained.

Butterworth Filtering

Mackin et al. [46,61,62] also used a 2D second-order low-pass Butterworth filter to reduce the information discrepancy between images reconstructed with different voxel sizes. They showed that the greatest agreement among images with differing pixel sizes was achieved using a cutoff frequency of 75Hz after 1mm resampling. Therefore, in the current study, radiomic features were calculated after filtering with a cutoff frequency of 75Hz both with and without resampling.

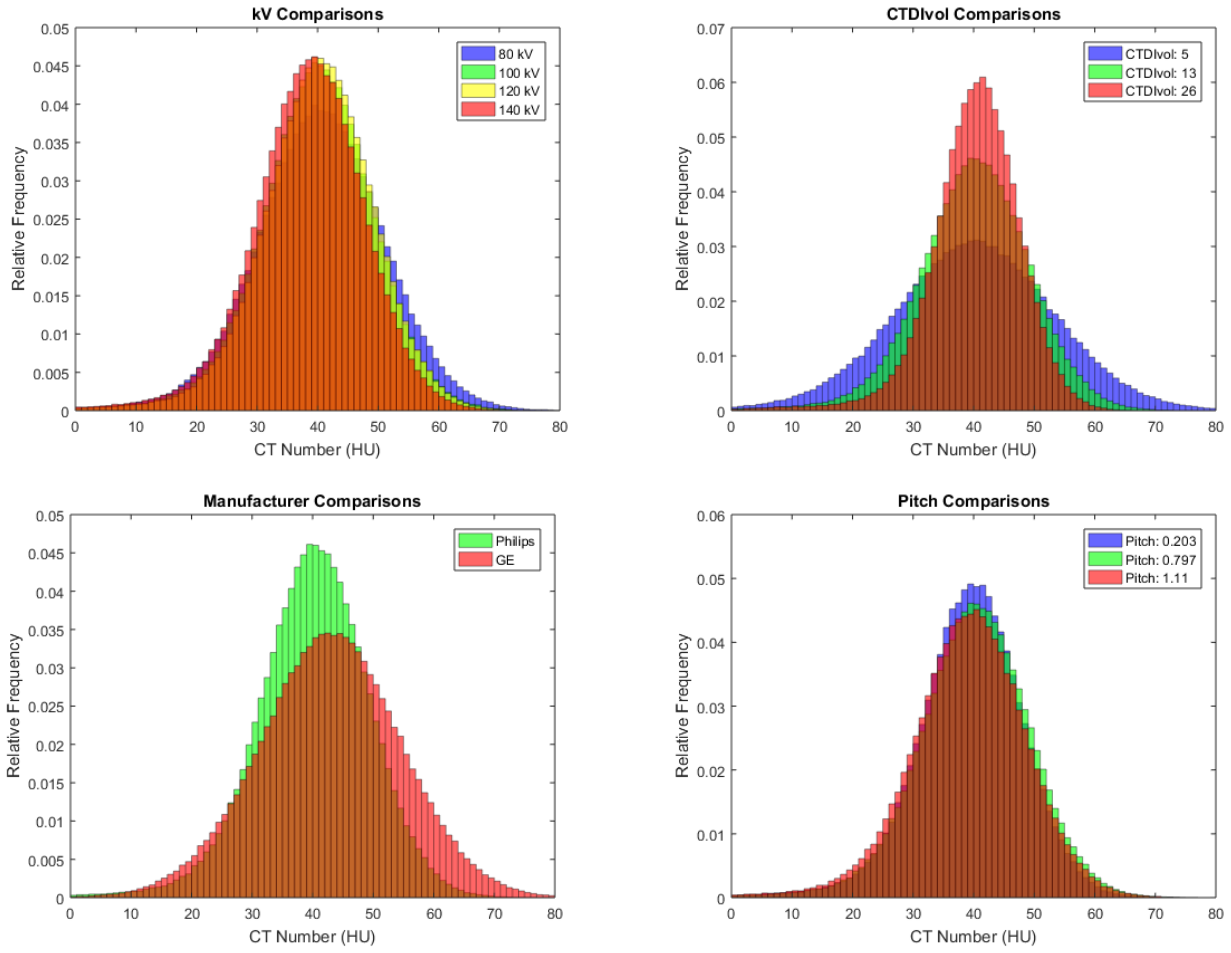
ComBat Harmonization

Johnson et al. [63] developed a method known as ComBat harmonization that was originally developed to correct for the batch effects introduced by acquiring genetic data on different days, with different equipment, or by different people. In the current study, the batch effects are variations introduced by differences in image acquisition and reconstruction

parameters. ComBat harmonization is an empirical Bayes estimation method that models the variation in radiomic feature values and estimates the effects introduced by these differences. This method has previously been shown to harmonize radiomic features extracted from PET and CT scans without altering the biological information and is also capable of preserving the imaging biomarkers indicative of the particular classification task [47,63,64].

2.3 Results

The varying pixel value distributions for each of the imaging parameters are shown in Figure 2.2.



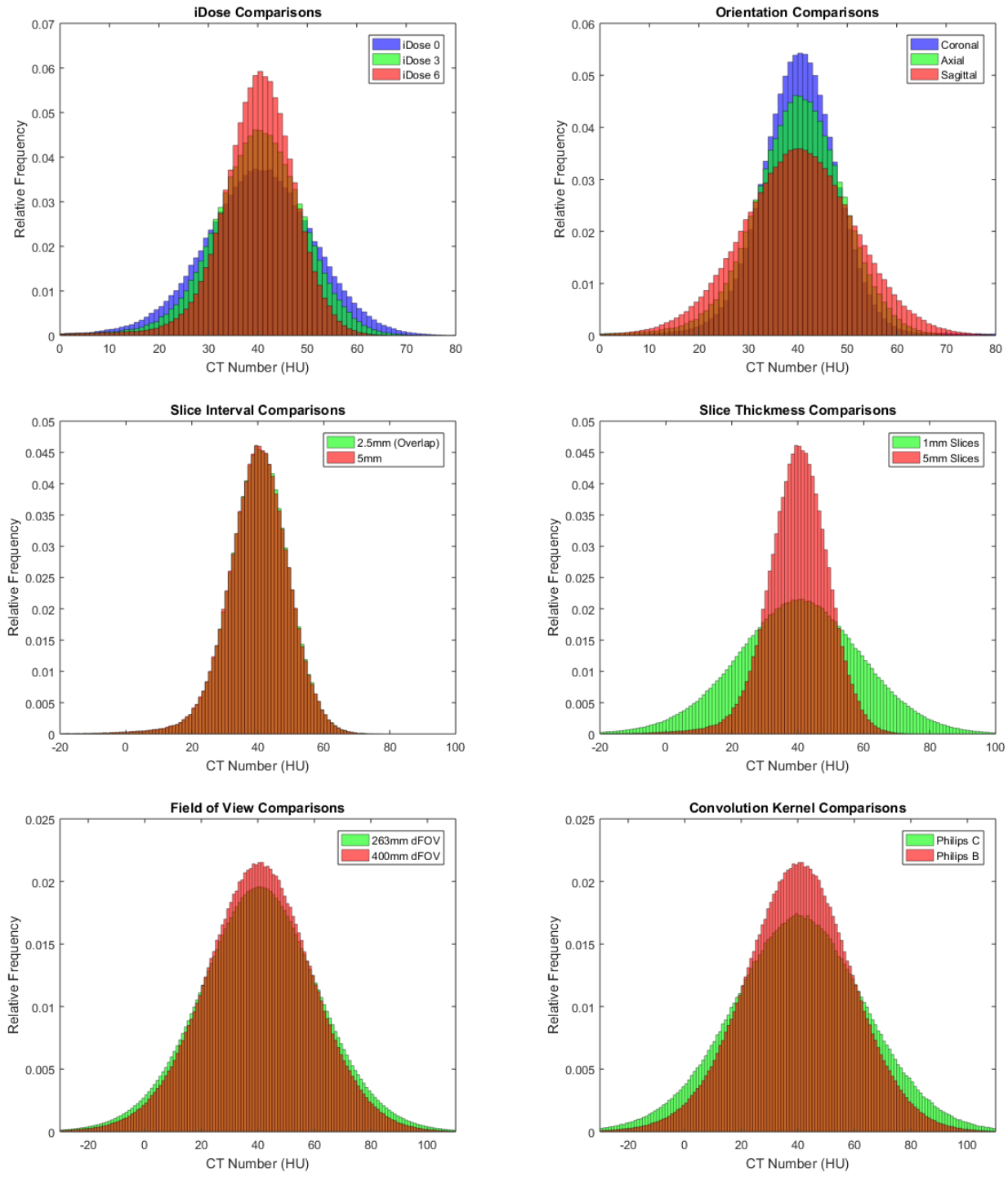


Figure 2.2: Histograms comparing the CT number distribution for each imaging protocol outlined in Table 2.1.

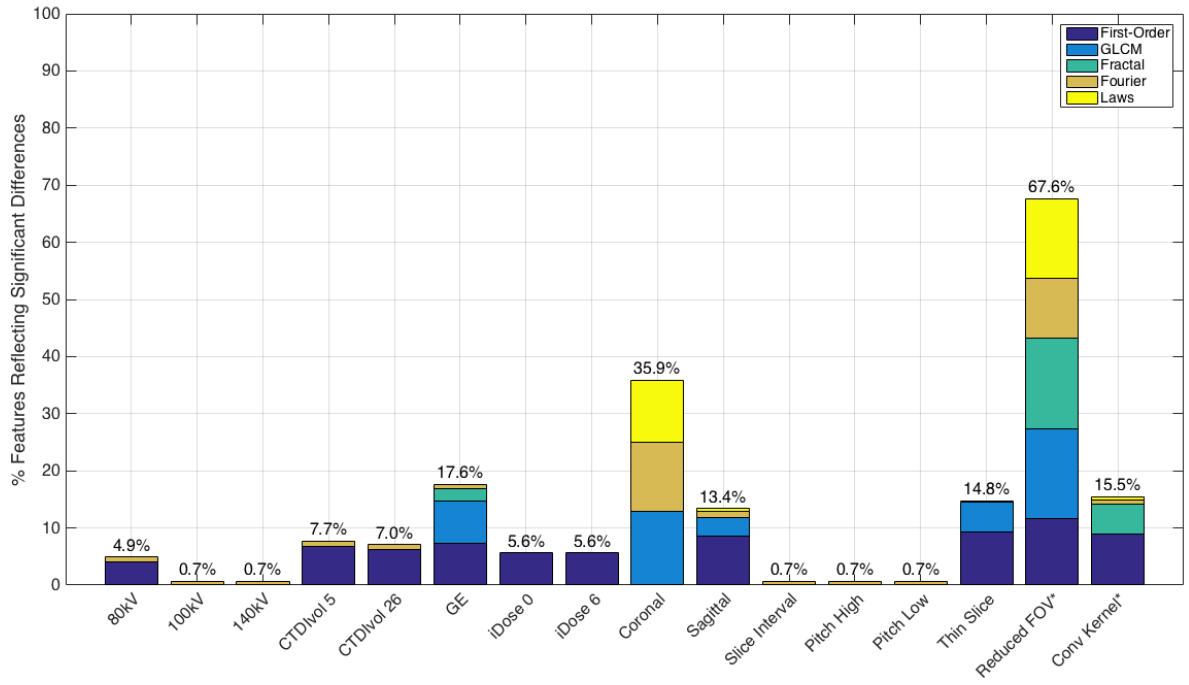


Figure 2.3: The relative number of features (%) that reflected significant differences between the modified scans and the corresponding reference scan ($p<0.0004$) along with the relative number of features reflecting significant differences for each feature category. Feature categories were scaled to account for the differing number of features in each category. All scans were compared to the 'Reference' scan other than the modified scans with reduced FOV and altered convolution kernel, which were compared to the 'ThinSlice' scan and indicated by the asterisk.

The relative number of features reflecting significant differences between the modified scan and the corresponding reference scan without harmonization is shown in Figure 2.3. The greatest number of features reflecting significant differences resulted from reducing the FOV (67.6%) and using coronal instead of axial slices (35.9%). The protocols resulting in the smallest number of sensitive features were characterized by changes in pitch, slice interval, and smaller changes in tube voltage (100kV and 140kV protocols), each with only 0.7% of features reflecting significant differences. However, as tube voltage deviated more from that of the reference scan, decreasing from 120 kV to 80 kV, the relative number of features reflecting significant differences increased from 0.7% to 4.9%.

The relative differences in radiomic features between reference and modified scans, ΔFV,

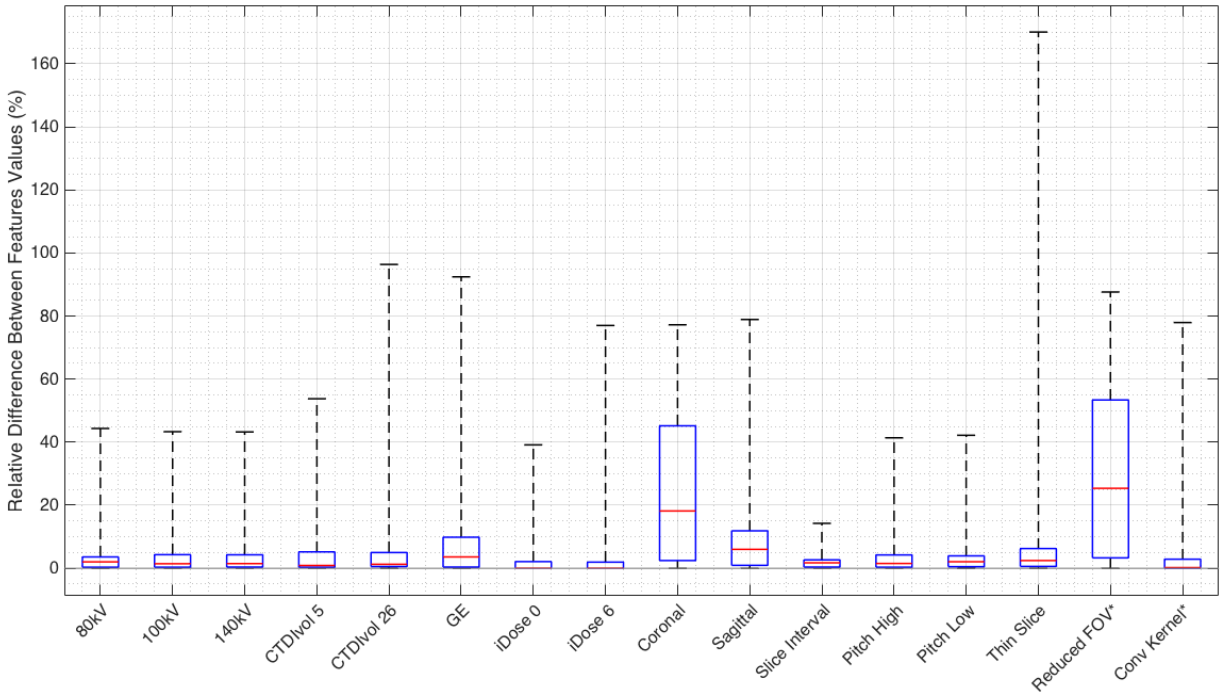


Figure 2.4: Boxplots illustrating the relative difference across radiomic features for each modified scan. The ends of boxes correspond to the first and third quartiles, while the ends of the whiskers correspond to the maximum and minimum values. All scans were compared to the 'Reference' scan other than the modified scans with reduced FOV and altered convolution kernel, which were compared to the 'ThinSlice' scan and indicated by the asterisk.

are shown in Figure 2.4.

The relative number of features reflecting significant differences when each of the various harmonization methods are used is shown in Figure 2.5. Histogram normalization (darker blue bars in Figure 2.5) reduced or maintained the number of features sensitive to differences in all image acquisition and reconstruction parameters. In comparison, resampling combined with Butterworth filtering (orange bars in Figure 2.5) increased the number of features sensitive to differences in tube voltage and slice interval, but reduced the number of features sensitive to differences in iDose level to zero. ComBat harmonization reduced the number of sensitive features to zero for all image acquisition and reconstruction parameters.

To assess the effect of the harmonization methods on each feature category, Table 2.2 illustrates the relative number of features from each category reflecting significant differences

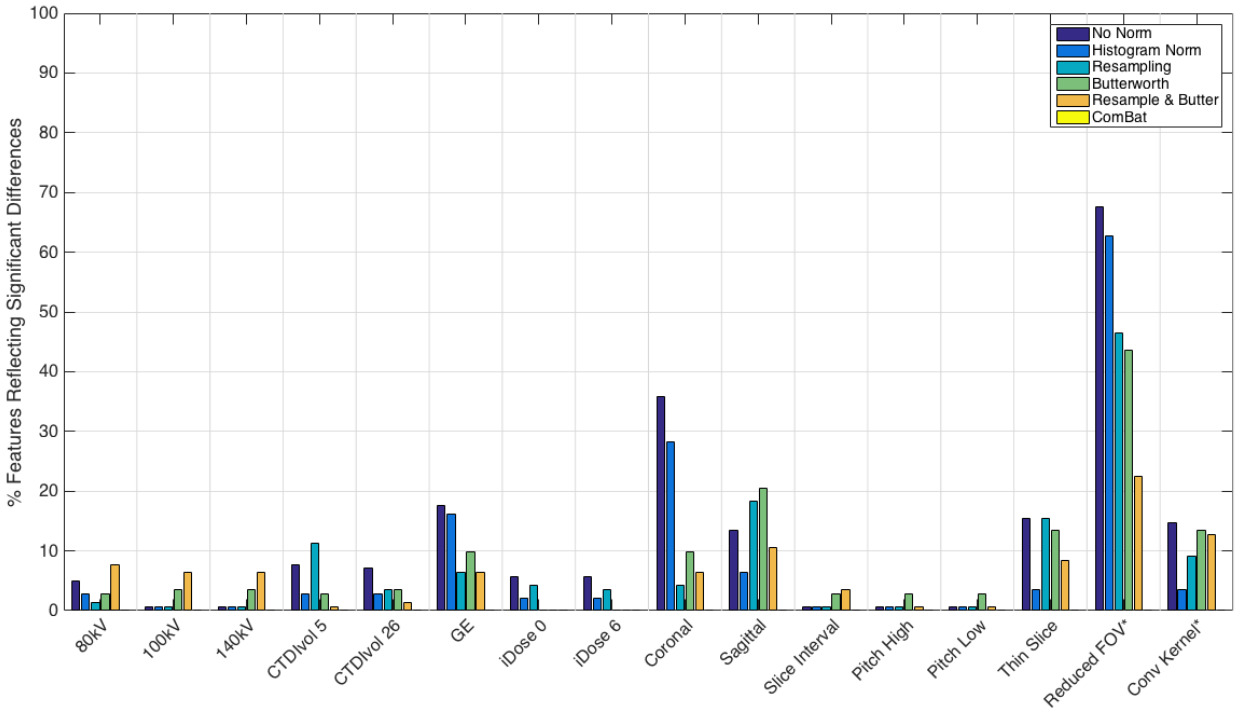


Figure 2.5: Relative number of features reflecting significant differences between the modified scans and the corresponding reference scan when each of the harmonization methods was implemented. All scans were compared to the 'Reference' scan other than the modified scans with reduced FOV and altered convolution kernel, which were compared to the 'ThinSlice' scan and indicated by the asterisk. Note, no yellow bars are shown because ComBat harmonization resulted in none of the features reflecting significant differences.

before and after harmonization.

Table 2.2: The relative number of features reflecting significant differences before and after each of the harmonization methods was used for first-order (F-O), GLCM, fractal, Fourier, and Laws' filter features. Green cells indicate fewer features reflecting significant differences when a particular harmonization method was used, while red cells indicate more features reflecting significant differences.

	No Harmonization						Histogram Normalization					
Protocol Name	*F-O*	*GLCM*	*Fractal*	*Fourier*	*Laws'*	*Total*	*F-O*	*GLCM*	*Fractal*	*Fourier*	*Laws'*	*Total*
80kV	27.3%	0.0%	0.0%	5.9%	0.0%	4.9%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%
100kV	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%
140kV	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%
CTDIvol5	45.5%	0.0%	0.0%	5.9%	0.0%	7.7%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%
CTDIvol26	40.9%	0.0%	0.0%	5.9%	0.0%	7.0%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%

Table 2.2 Continued: The relative number of features reflecting significant differences before and after each of the harmonization methods was used for first-order (F-O), GLCM, fractal, Fourier, and Laws' filter features. Green cells indicate fewer features reflecting significant differences when a particular harmonization method was used, while red cells indicate more features reflecting significant differences.

GE	63.6%	64.3%	20.0%	5.9%	0.0%	17.6%	54.5%	64.3%	20.0%	5.9%	0.0%	16.2%
PitchHigh	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%
PitchLow	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%
iDose 0	36.4%	0.0%	0.0%	0.0%	0.0%	5.6%	13.6%	0.0%	0.0%	0.0%	0.0%	2.1%
iDose 6	36.4%	0.0%	0.0%	0.0%	0.0%	5.6%	13.6%	0.0%	0.0%	0.0%	0.0%	2.1%
Coronal	0.0%	50.0%	0.0%	47.1%	42.9%	35.9%	0.0%	50.0%	0.0%	47.1%	29.8%	28.2%
Sagittal	54.5%	21.4%	0.0%	5.9%	3.6%	13.4%	22.7%	21.4%	0.0%	5.9%	0.0%	6.3%
Slice Interval	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%
ThinSlice	63.6%	35.7%	0.0%	0.0%	2.4%	15.5%	22.7%	0.0%	0.0%	0.0%	0.0%	3.5%
ThinSlice_ReducedFOV	59.1%	78.6%	80.0%	52.9%	70.2%	67.6%	27.3%	78.6%	80.0%	64.7%	67.9%	62.7%
ThinSlice_ConvKernel	68.2%	0.0%	40.0%	5.9%	4.8%	14.8%	13.6%	0.0%	20.0%	5.9%	0.0%	3.5%
	Resampling						Butterworth Filtering					
Protocol Name	*F-O*	*GLCM*	*Fractal*	*Fourier*	*Laws'*	*Total*	*F-O*	*GLCM*	*Fractal*	*Fourier*	*Laws'*	*Total*
80kV	4.5%	0.0%	0.0%	5.9%	0.0%	1.4%	9.1%	0.0%	0.0%	5.9%	1.2%	2.8%
100kV	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	4.5%	0.0%	0.0%	5.9%	3.6%	3.5%
140kV	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	4.5%	0.0%	0.0%	5.9%	3.6%	3.5%
CTDIvol5	31.8%	57.1%	0.0%	5.9%	0.0%	11.3%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%
CTDIvol26	18.2%	0.0%	0.0%	5.9%	0.0%	3.5%	18.2%	0.0%	0.0%	5.9%	0.0%	3.5%
GE	36.4%	0.0%	0.0%	5.9%	0.0%	6.3%	27.3%	0.0%	20.0%	5.9%	7.1%	9.9%
PitchHigh	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%
PitchLow	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%
iDose 0	22.7%	7.1%	0.0%	0.0%	0.0%	4.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iDose 6	22.7%	0.0%	0.0%	0.0%	0.0%	3.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Coronal	0.0%	0.0%	20.0%	29.4%	0.0%	4.2%	0.0%	0.0%	0.0%	5.9%	15.5%	9.9%
Sagittal	18.2%	42.9%	0.0%	5.9%	17.9%	18.3%	4.5%	0.0%	0.0%	5.9%	32.1%	20.4%
Slice Interval	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	13.6%	0.0%	0.0%	5.9%	0.0%	2.8%
ThinSlice	36.4%	57.1%	20.0%	5.9%	4.8%	15.5%	18.2%	7.1%	0.0%	5.9%	15.5%	13.4%
ThinSlice_ReducedFOV	22.7%	57.1%	40.0%	29.4%	54.8%	46.5%	27.3%	35.7%	80.0%	52.9%	45.2%	43.7%

Table 2.2 Continued: The relative number of features reflecting significant differences before and after each of the harmonization methods was used for first-order (F-O), GLCM, fractal, Fourier, and Laws' filter features. Green cells indicate fewer features reflecting significant differences when a particular harmonization method was used, while red cells indicate more features reflecting significant differences.

ThinSlice_ConvKernel	22.7%	57.1%	0.0%	0.0%	0.0%	9.2%	9.1%	7.1%	0.0%	0.0%	19.0%	13.4%
	Resampling and Filtering						ComBat					
Protocol Name	*F-O*	*GLCM*	*Fractal*	*Fourier*	*Laws'*	*Total*	*F-O*	*GLCM*	*Fractal*	*Fourier*	*Laws'*	*Total*
80kV	31.8%	0.0%	20.0%	5.9%	2.4%	7.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
100kV	31.8%	0.0%	20.0%	5.9%	0.0%	6.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
140kV	31.8%	0.0%	20.0%	5.9%	0.0%	6.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CTDIvol5	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CTDIvol26	4.5%	0.0%	0.0%	5.9%	0.0%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
GE	31.8%	0.0%	20.0%	5.9%	0.0%	6.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PitchHigh	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PitchLow	0.0%	0.0%	0.0%	5.9%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iDose 0	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iDose 6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Coronal	27.3%	0.0%	20.0%	5.9%	1.2%	6.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Sagittal	27.3%	0.0%	20.0%	5.9%	8.3%	10.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Slice Interval	9.1%	0.0%	0.0%	5.9%	2.4%	3.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ThinSlice	0.0%	0.0%	0.0%	5.9%	13.1%	8.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ThinSlice_ReducedFOV	27.3%	0.0%	40.0%	29.4%	22.6%	22.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ThinSlice_ConvKernel	9.1%	7.1%	0.0%	0.0%	17.9%	12.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

As shown in Table 2.2., first-order features were the most affected by histogram normalization, resampling, filtering, or resampling and filtering combined. Histogram normalization increased the number of sensitive features for Fourier features when comparing the 'ThinSlice_ReducedFOV' scan to the 'ThinSlice' scan, while resampling, filtering, and resampling and filtering combined showed mixed results among all feature categories and imaging parameters. No feature from any category reflected significant differences after ComBat harmonization.

2.4 Discussion

The results from this study illustrate that radiomic features extracted from CT scans of a cadaveric liver can greatly vary when these scans are acquired and reconstructed with different parameters; however, the degree of this variability is highly dependent on the parameters that are altered. The percentage of features that reflected significant differences are shown in Figure 2.3. Reducing the FOV and using coronal instead of axial views resulted in the greatest differences in features with 67.6% and 35.9% of features reflecting significant differences, respectively. Slight differences in tube voltage, slice interval, and pitch resulted in the smallest discrepancy in features with changes in any of these parameters resulting in 0.7% of features reflecting significant differences. This is in part due to the similarity in the pixel value distributions shown in Figure 2.2. The dependence of these features on each parameter also varied with feature category. For example, while only about 7% of features were sensitive to differences in CT dose index (CTDIvol), about 90% of these sensitive features were first-order features (as shown by the dark blue component in Figure 2.3). The remaining feature categories were relatively robust to differences in CTDIvol. Given that CTDIvol, an estimation of patient dose, greatly affects image noise, it is not surprising that first-order features vary with this parameter.

Additionally, only first-order features were sensitive to changes in iDose level, a Philips iterative reconstruction technique that focuses on quantum noise reduction. In comparison, all feature categories were about equally sensitive to changes in FOV as shown in the equal distribution of colors in Figure 2.3. Because decreased FOV results in smaller pixels without a comparable increase in tube current, the image noise was increased. In addition, the anatomy is distributed over a greater number of pixels, and the combination of increased noise and altered pixel distribution affected several first-, second-, and higher-order features [42,54]. While the coronal and sagittal slice orientation also resulted in a large percentage of sensitive features, 35.9% and 13.4%, respectively, the distribution of features varied among first-order, GLCM, Fourier, and Laws' filter features. It is expected that changing the slice

orientation will affect the features because not only does the anatomy contained within each slice differ, but the pixel dimensions differ in the axial, sagittal, and coronal planes for each of the three scans (see Figure 2.1). Additionally, first-order features were affected by the majority of imaging parameters (i.e., 10 of 16), whereas the other feature categories were more robust. GLCM and Laws' filter features were both affected in five comparisons, while fractal features were affected in three comparisons. Fourier features were affected in 13 of the 16 comparisons; however, for 11 of these comparisons, only one Fourier feature (6% of Fourier features calculated) reflected significant differences.

While changes in image acquisition and reconstruction parameters affected the number of features that reflected significant differences between reference and modified scans, these differences typically did not substantially change the feature values themselves as shown in Figure 2.4. The majority of features differed by less than 1% for 4 of the 16 comparisons (CTDIvol5, iDose 0 and 6, and ThinSlice_ConvKernel), and only changing the view from axial to coronal or sagittal and reducing the FOV resulted in median relative differences greater than 5%. Comparing the relative number of features reflecting significant differences (Figure 2.3) and the relative differences in feature values between reference and modified scans (Figure 2.4) illustrates that differences in imaging parameters may result in consistent biases in feature values between these scans; however, these biases may be small relative to the feature values themselves. For example, when calculating the difference in first-order entropy between ThinSlice_ReducedFOV and ThinSlice scans, the average difference in entropy across slices was 0.161 ± 0.142. This difference indicates that, while the bias in this first-order entropy between these two scans is small (about one standard deviation on average), the bias is consistent across scans. This trend is expected given reducing the FOV decreases the sampling statistics allocated to each pixel, which increases the quantum noise within the image. Because first-order entropy is reflective of image noise, it is not surprising that decreasing the FOV systematically increases the value of this feature across slices.

While the values of most features were not greatly affected by each imaging parame-

ter, other features differed by up to 170% (Figure 2.4). In these instances, features that quantify image noise and pixel value distribution such as first-order standard deviation and entropy were more susceptible to imaging parameters that affected the image noise (e.g., iDose reconstruction level, CTDIvol, and convolution kernel).

The harmonization methods had varying effects when mitigating differences in feature values between modified and reference scans as shown in Figure 2.5 and Table 2.2. Combining voxel size resampling and Butterworth filtering increased the number of features sensitive to differences in tube voltage. Conversely, the same harmonization method reduced the number of sensitive features to zero when comparing scans obtained with different iDose levels. For all scans that had more than 1% of features reflecting significant differences before harmonization, histogram normalization reduced the number of sensitive features. This discrepancy indicates that different methods may be better suited to harmonize different feature categories. For example, resampling combined with Butterworth filtering was initially utilized to mitigate the effects of differing pixel sizes on radiomic features, so this method is not necessarily expected to harmonize features for scans acquired with different tube voltages or slice interval; however, resampling and filtering reduced the number of features sensitive to differences in FOV by 64.4% supporting results previously reported in the literature [46]. In addition, histogram normalization had the greatest effect on first-order features (Table 2.2), while the remaining feature categories were relatively unchanged. Consequently, this method would be best suited to mitigate the effects of imaging parameters that primarily alter first-order features such as differences in iDose reconstruction levels or CTDIvol.

In comparison to other harmonization methods, ComBat harmonization reduced the number of features reflecting significant differences to zero regardless of the parameter changed. Orlhac et al. [64] extracted 40 radiomic features from CT scans of a phantom composed of various materials such as wood, cork, rubber particles, and plastics of various densities. This study reported that nearly all features extracted from any material were sensitive to differences in imaging parameters such as reconstruction kernel, pitch, and

tube current. When ComBat harmonization was used, none of the features reflected significant differences for all phantom materials, validating the results found in the current study when human tissue was used. Previous studies have partially quantified the dependence of radiomic features on image acquisition and reconstruction parameters and the effects of individual harmonization methods. In contrast, this study compares the variability of radiomic features across parameters and the degree of harmonization due to several harmonization methods using CT scans of human tissue. The current study also illustrates that ComBat harmonization can mitigate differences in feature values when these features are extracted from images of human tissue acquired with different CT image acquisition and reconstruction parameters.

This study contained several limitations. First, only one scan was acquired for each acquisition and reconstruction parameter. Each slice was used as a distinct analyzable unit, and because the various scans often contained a different number of slices, paired statistical comparisons could not be made. For example, to compare the agreement in feature values between scans, the average feature value across slices was compared. The concordance correlation coefficient (CCC) would have compared correlation among features from multiple scans and would have been more powerful than comparing the relative differences in feature values. Additionally, while the current study quantifies how many features reflect significant differences between image acquisition and reconstruction parameters, we cannot determine whether these differences are clinically significant when applied to a particular clinical task. This work illustrated that changing the iDose level affects a number of first-order features when images of the same liver are acquired; however, these differences may not be clinically relevant when applied to a detection or diagnosis task using an ensemble of patients. If the variability in feature values due to differences in iDose level is much less than the variability in feature values due to differences in patient anatomy, then biases introduced by differences in imaging parameters may be inconsequential. Additionally, assessing the changes in texture over time (i.e., delta radiomics) while controlling for patient variability has been shown to

remove the biases introduced by differences in radiomic software and may also mitigate differences in texture due to variable imaging parameters [37].

An additional limitation is that scans were acquired with changes in more than a single parameter and retrospectively analyzed such as those for the 'ThinSlice_ReducedFOV' and 'ThinSlice_ConvKernel' protocols. In the future, it would be useful to repeat the scans but isolate changes to a single parameter in each scan. Future studies should also investigate whether the results reported here are generalizable to other tissue types and imaging modalities. Mackin et al. [35] showed that radiomic features reflected greater dependence on tube current when features were extracted from CT scans of more homogeneous tissues similar to liver tissue. Therefore, the dependence of radiomic features and the potential of each harmonization method may change when different tissues or materials are analyzed

Finally, ComBat harmonization altered the feature values such that none of the features reflected significant differences between reference and modified scans. While this result is supported by the literature, future studies should determine the impact of Combat harmonization and the other harmonization methods on classification ability when applied to a detection or diagnosis task. Histogram normalization modifies the pixel values in an ROI such that the minimum and maximum pixels are the same for all ROIs regardless of disease status. Consequently, this harmonization method negates the utility of these features for any clinical application because they were be made the same for all normalized images. In Chapter 5, the effect of ComBat harmonization on classification performance is assessed; however additional studies should investigate whether the remaining harmonization methods outlined in this chapter affect classification ability.

2.5 Conclusion

This study investigated the effects of several image acquisition and reconstruction parameters on radiomic features from various feature categories and compared five methods to mitigate these effects. Reducing the FOV and changing the image reconstruction plane

from axial to coronal resulted in the greatest number of significantly different features, while changes in tube voltage, pitch, and slice interval resulted in the least. Histogram normalization was shown to reduce or maintain the number of sensitive features for all image acquisition and reconstruction parameters, while voxel size resampling and Butterworth filtering reduced the number of sensitive features for 10 out of 16 comparisons. ComBat harmonization eliminated the number of sensitive features for all image imaging parameters. The current study illustrates that, while radiomic features are dependent on image acquisition and reconstruction parameters, these dependencies may be mitigated using the appropriate harmonization method to control for the effects of variable imaging parameters. With these harmonization techniques, particularly ComBat harmonization, future radiomics studies may be more reproducible and allow for larger databases of potentially disparate patient data, resulting in the greater translation of radiomics research into clinical practice.

CHAPTER 3

VARIATION IN ALGORITHM IMPLEMENTATION ACROSS RADIOMICS SOFTWARE

3.1 Introduction

The previous chapter assessed the variability in radiomic features introduced by differences in the image acquisition process. Extrapolating on this, the effect of the radiomic feature calculation process on radiomic features was also assessed. Because of the growing promise radiomics has shown and the large amounts of imaging data available, investigators have become more interested in quantifying texture to increase the amount of information extracted from medical images and to limit variability among radiologists [1,2,24,25,65-73]. Many research groups have developed in-house and freely-available radiomics software packages to allow for the advancement of radiomics research. These packages, however, are often used with a "one-size-fits-all" approach without considering the underlying mechanisms embedded in the algorithms that may result in variations among packages. Such variations could be caused by differences in image preprocessing, differences in the algorithms used to calculate features, or differences in algorithm implementation. Additionally, radiomics packages are often used to analyze images of one specific imaging modality, anatomic location, or tissue type, when the software used was designed to analyze another type of image. Radiomic features have been shown to vary substantially based on differences in image acquisition parameters, reconstruction algorithms, and gating techniques, and these differences may be exacerbated when computed with different packages [2,30,36,41,53,56,74-78].

A number of studies have noticed this lack of harmonization among radiomics research and have called for greater standardization [38-44,79,80]. Some of these studies investigated the variability in radiomics research due to the differences in the feature calculation process and attempted to standardize this process to obtain more reproducible radiomic features. However, complete standardization has not fully obtained, and these studies were hindered

by using inadequate medical images to assess their feature calculation software. Therefore, the purpose of this study was to compare two in-house radiomics software packages and three freely-available software packages using clinical images of various anatomic regions and imaging modalities. This study also aimed to determine the sources of these variations among the radiomics software analyzed.

3.2 Methods and Materials

3.2.1 Imaging Data

Cranial-caudal (CC) digital mammography, head and neck (HN) computed tomography (CT) images, and contrast-enhanced T2-weighted breast magnetic resonance imaging (MRI) scans were obtained through the Human Imaging Research Office (HIRO) under institutional review board approval [81]. The breast MRI database consisted of two serial scans for each patient acquired between 6 and 12 months apart. Image parameters and patient information are shown in Table 3.1. For the mammography and HN CT scans, pixel information was extracted from a single region of interest (ROI) in each image. Mammography ROIs (256×256 pixels) contained normal breast parenchyma, while HN CT ROIs contained manually segmented tumor (mean number of pixels: 1102; range: 174-2819). For breast MRI scans, ROIs (32×32 pixels) were placed in a single slice in the first image acquisition, and corresponding ROIs were anatomically matched in the second image acquisition using common landmarks such that ROIs contained similar anatomy for both image acquisitions. Example ROIs from each database are shown in Figure 3.1.

Table 3.1: Patient and scan characteristics.

	Mammography	Head and Neck CT
Number of Patients	40	39
Number of Scans	40	39
Mean Pixel Spacing (range) (mm)	0.1 (0.1 - 0.1)	0.536 (0.424 - 0.688)
Scanner Manufacturer	GE Senographe 2000D	Philips Brilliance (n = 35) Philips iCT 256 (n = 3) Siemens Biograph 64 (n = 1)
	BreastMRI	
	Time Point 1	Time Point 2
Number of Patients	45	45
Number of Scans	45	45
Mean Slice Thickness (range) (mm)	2.0 (2.0 - 2.0)	1.98 (1.6 - 2.0)
Mean Repetition Time (range) (sec)	2.0 (2.0 - 2.0)	2.0 (2.0 - 2.0)
Mean Echo Time (range) (msec)	379.8 (367.9 - 396.5)	372.6 (210.0 - 400.1)
Mean Field Size (range) (pixels)	434×434 (384×384 - 480×480)	435×435 (384×384 - 480×480)
Mean Pixel Size (range) (mm)	0.766 (0.708 - 0.787)	0.769 (0.708 - 0.792)

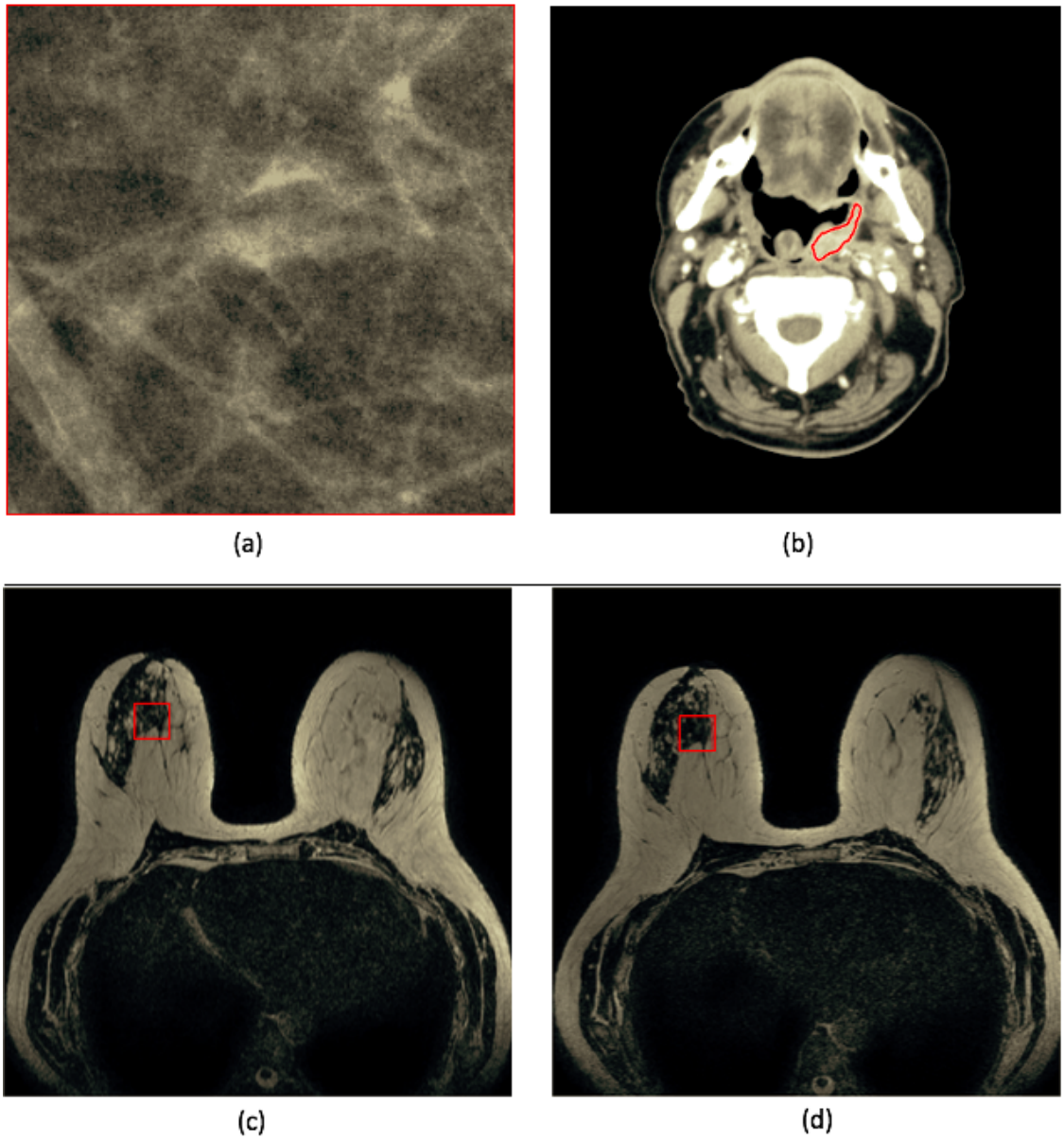


Figure 3.1: Example ROIs depicting a 256×256-pixel mammography ROI (a) and a head and neck ROI containing contoured tumor (b). An example depicting a 32×32-pixel breast MRI ROI placed in the first image acquisition (c) and anatomically matched in the second image acquisition (d).

3.2.2 Radiomics Software

Five radiomics software packages were utilized for this study. Two packages had been developed in-house by independent research labs at the University of Chicago (A1 and A2)

Table 3.2: Number of directionally-independent features per feature category that can be calculated by each radiomics package.

Feature Category	A1	A2	IBEX	MaZda	Pyradiomics
Shape			18	73	10
First-Order Histogram	22	18	24	9	19
Intensity Histogram Gaussian Fit			5		
Absolute Gradient				5	
Run-Length Matrix			11	5	16
Neighborhood Intensity Difference			5		5
Co-Occurrence Matrix	22	14	22	11	24
Autoregressive Model Parameters				5	
Wavelet				20	
Fractal	5	25			
Fourier	17	22			
Laws'	84				

[53,81-84], and three were freely-available packages: MaZda v4.6 (Institute of Electronics, Technical University of Lodz, Poland) [85-88], IBEX v1.0 beta (The University of Texas MD Anderson Cancer Center) [89], and PyRadiomics v2.0.0 [90]. The three packages from outside our institution were chosen because they were freely available at the initiation of this study, and they had been cited in a number of recent publications. Each package was capable of calculating several classes of radiomic features including first-order histogram features, fractal features, Fourier features, and gray-level run length matrix features; however, only first-order histogram features and second-order gray-level co-occurrence matrix (GLCM) features were common among all five packages as shown in Table 3.2.

In an abstract from the IBSI, the union of all features across all packages was compared [43,44]. In the current study, only the features shared by all five packages with the same naming conventions were compared. When comparing breast MRI features, relative differences in feature values were calculated (i.e., delta radiomics) between image acquisitions for

each patient using the first image acquisition as the reference:

$$\Delta FV_{f,p} = \frac{FV_{2,f,p} - FV_{1,f,p}}{FV_{1,f,p}} \tag{3.1}$$

Here, $FV_{1,f,p}$ and $FV_{2,f,p}$ are the feature values for feature f from patient p for the first and second image acquisitions, respectively. The features common among all five packages are shown in Table 3.3, and definitions for each feature can be found in the appendix. Because many breast MRI ROIs had minimum pixel values of zero, the relative difference in the "minimum" feature between time points was undefined. To bypass this issue, the difference between maximum and minimum pixel values was calculated, and the range was used to assess the breast MRI database, while minimum was used to assess the mammography and HN CT databases.

Table 3.3: First- and second-order radiomic features common among all five packages. First-order minimum was used in assessing the mammography and head and neck CT databases, while first-order range was used in assessing the breast MRI database.

First-Order Histogram Features	**Second-Order GLCM Features**
Maximum	Entropy
Minimum/Range	Contrast
Mean	Sum Average
Standard Deviation	Sum Variance
Skewness	Sum Entropy
Kurtosis	Difference Entropy

3.2.3 Sources of Feature Variation: GLCM Parameters

GLCM features first were calculated using the package-specific default GLCM parameters. These parameters included the gray-level limits, the dimensions of the GLCM, and the directions used in the final average of the GLCM feature values. The distance between neighboring pixel values was 1 for all packages, and the GLCM features were all normalized

Table 3.4: Package-specific default GLCM parameters and GLCM parameters that were modified to maximize consistency among radiomics packages.

Package-Specific Default GLCM Parameters					
GLCM Parameter	*A1*	*A2*	*IBEX*	*MaZda*	*Pyradiomics*
Gray-Level Limits	[-1500,1500]	[Min,Max]	[Min,Max]	[1,4096]	[Min,Max]
Number of Gray Levels	3001	64	(Max PV - Min PV)	256	$\frac{\text{Max PV - Min PV}}{25}$
Number of Directions	4	4	8	4	4
Consistent* GLCM Parameters					
GLCM Parameter	*A1*	*A2*	*IBEX*	*MaZda*	*Pyradiomics*
Gray-Level Limits	[Min,Max]	[Min,Max]	[Min,Max]	[1,4096]†	[Min,Max]
Number of Gray Levels	64	64	64	64	64
Number of Directions	4	4	8	4†	4

*Parameters were modified to maximize consistency across packages.
†Note: These parameters could not be modified.
PV = Pixel Value

by the number of pixels within the ROI. Next, these parameters were modified to allow for the greatest possible consistency among packages. The package-specific default along with the consistent GLCM parameters are shown in Table 3.4. Due to limitations in the customizability of the MaZda interface, the gray-level limits and the number of directions could not be modified to match the other packages. Features were calculated by each software package for each ROI of each image.

3.2.4 Sources of Feature Variation: Algorithm Implementation

Differences in algorithm implementation and ROI processing were investigated to determine sources of variation among software packages. While the underlying mathematical equations used in all software packages are expected to be computationally the same, the interpretation and implementation of these formulas may vary from one package to another.

Packages A2, MaZda, IBEX, and Pyradiomics cited Haralick [14], while A1 cited publications by Felipe [91] and the *Handbook of Computer Vision Applications* [15] for the equations used for feature calculation; however, the algorithmic implementation of these equations may vary due to differences in notation, equation representation, and implementation strategies. Furthermore, the *Handbook of Computer Vision Applications* cites the Haralick paper for its equations after modifying the notation and imposing some corrections and conditions on the equations used. The source code for all packages other than MaZda was investigated for differences between algorithm implementation. MaZda was not open source, so the source code could not be analyzed. Pyradiomics is also the only package of the five held in compliance with the feature definitions outlined by the IBSI.

To remove the effects of ROI preprocessing and GLCM parameter variability for each package, feature functions were extracted from the two in-house packages as well as IBEX and Pyradiomics such that only the functions used to calculate the individual features were investigated. These functions were used to calculate features directly on a single mammographic ROI. Individual feature values were compared across packages to determine differences in algorithm implementation. These equations were extracted after GLCM construction such that differences in the GLCMs across packages did not affect the resultant feature values.

3.2.5 Statistical Analysis

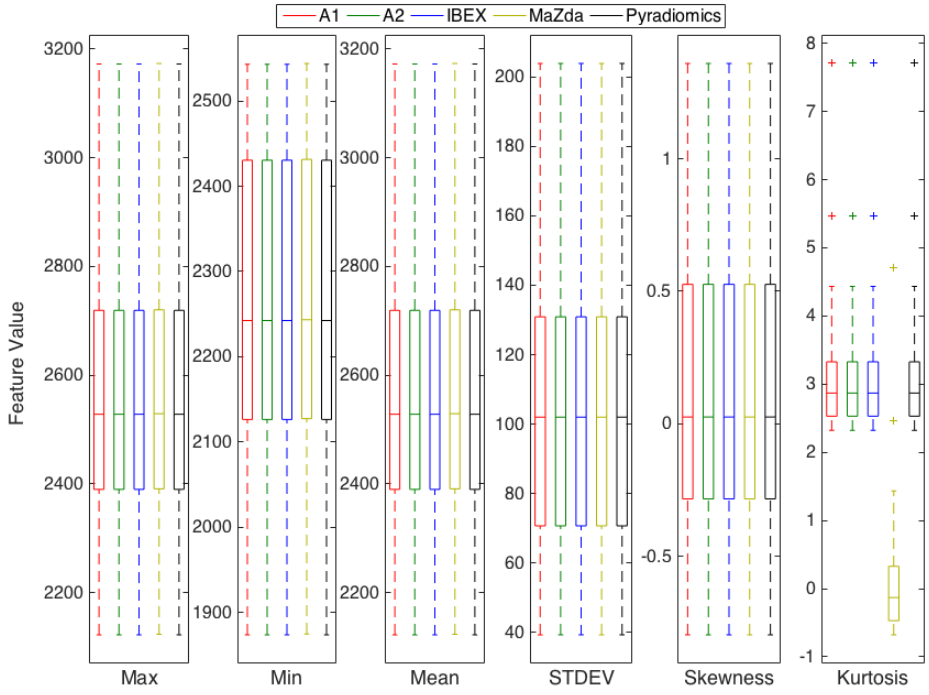
After determining that data were not normally distributed using the Shapiro-Wilk test, nonparametric repeated measures Friedman tests were used to test for significance among features calculated with each software package. The null hypothesis was that all features calculated using each software packages were the same and sampled from the same population. Significance was assessed at the $\alpha = 0.05$ level using Bonferroni correction to account for the 12 features evaluated ($p<0.00417$).

The intraclass correlation coefficient (ICC) was used to assess the agreement of radiomic feature values among packages with package-specific and consistent GLCM parameters using

the two-way mixed effect model illustrating the absolute agreement of the feature values across packages (i.e., ICC(A,1)) [92]. The ICC quantifies the absolute agreement between sets of data by comparing the variability in feature values across software packages to the variability in values across patients. ICC values are stratified to indicate "excellent" (ICC > 0.9), "good" ($0.9 \geq$ ICC > 0.75), "moderate" ($0.75 \geq$ ICC > 0.5), or "poor" (ICC ≤ 0.5) agreement [93].

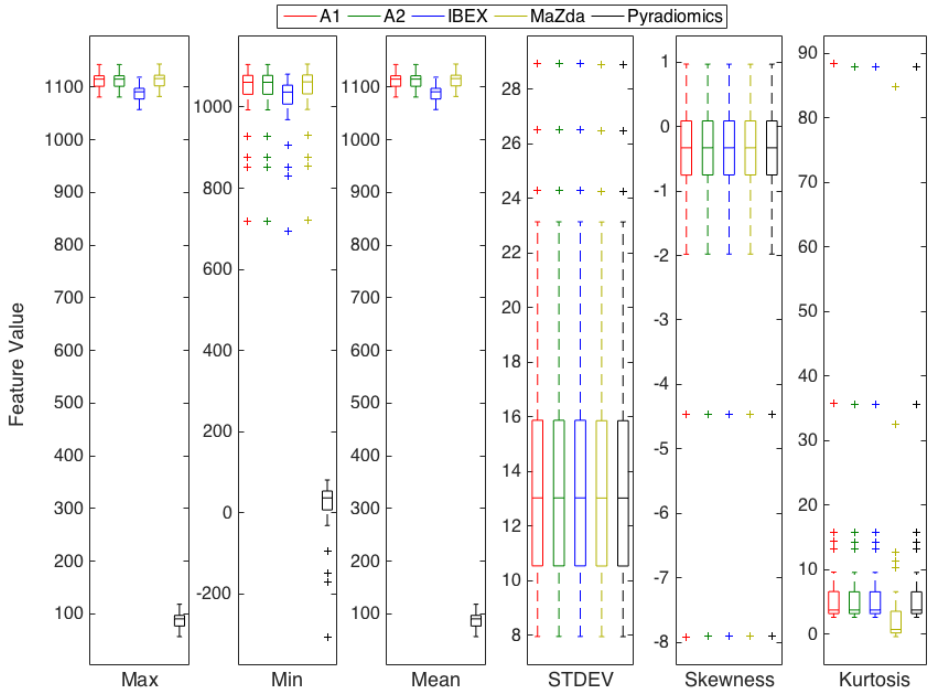
3.3 Results

First-order gray-level features and second-order GLCM features were generated using the 40 mammography ROIs, 39 HN CT ROIs, and 45 breast MRI ROI pairs as input for each of the five radiomics software packages. Boxplots depicting the distributions of the calculated features among all five packages are shown in Figure 3.2.

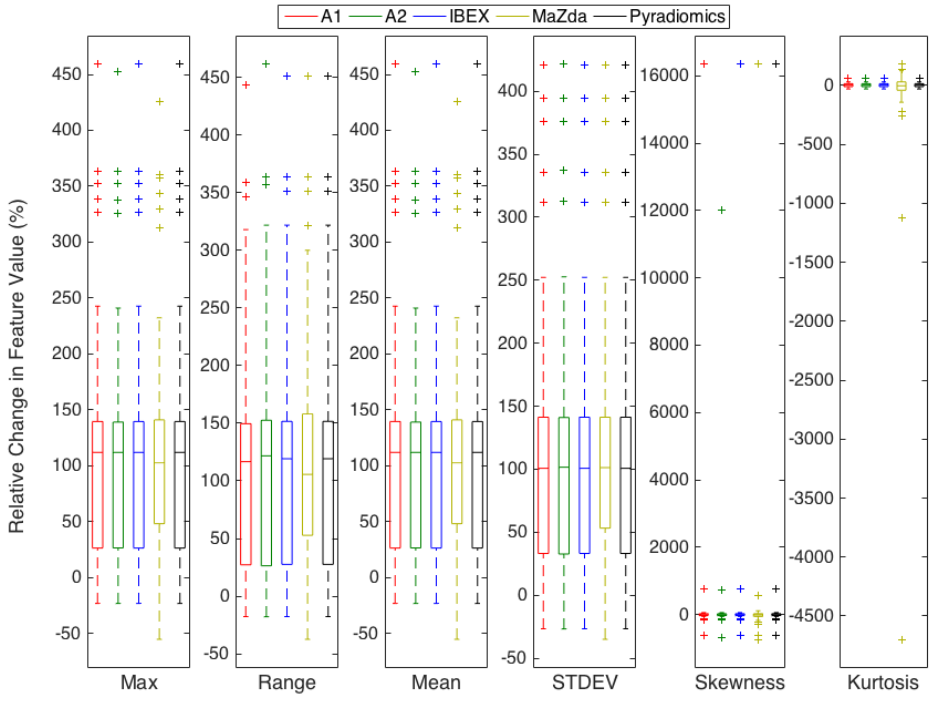


(a) Mammography

Figure 3.2: Distribution of first-order features calculated on the mammography (a), HN CT (b), and breast MRI ROIs (c). Boxes extend from the first to the third quartile with the median represented by the centerline. Outliers are indicated by +.



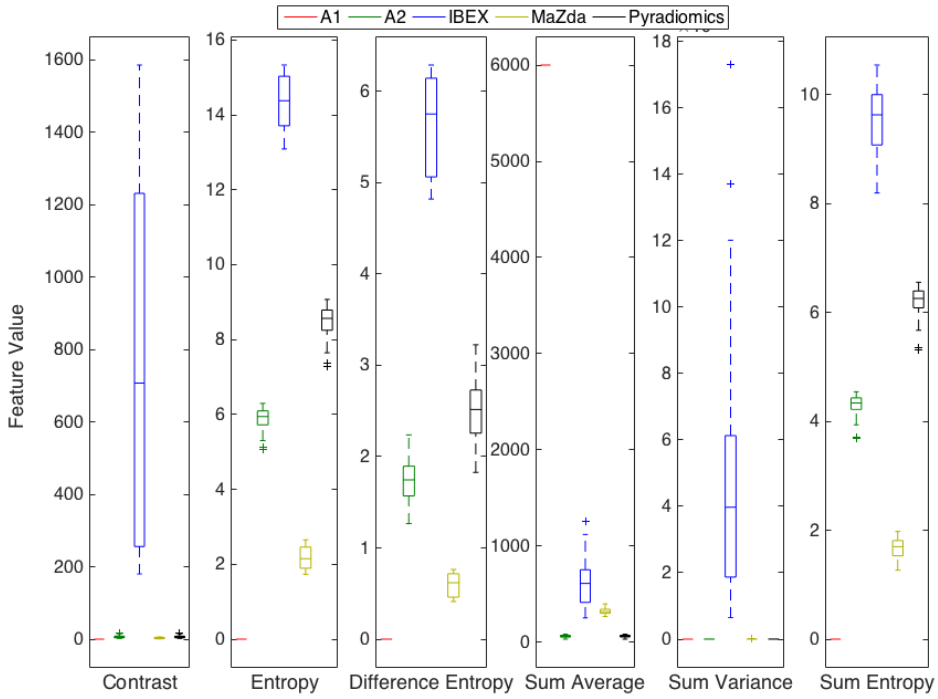
(b) Head and Neck CT



(c) Breast MRI

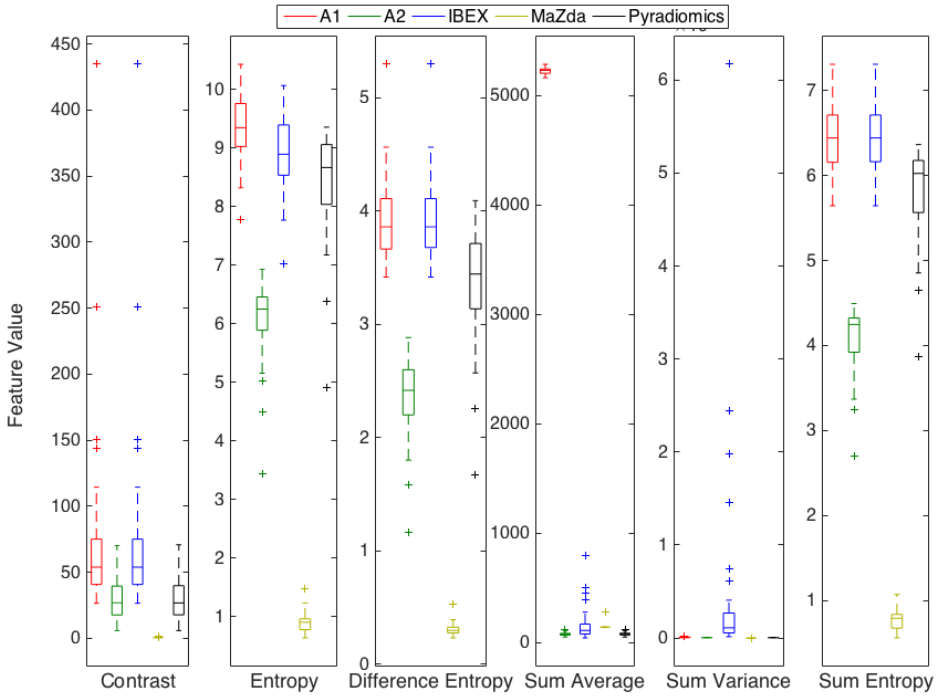
Figure 3.2 Continued: Distribution of first-order features calculated on the mammography (a), HN CT (b), and breast MRI ROIs (c). Boxes extend from the first to the third quartile with the median represented by the centerline. Outliers are indicated by +.

Second-order GLCM features were calculated using the package-specific default GLCM parameters with the feature value distributions shown in Figure 3.3.

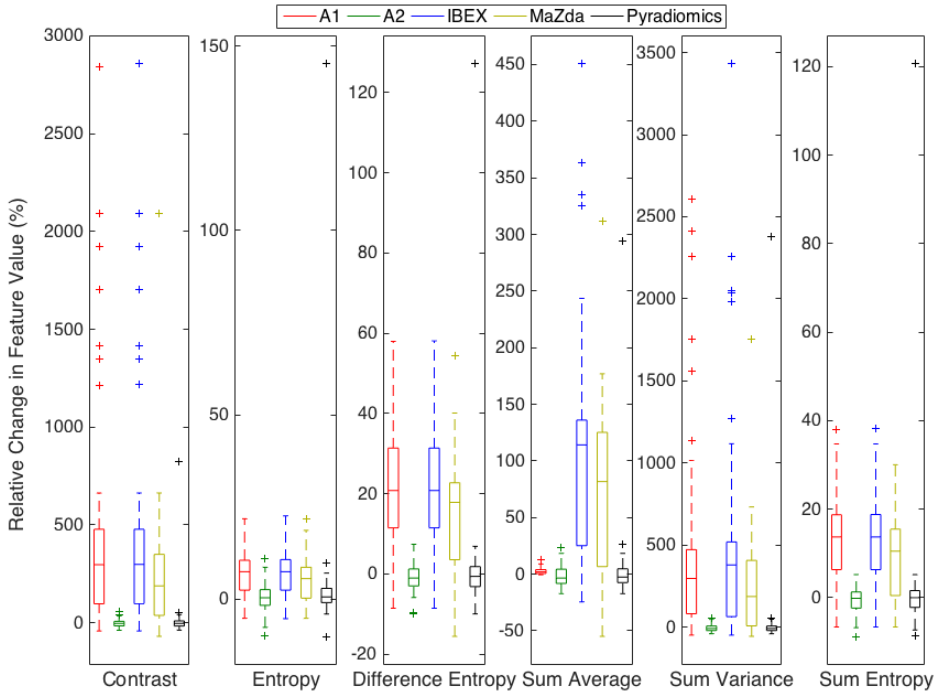


(a) Mammography

Figure 3.3: Distribution of GLCM features calculated on the mammography (a), HN CT (b), and breast MRI ROIs (c) when calculated using the package-specific default GLCM parameters outlined in Table 3.4.



(b) Head and Neck CT



(c) Breast MRI

Figure 3.3 Continued: Distribution of GLCM features calculated on the mammography (a), HN CT (b), and breast MRI ROIs (c) when calculated using the package-specific default GLCM parameters outlined in Table 3.4.

Table 3.5: The p-values resulting from the nonparametric Friedman tests comparing radiomic features across packages, and ICCs illustrating agreement in features among packages. Second-order features were calculated using package-specific default GLCM parameters. Features reflecting significant differences are highlighted in red (p<0.004).

	Mammography		Head and Neck CT		Breast MRI	
Feature	*p-Value*	*ICC*	*p-Value*	*ICC*	*p-Value*	*ICC*
Maximum	<0.004	0.999	<0.004	0.002	<0.004	0.934
Minimum	<0.004	0.997	<0.004	0.027	NA	NA
Range	NA	NA	NA	NA	<0.004	0.935
Mean	<0.004	0.998	<0.004	0.001	<0.004	0.964
Standard Deviation	<0.004	1.000	<0.004	1.000	0.301	0.977
Skewness	0.791	1.000	0.005	1.000	0.907	0.984
Kurtosis	<0.004	0.346	<0.004	0.991	0.514	0.006
GLCM Contrast	<0.004	0.001	<0.004	0.131	<0.004	0.299
GLCM Entropy	<0.004	0.002	<0.004	0.007	<0.004	0.125
GLCM Sum Entropy	<0.004	0.003	<0.004	0.006	<0.004	0.232
GLCM Sum Average	<0.004	<0.001	<0.004	<0.001	<0.004	0.127
GLCM Sum Variance	<0.004	<0.001	<0.004	0.002	<0.004	0.314
GLCM Difference Entropy	<0.004	0.006	<0.004	0.009	<0.004	0.226

ICC values and p-values assessing differences in feature values across packages are shown in Table 3.5. When assessing significant differences in features among software packages, 11 of the 12 features differed significantly for both the mammography and HN CT databases, while 9 of the 12 features differed significantly for the breast MRI database. All second-order features reflected significant differences for all three databases.

While most first-order features showed significant differences for the mammography and HN CT databases, the ICCs for all first-order mammography features besides kurtosis demonstrated excellent agreement. This indicates that while systematic biases were introduced due to differences in each of the packages resulting in significant differences, the

magnitude of these biases are small relative to the feature values themselves. Therefore, the ICC still reflected excellent agreement among packages for these features. Among HN CT ROIs, maximum, minimum, and mean showed poor agreement, whereas the remaining first-order features all showed excellent agreement. All second-order features for mammography, HN CT, and breast MRI ROIs showed poor agreement; however, features extracted from the breast MRI ROIs typically had ICC values that were greater than those from mammography or HN CT ROIs.

Plots showing values for the kurtosis and GLCM entropy across the 39 HN CT ROIs are shown in Figure 3.4, which demonstrate excellent and poor agreement across packages, respectively. In the scatter plot depicting the feature distributions for kurtosis, the differences in feature values among packages for each patient are small relative to the variation in feature values among patients. Additionally, the differences in feature values introduced by differences among radiomics packages were consistent, resulting in significant differences when using Friedman tests; however, because the bias was small compared to the variation among patients, the ICC is close to 1 reflecting excellent agreement. In contrast, the differences in GLCM entropy for each patient across packages are large and systematic, resulting in significant differences and an ICC value near 0 reflecting poor agreement.

3.3.1 Variation in Image Importation and Preprocessing

In both the A1 and A2 packages, the raw images are imported with no preprocessing or normalization applied before features are calculated. MaZda, however, applies a default normalization such that the value of each pixel is increased by one. The MaZda user manual [85] states that this is done to keep consistency with the equations presented in Haralick [94]. Additionally, MaZda was not designed to accommodate negative pixel values, and the resulting pixel values are greatly dependent on whether or not the image was stored as signed or unsigned because of the way signed and unsigned images differ in their bit allocation. In comparison, pixel values in IBEX are not dependent on whether the image was stored as

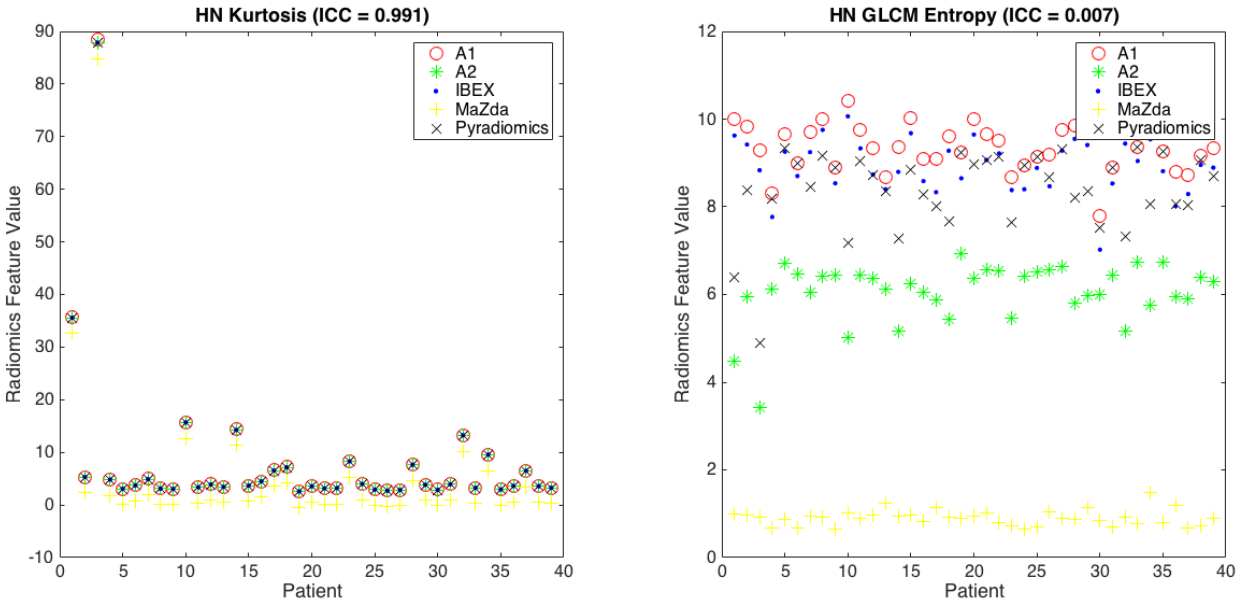


Figure 3.4: Scatter plots illustrating the agreement of features across packages. HN kurtosis showed excellent agreement (ICC = 0.991) because the variability in feature values among packages is much less than the variability in feature values among patients, while HN GLCM entropy showed poor agreement (ICC = 0.007). Because of the consistent bias introduced in the feature distributions, HN kurtosis is still significantly different when calculated using different radiomics packages despite the strong agreement reflected by the ICC for HN kurtosis.

signed or unsigned, but negative pixel values are truncated at zero while non-negative pixel values retain their original value. Additionally, IBEX imports images using the RescaleSlope and RescaleIntercept tags from the DICOM header in the following manner:

$$ImageData = (ImageData) * RescaleSlope + RescaleIntercept + 1000 \qquad (3.2)$$

The RescaleIntercept tag for a standard CT scan typically has a value of -1024, resulting in the value of each pixel in the image being reduced by 24. These trends can be seen in the boxplots for the maximum, minimum, and mean in Figure 3.2.b. Pyradiomics uses the same preprocessing but does not add 1000 to the pixel values, so Pyradiomics automatically adjusts the pixel values from CT scans to Hounsfield units (HU). The features maximum, minimum, and mean are therefore all reduced by a constant value of 1024 also shown in Figure

Table 3.6: Differences in image importation characteristics.

	A1	A2	IBEX	MaZda	Pyradiomics
Imported image dependent on image being signed or unsigned			✓		
Capable of importing negative pixel values	✓	✓			✓
Capable of performing calculations using negative pixel values without manual preprocessing	✓		✓	✓	✓
Capable of performing calculations using negative pixel values with manual preprocessing	✓	✓	✓	✓	✓
Uses DICOM header in preprocessing				✓	✓

3.2.b, and the corresponding ICC values reflect poor agreement. MaZda, in comparison, does not consider any information contained in the DICOM header, resulting in fundamentally different feature values than when analyzed in IBEX. Differences in image importation are summarized in Table 3.6.

3.3.2 Variation in Algorithm Implementation

First- and second-order feature values for the single mammography image when feature functions were extracted from the packages A1, A2, IBEX, and Pyradiomics are shown in Table 3.7.

When calculated by isolating feature functions from preprocessing steps, most features show strong agreement among packages. Sum variance is shown to greatly differ between IBEX and the remaining packages; however, A1 references Jahne et al. [15] for this equation, which incorporates the value of the sum average in its calculation, whereas IBEX references Haralick et al. [14], which instead incorporates the value of the sum entropy. It is stated in

Table 3.7: Feature values for a single mammography image when feature algorithms are extracted from packages A1, A2, IBEX, and Pyradiomics.

Features	A1	A2	IBEX	Pyradiomics
Maximum	3.161×10^3	3.161×10^3	3.161×10^3	3.161×10^3
Minimum	2.123×10^3	2.123×10^3	2.123×10^3	2.123×10^3
Mean	2.546×10^3	2.546×10^3	2.546×10^3	2.546×10^3
Standard Deviation	1.526×10^2	1.526×10^2	1.526×10^2	1.526×10^2
Skewness	4.390×10^{-1}	4.390×10^{-1}	4.390×10^{-1}	4.390×10^{-1}
Kurtosis	2.967×10^1	2.967×10^1	2.967×10^1	2.967×10^1
GLCM Contrast	1.814×10^{12}	1.814×10^{12}	1.814×10^{12}	1.814×10^{12}
GLCM Entropy	-1.889×10^9	-1.309×10^9	-1.889×10^9	-1.889×10^9
GLCM Sum Entropy	-3.104×10^9	-2.152×10^9	-3.104×10^9	-3.104×10^9
GLCM Sum Average	4.271×10^{10}	4.271×10^{10}	4.271×10^{10}	4.271×10^{10}
GLCM Sum Variance	3.044×10^{29}	3.044×10^{29}	1.608×10^{27}	3.044×10^{29}
GLCM Difference Entropy	-3.269×10^9	-2.266×10^9	-3.269×10^9	-3.269×10^9

Jahne et al. [15] that this discrepancy is thought to be a typographical error. Packages A2 and Pyradiomics also reference Haralick et al. [14] for this feature, but both packages use sum average in the calculation of sum variance.

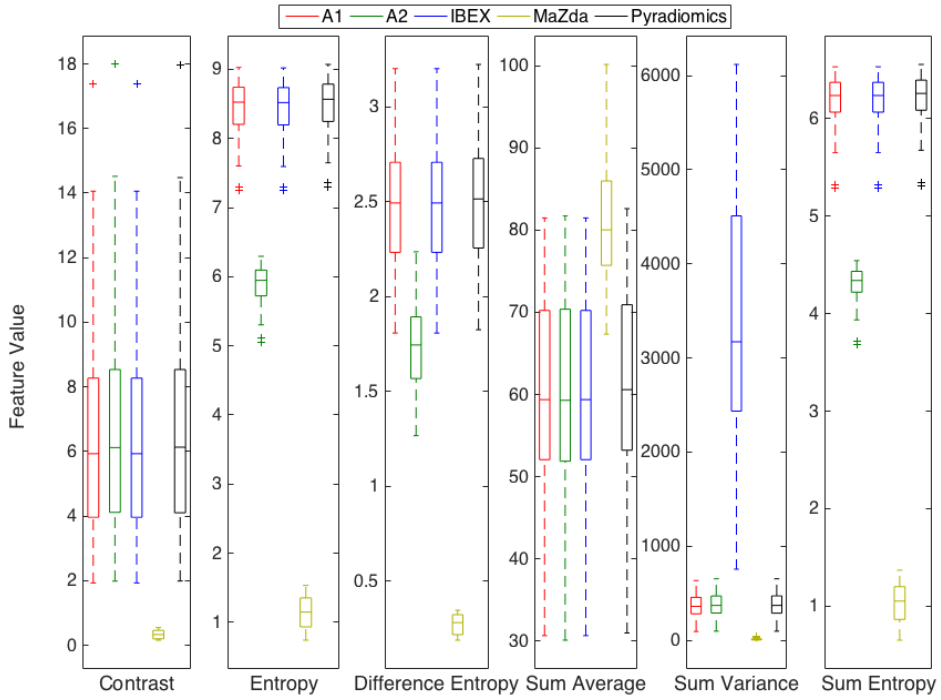
Differences in values for GLCM entropy, sum entropy, and difference entropy between package A2 and the remaining packages arise from different entropy definitions. While A1, IBEX, and Pyradiomics use a logarithm with base 2 in this calculation, A2 uses a natural logarithm. When these features from A2 are scaled by a ratio incorporating the two logarithm bases, the values of entropy, sum entropy, and difference entropy agree with values calculated by the remaining packages to within four significant digits.

3.3.3 Variation in Naming Conventions

While some features with a common name have different algorithmic implementations in different software packages, other features use the same equation (and potentially the same implementation) but are known by a number of different names. The individual features analyzed in this study were those that had common naming conventions among software packages. This feature set might have been larger had common naming conventions been used to describe common mathematical calculations. As an example, the literature, as well as the notes in some published Matlab functions, show that GLCM energy can also be referred to as "uniformity," "uniformity of energy," and "angular second moment" [94,95]. The same Matlab functions refer to GLCM contrast as "variance" or "inertia." Also, the GLCM homogeneity used in A1 was identified to be identical to the inverse difference moment outlined in Haralick [14]; however, the homogeneity in A1 uses the absolute value of the involved differences rather than the square of that difference [94]. Despite the underlying code differing greatly in one package, the computed GLCM absolute value and GLCM difference average feature values were identical for all patients, indicating that these features may be equivalent; however, variations in naming conventions may be difficult to identify when both feature names and algorithm implementation differ among software packages. Pyradiomics outlines several features that are often referred to with different names. For example, GLCM homogeneity 1 and 2 are mathematically identical to GLCM inverse difference and GLCM inverse difference moment, respectively, and GLCM sum variance is also known as GLCM cluster tendency [90]. Finally, it can be seen in Figure 3.2 that the kurtosis calculated by MaZda is exactly three less than the kurtosis calculated by the remaining packages for each ROI. This is because MaZda instead calculates the kurtosis that exceeds that of a Gaussian distribution (i.e., the excess kurtosis), which has a value of about three. This discrepancy in naming convention is not stated explicitly in MaZda's interface.

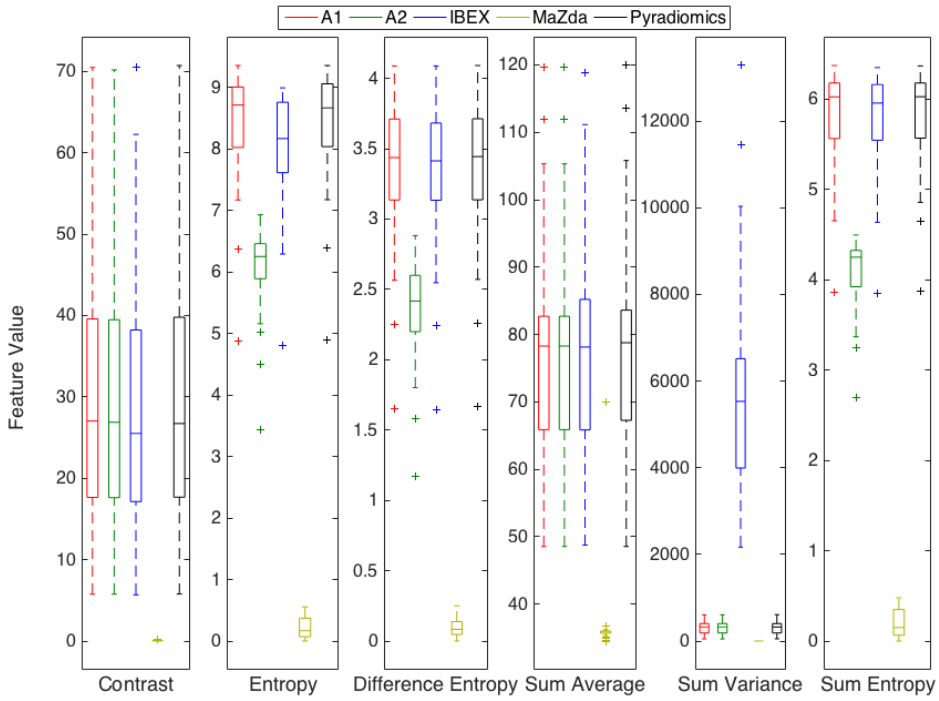
3.3.4 Variation in GLCM Parameters

When GLCM parameters were modified to those shown in Table 3.4 to maximize the consistency among radiomics packages, the resulting feature distributions are shown in Figure 3.5. Compared with the distributions shown in Figure 3.3, the ranges in the feature values are greatly reduced for the mammography, HN CT, and breast MRI images, indicating greater agreement among packages. When using the modified GLCM parameters for all three databases, ICC values typically increased (Table 3.8) compared with those calculated using default GLCM parameters but remained less than 0.5 indicating poor agreement. In addition, all second-order features were still significantly different across packages for all three image databases.

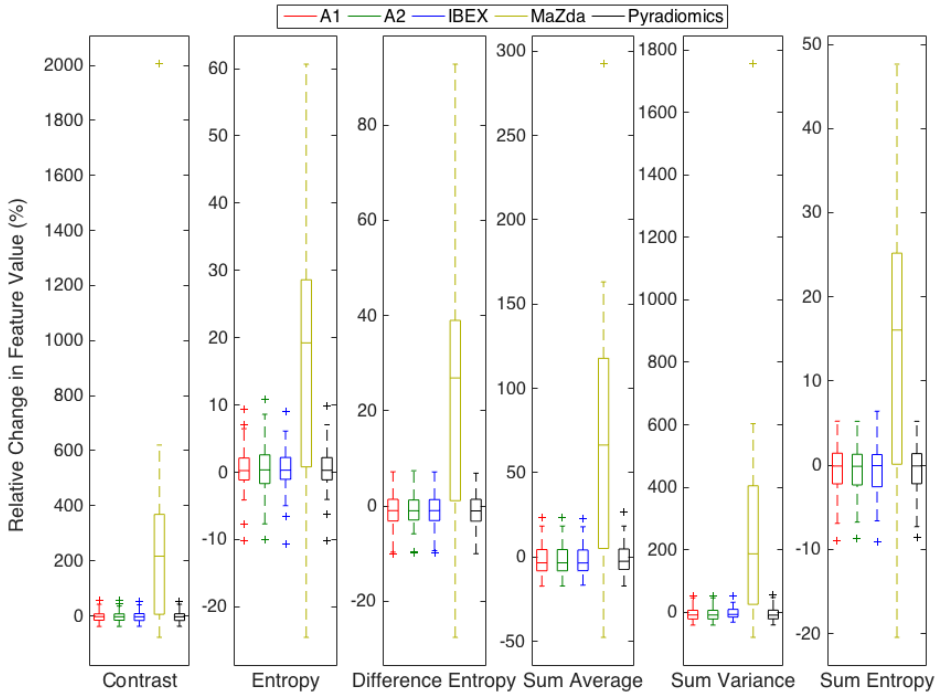


(a) Mammography

Figure 3.5: Distribution of GLCM features calculated on the mammography (a), HN CT (b), and breast MRI ROIs (c) when calculated using the consistent GLCM parameters. Boxes extend from the first to the third quartile with outliers indicated by +.



(b) Head and Neck CT



(c) Breast MRI

Figure 3.5 Continued: Distribution of GLCM features calculated on the mammography (a), HN CT (b), and breast MRI ROIs (c) when calculated using the consistent GLCM parameters. Boxes extend from the first to the third quartile with outliers indicated by +.

Table 3.8: The p-values resulting from the nonparametric Friedman tests comparing radiomic features across packages, and ICCs illustrating agreement in features among packages. Second-order features were calculated using consistent GLCM parameters. Significant differences and agreement based on ICC values were assessed with and without MaZda. Features reflecting significant differences are highlighted in red (p<0.004).

		Mammography		Head and Neck CT		Breast MRI	
	Feature	*p-Value*	*ICC*	*p-Value*	*ICC*	*p-Value*	*ICC*
With MaZda	GLCM Contrast	<0.004	0.421	<0.004	0.408	<0.004	0.036
	GLCM Entropy	<0.004	0.009	<0.004	0.032	<0.004	0.157
	GLCM Sum Entropy	<0.004	0.010	<0.004	0.025	<0.004	0.110
	GLCM Sum Average	<0.004	0.410	<0.004	0.283	<0.004	0.059
	GLCM Sum Variance	<0.004	0.018	<0.004	<0.001	<0.004	0.027
	GLCM Difference Entropy	<0.004	0.062	<0.004	0.061	<0.004	0.062
Without MaZda	GLCM Contrast	<0.004	0.999	<0.004	0.991	0.408	0.999
	GLCM Entropy	<0.004	0.083	<0.004	0.347	0.408	0.982
	GLCM Sum Entropy	<0.004	0.075	<0.004	0.254	0.005	0.997
	GLCM Sum Average	<0.004	0.998	<0.004	0.995	0.011	0.994
	GLCM Sum Variance	<0.004	0.025	<0.004	<0.001	0.756	0.840
	GLCM Difference Entropy	<0.004	0.408	<0.004	0.468	0.896	0.997

Because MaZda limited the GLCM parameters that could be customized, the analysis was repeated while excluding MaZda. When using the modified GLCM parameters, all second-order features still showed significant differences for both mammography and HN CT ROIs, while none of the features reflected significant differences for the breast MRI database as shown in Table 3.8. While ICCs excluding MaZda increased for every feature extracted from mammography and HN CT ROIs, only two features (GLCM Contrast and Sum Average) for both databases increased in value to exceed 0.9 indicating excellent agreement. When analyzing the breast MRI database, ICC values decreased for five of the six second-order features when consistent GLCM parameters were used. This is because the relative changes

in second-order features were all relatively close to zero, and the variability among ROIs was relatively small. Because the ICC is proportional to the between-patient variability, smaller differences in features among patients results in spuriously low ICC values. The consistency in feature values among patients result in ICC values being relatively noisy and not entirely reflective of the agreement in these relative differences. Although, when MaZda was excluded from the analysis, the ICC values for all second-order features reflected good or excellent agreement when consistent GLCM parameters were used.

3.4 Discussion

This study demonstrated and quantified differences in computed radiomic feature values among five radiomics software packages due to various sources of discrepancy. These sources of variation include differences in image importation and preprocessing, algorithm implementation, as well as GLCM and feature-specific parameters. While many first-order features showed relatively good agreement across packages, nearly all features significantly differed. All second-order features showed poor agreement and differed significantly when using package-specific default GLCM parameters. Therefore, when these radiomic features are used for predictive modeling, computer-aided diagnosis, or image segmentation, for example, the results could greatly differ depending on the software being used. Subsequently, the results from studies that use one particular package potentially may not correlate with studies that rely on a different package. Furthermore, if the same package is used, results may still not agree if feature parameters (e.g., GLCM parameters) are not consistent across these studies.

The effects of some of the differences in software packages may be mitigated by using a delta-radiomics architecture and calculating the differences in feature values between images. When analyzing the relative changes in feature values with the breast MRI database, none of the second-order features reflected significant differences when consistent GLCM parameters were used and when MaZda was not included in the analysis, while all second-order features

reflected significant differences for the mammography and HN CT databases under the same conditions (Table 3.8). When using delta radiomics to assess the relative change in feature values between time points, differences in software packages have the potential of being factored out. For example, the differences in log bases when calculating GLCM entropy between packages A2 and the remaining packages introduce a scaling effect on these features as shown in Figure 3.5.a and 3.5.b. When the relative changes in feature values were calculated, these scaling effects were removed (Figure 3.5.c). Although, calculating relative differences also has the potential to emphasize differences among packages. For example, one patient had a value of kurtosis very close to zero for the first image acquisition when calculated with MaZda resulting in the relative change in kurtosis between time points being greater than 4500% (outlier in Figure 3.2.c). This outlier did not exist for the remaining packages because of the differences in their respective definitions of kurtosis: MaZda calculated excess kurtosis while the remaining packages did not. This difference was quantified by an ICC of 0.006. Therefore, researchers should fully understand the definitions and equations for the features they include in their studies to avoid such deviations.

The exchange and comparison of radiomics software may allow for a standardization of these software packages resulting in more translatable radiomics-based research. For example, it was found by comparing the feature values across packages that package A1 had an error in the normalization of the GLCMs before feature calculation. Therefore, software errors may be revealed through a comparison of results and code from one institution to another, and it may be more likely to address previously unknown errors. Use of a standard set of "calibration" cases and reporting of the resulting feature values such as those provided by the IBSI could serve as a tool by which to validate and commission new radiomics software [43]. In addition, an editorial from Valliéres et al. [29] summarizes the elements of the radiomics workflow that may also result in variations and practices that could be used when incorporating radiomics into research. For example, the Radiomics Ontology offers a means of consistently reporting aspects of the radiomics workflow including radiomic

features, segmentation algorithms, and image filters. This editorial refers to the Responsible Research and Innovation website for guidelines regarding the effective reporting of research methods and results.

A number of studies have recognized and reported the need for standardizing the radiomics pipeline [38-41]. Consequently, the IBSI worked towards standardizing radiomics research by compiling an extensive manual of recommended feature definitions and image processing protocols. The collection of 19 institutions included in the IBSI used these recommendations to iteratively modify the feature extraction process when using a shared digital phantom and eventually a small set of CT scans from lung cancer patients. Features were considered standardized if 50% of the contributors produced the same feature value. Through this process, agreement was achieved for 99.4% and 96.4% of features extracted from the digital phantom and CT scans, respectively [44,79]. While the institutions included in the IBSI have increased the homogeneity of their radiomics workflow, the current study illustrates that the field would benefit from a broader standardization effort that captures institutions using both propriety in-house radiomics packages as well as freely-available open-source packages. The IBSI has established an important role in the standardization of feature definitions and radiomics algorithms. The goal of the present study was not to duplicate the IBSI effort or offer recommendations outside of those established by the IBSI but rather to quantify the differences in radiomic features computed from real-world clinical images by radiomics software packages that have been the basis for numerous publications. These findings provide additional support for the goals that the IBSI seeks to achieve and quantifies the sources of variation that are highlighted here and also by the IBSI.

This investigation included a few limitations that introduced a degree of uncertainty in the results while also indicating areas that may require attention while working towards standardizing radiomics research. The source code for MaZda was not available, making it difficult to investigate the underlying mechanics and isolating the components such as preprocessing and algorithm implementation. To facilitate reproducible research, freely-

available radiomics packages may want to increase the transparency of their methods by making the source code available to the public for comparison. Also, MaZda does not allow for automated ROI processing, so for robust prediction models that include hundreds or thousands of images, manual feature extraction could take several hours and introduce a large degree of human error. IBEX also did not inherently allow for automated feature extraction for multiple images while using altered GLCM parameters; however, the IBEX source code could be used to create an automated feature extraction function. Radiomics packages developed in the future should consider automating the feature extraction process while allowing the user to customize the feature calculation parameters such as those involved in constructing the GLCMs. Furthermore, when publishing results obtained from radiomics research, any relevant material required to reproduce the work, such as feature definitions or GLCM parameters, should be included in manuscript appendices or supplemental material for transparency.

Additional limitations of this study that hindered comparisons among packages included the inconsistency in computed radiomic features and the inability of some packages to calculate features in three dimensions. Of the hundreds of features that could be calculated among the various packages, the only feature classes all five software packages had in common were the first-order histogram features and GLCM features. Among these categories, only six features from each class were common among all five packages. This illustrates that features should be translatable across radiomics software packages by using the feature definitions supplied by the IBSI [43]. These definitions should also be used to allow for calculations in both two and three dimensions. For instance, because some packages included in this study could not compute three-dimensional features, comparison to IBSI digital phantom data was not possible [43].

Future work should incorporate additional radiomics packages to further test the variability of the resultant feature values. The first- and second-order features used in this study were chosen because they were the only twelve features that all five packages had in common;

however, using additional package combinations could allow for a larger number of studied features.

Additional studies should directly investigate the effects of analyzing images from additional imaging modalities such as positron emission tomography (PET). Radiomics packages may have been developed to process a particular type of image from a specific imaging modality. MaZda was originally developed to extract features from MRI scans with a particular range of pixel values, whereas A1 was originally developed to study lung CT scans in Hounsfield units. Therefore, the package-specific default GLCM parameters for A1 used a gray-level limit of -1500 to 1500, while the gray-level limit for MaZda was determined automatically based on the bit depth of the MRI ROIs. Investigating images from additional imaging modalities and additional tissue types could offer insight into how different packages behave under various circumstances.

3.5 Conclusion

An analysis of the variability in five radiomics software packages was performed to determine sources of discrepancies in computed radiomic features among packages. Inconsistencies in image importation, algorithm implementation, and GLCM parameters were investigated. The vast majority of features demonstrated significant differences in computed values across packages, however, most first-order features showed excellent agreement based on ICC. Second-order features had relatively poor agreement among packages when assessed by the ICC values; however, these differences could be reduced by comparing the relative change in feature values between time points. When GLCM parameters were modified to provide greater consistency across packages, ICC values increased but only two features reflected excellent agreement for the mammography and HN CT databases (GLCM contrast and GLCM sum average). Investigators should therefore use caution when adopting new radiomics packages and incorporating them into their research, ensuring the software used is appropriate for the images being studied. Investigators should also fully disclose the underly-

ing calculation parameters so that results from one radiomics-based study may be translated to other studies. Additional collaboration with groups such as the IBSI should be conducted to achieve greater harmonization of radiomics methods with direct clinical application across a greater number of institutions.

CHAPTER 4
EFFECTS OF VARIABILITY IN RADIOMICS SOFTWARE PACKAGES ON CLASSIFYING PATIENTS WITH RADIATION PNEUMONITIS

4.1 Introduction

The effects of the various components of the radiomics workflow on the final radiomic feature values has been well documented [2,30,36,37,41,56,74-77]; however, these studies have not determined whether these differences are clinically significant. Certain image acquisition and reconstruction parameters have were shown to introduce biases into the resultant feature values, but it is unclear whether these biases translate into differences in classification ability when applied to a particular clinical task. The previous chapter also reported that using different radiomics software can alter the resultant feature values by several orders of magnitude, but how these differences affect classification performance has not yet been quantified.

Many freely-available software packages have yet to become standardized to any particular reference despite the growing recognition to do so resulting in continued variability in radiomics research across institutions [37-44,55,80]. The previous chapter expanded on these studies and quantified the variability in radiomic feature values when different radiomics software packages were used to analyze medical images of different imaging modalities and tissue types. Quantifying the differences in radiomic feature values and understanding the sources of these deviations are important; however, the effects of this variability when applied to a clinical task are currently unknown. The purpose of the present study was to use three radiomics software packages to extract feature values from serial thoracic CT scans of patients receiving radiation therapy. A delta-radiomics architecture was implemented by using the changes in radiomic features over time to classify patients with and without ra-

diation pneumonitis (RP), and classification ability associated with each radiomics package was compared.

4.2 Methods and Materials

4.2.1 Imaging Data

A retrospective database of serial thoracic CT scans was acquired from 105 patients receiving radiation therapy (RT) for esophageal cancer. Each patient underwent two high-resolution diagnostic CT scans, with the first scan acquired prior to treatment and the second acquired no more than four months after treatment (Table 4.1) [53,96].

A treatment planning scan and the associated dose map obtained from treatment planning were also acquired for each patient. Dose maps were generated using heterogeneity corrections using a Pinnacle (Philips Medical Systems, Andover, MA) treatment planning system for photon therapy or Eclipse (Variation Medical Systems, Palo Alto, CA) treatment planning system for proton therapy. Patients were monitored for up to 6 months after treatment, and using all available documentation and imaging, RP status was determined through consensus of three clinicians using Common Toxicity Criteria for Adverse Effects, version 4 (CTCAE v4) [97]. Each patient was assigned a binary value reflecting RP status: 1 for patients with RP (grade$\geq$2) or 0 for patients without RP (grade$<$2) with examples shown in Figure 4.1.

Table 4.1: Patient, treatment, and image characteristics represented as the number of patients belonging to that category and the relative number of patients belonging to that category represented as a percentage in parentheses. (MRD: mean ROI dose; MLD: mean lung dose)

	Parameter Total	With RP	Without RP
No. of Patients	105 (100%)	20 (19%)	85 (81%)
Gender			
Male	88 (84%)	17 (85%)	71 (84%)
Female	17 (16%)	3 (15%)	14 (16%)
Median age	63 yrs	65 yrs	62 yrs
(range)	(27-81) yrs	(48-81) yrs	(27-79) yrs
Smoking History			
Current	15 (14%)	2 (10%)	13 (15%)
Former	68 (65%)	13 (65%)	55 (65%)
Never	22 (21%)	5 (25%)	17 (20%)
Treatment Modality			
IMRT	56 (53%)	9 (45%)	47 (55%)
3D-CRT	17 (16%)	4 (20%)	13 (15%)
Proton	32 (31%)	7 (35%)	25 (30%)
Treatment Dose Parameters			
Median Prescribed Dose	50.4 Gy	50.4 Gy	50.4 Gy
(range)	(45-66) Gy	(48.6-63) Gy	(36-66) Gy
Median No. of Fractions	28	28	28
(range)	(25-33)	(27-28)	(25-33)
Median MLD	9.6 Gy	10.5 Gy	9.4 Gy
(range)	(2.5-18.7) Gy	(2.9-15.2) Gy	(2.5-38.4) Gy
Median Lung V20	16.6%	15.5%	16.3%
(range)	(3.7-38.4)%	(4.8-31.3)%	(3.7-38.4)%
Median MRD	37.9 Gy	38.2 Gy	37.8 Gy
(range)	(31.2-44.7) Gy	(34.7-42.6) Gy	(31.2-44.7) Gy
Incidence of RP			
Grade 0	38 (36%)		
Grade 1	47 (45%)		
Grade 2	11 (10%)		
Grade 3	5 (5%)		
Grade 4	3 (3%)		
Grade 5	1 (1%)		

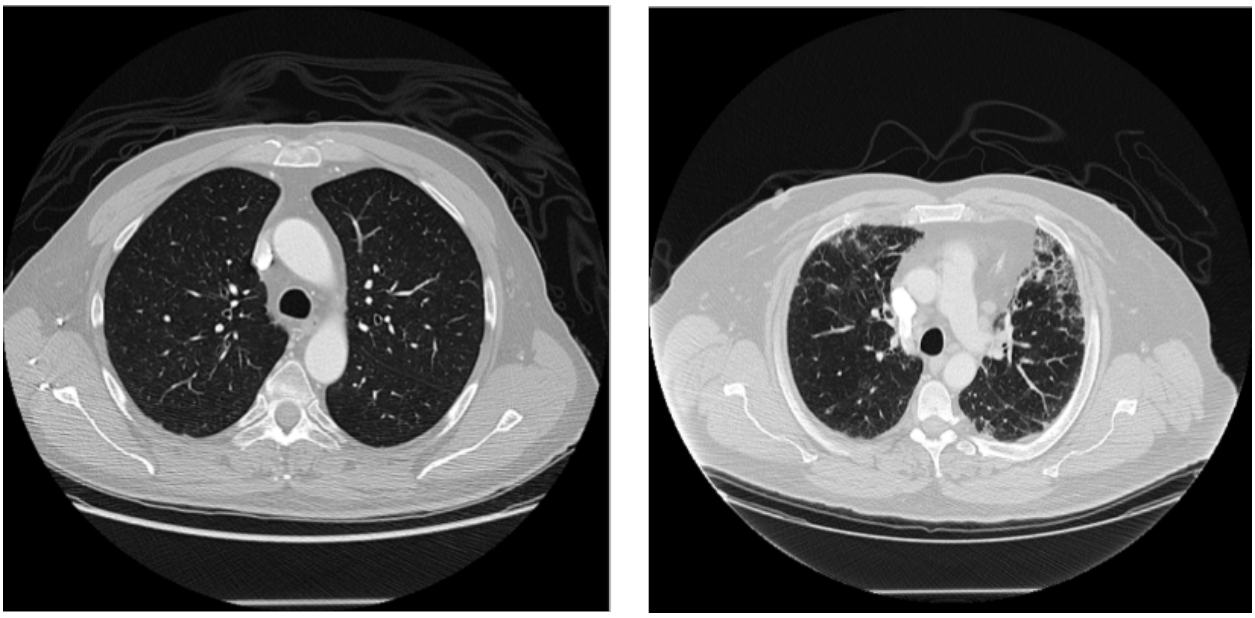


Figure 4.1: CT scans illustrating the differences in texture for patients without symptomatic radiation pneumonitis (RP grade: 0; left) and with symptomatic radiation pneumonitis (RP grade: 5; right), which appears as higher intensity pixels. (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

The patient- and treatment-specific variables shown in Table 4.1 and their association with RP status were evaluated using the Chi-squared test for nominal categorical variables and the Mann-Whitney U-test for continuous variables. Significance was assessed at the 0.05 level after correcting for multiple comparisons using Bonferroni ($p<0.006$).

The pre-RT, post-RT, and treatment planning CT scans were segmented using a semi-automated lung segmentation method and manually modified if necessary. Segmented post-RT scans were deformably registered to the pre-RT scans using the demons-based Plastimatch v1.5.12-beta software [98], resulting in a vector map that matched corresponding anatomy between image acquisitions. The post-RT CT scans were not deformed themselves to preserve the texture of the captured tissue structure.

4.2.2 Feature Calculation

Regions of interest (ROIs) 32×32 pixels in size (range in physical size: 20.0×20.0 mm^2 to 31.3×31.3 mm^2) were randomly placed in the lung volume of the pre-RT scan for each patient without overlap and with a maximum of 10 ROIs placed in each axial slice. Corresponding ROIs were anatomically matched in the post-RT scan using the vector map obtained from deformable registration as has been outlined in previous studies (Figure 4.2) [53,78,96]. To summarize these methods, the post-RT scan was not deformed thus preserving the texture of the image, but the vector map obtained from registration was used to anatomically match corresponding ROIs between time points.

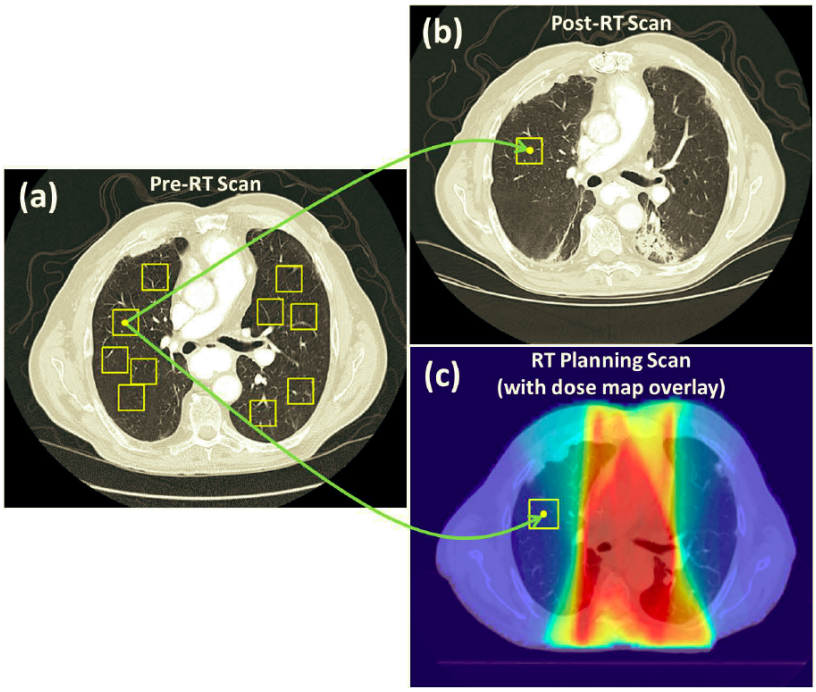


Figure 4.2: ROIs are randomly placed in the lung volume of the pre-RT scans (a), and the vector map obtained from deformable registration anatomically matches ROIs in the post-RT scan (b). The vector map obtained from deformably registering the treatment planning scan (c) is used to match ROIs in the pre-RT scan to the anatomical locations in the treatment planning dose map, assigning a dose distribution to each ROI. Only ROIs placed in high-dose regions (≥30 Gy) were used. (Reprinted with permission Cunliffe et al. Int. J. Radiation Oncol. Biol. Phys. 91(5) 2015).

Additionally, the treatment planning scan was deformably registered to match the corresponding dose map to the pre-RT scan and enable calculation of the average planned radiation doses within each pre-RT ROI. Only ROIs placed in high-dose regions ($\geq$30 Gy) were extracted given previous results showed that ROIs extracted from these regions were more predictive of RP development [53,96].

Three radiomics packages were used for analysis: one in-house package, A1, (developed at the University of Chicago) [53] and two open-source packages, IBEX v1.0 beta (The University of Texas MD Anderson Cancer center) [89] and PyRadiomics v.2.0.0 [90]. These packages were used because they were the only packages that were freely available with source code at the initiation of this study. Each package could also be automated to process multiple ROIs in a single setting, and these packages had been cited in the literature. Each radiomics package was used to calculate all 2-dimensional features common among the three packages that were also previously shown to be robust to deformable registration [31,78]. These features consisted of four first-order histogram features and four second-order gray-level co-occurrence matrix (GLCM) features (Table 4.2) with feature definitions outlined in the appendix. While first-order features quantify the various attributes of the gray-level histogram of the pixel values, GLCM features characterize the spatial distribution of pixels within an ROI. The construction of the GLCMs prior to feature calculation can vary with the parameters used to describe these matrices such as the gray-level limits, the number of gray levels, and the pixel binning [14,94]. Given that many radiomics studies do not report the GLCM parameters used, GLCM features were computed using the package-specific default parameters reported in the previous chapter (Table 3.4). For every feature computed with each radiomics package, a logistic regression model was constructed to classify patients with RP.

Table 4.2: First- and second-order radiomic features common among all three packages and also robust to deformable registration.

First-Order Histogram Features	**Second-Order GLCM Features**
Mean	Sum Average
Mimum	Sum Entropy
Median	Difference Entropy
Entropy	Entropy

4.2.3 Single-Feature Logistic Regression Modeling (M_{Avg})

Based on the results from the previous chapter, it was suspected that using changes in features between image acquisitions may mitigate some of the differences in radiomics software. For each feature, the differences in feature values between image acquisitions were calculated using methods similar to those used in previous studies [53,96]:

$$\overline{\Delta FV}_{F,S,p} = \frac{1}{N_p}\sum_{r=1}^{N_p}(FV_{F,S,p,r}^{post-RT} - FV_{F,S,p,r}^{pre-RT}) \tag{4.1}$$

where $\overline{\Delta FV}_{F,S,p}$ is the average change in feature F calculated using software package S over all ROIs in patient p. N_p is the total number of ROIs placed in high-dose regions for patient p, and $FV_{F,S,p,r}^{pre-RT}$ and $FV_{F,S,p,r}^{post-RT}$ are the feature values computed in ROI r in the pre- and post-RT scans of patient p, respectively.

Using the dose map and the corresponding vector maps obtained from deformable registration, the mean ROI dose (MRD) for each patient was calculated using:

$$MRD_p = \frac{1}{N_p}\sum_{r=1}^{N_p} D_{p,r} \tag{4.2}$$

where $D_{p,r}$ is the average planned radiation dose within the pre-RT ROI r of patient p, and N_p is the total number of ROIs placed in the high dose-region of patient p. This resulted in one value of MRD for each patient.

A logistic regression model was developed using the pROC package in R v3.3.3 classifying

patients with and without RP. Previous studies have reported that treatment-specific dose parameters are correlated with RP development; however, these results vary across institutions [99,100]. Therefore, logistic regression models were constructed using MRD alone, and individual features were added to this model to determine whether the addition of these features significantly improved classification ability using receiver operating characteristic (ROC) analysis, with the area under the ROC curve (AUC) as the performance assessment metric. Individual radiomic features calculated using each package were then added to the logistic regression model.

$$RP \sim MRD + \overline{\Delta FV}_{F,S} \tag{4.3}$$

Here, RP is a binary classifier indicating whether or not a patient develops RP (grade$\geq$2), MRD is the mean ROI dose for each patient, and $\overline{\Delta FV}_{F,S}$ is the mean change in each of the eight radiomic features (F) calculated using each of the three radiomics software packages (S). Eq. 4.3 resulted in one model for each feature-package combination (24 models in total). Models of this form using a single feature with changes in feature values averaged over all ROIs for each patient are referenced to as M_{Avg} for clarity. Analysis of variance (ANOVA) was used with Chi-squared tests to determine whether the addition of each feature to the logistic regression model significantly improved classification ability over using MRD by itself. The treatment-specific parameter MRD was also replaced in the regression model with the mean lung dose (MLD) and the relative volume of the lung that received at least 20 Gy (V20) to determine whether these parameters resulted in a different set of features that were significantly associated with RP. Significance was assessed at the $\alpha = 0.05$ level after correcting for multiple comparisons using Bonferroni ($p<0.002$).

During logistic regression, patient data were randomly sampled so that 50% of the patients were used for training the regression model, while the remaining 50% were reserved for testing. Sampling was performed to maintain the ratio of RP-positive to RP-negative patients in both the training and testing sets. Random sampling was performed 1000 times,

and an AUC value was calculated for each iteration, resulting in a mean AUC value across iterations along with the corresponding 95% confidence intervals.

4.2.4 Multi-Feature Logistic Regression Modeling

To determine whether combinations of features significantly improved classification ability and whether feature combinations differed among packages, an additional feature was added to the logistic regression model using Eq. 4.4.

$$RP \sim MRD + \overline{\Delta FV}_{F1,S} + \overline{\Delta FV}_{F2,S} \tag{4.4}$$

Here, the subscripts $F1$ and $F2$ refer to the first and second feature added to the model, respectively, and S refers to the software package used to calculate these features, similar to that shown in Eq. 4.3. Models were first created using each of the eight individual features, and the seven remaining features were added to each of these models, resulting in a total of 56 feature combinations. Significance was assessed at the 0.05 level after correcting for the 56 comparisons per package ($p<0.0009$).

Because of the limited number of RP-positive patients in this dataset, the potential for overfitting was a concern. Therefore, the Akaike information criterion (AIC) was used to assess the relative quality of the models when the first and second features were included in the model compared with when only the first feature was included. The AIC quantifies model quality by balancing the potential for improved goodness of fit due to additional feature inclusion with the deficit introduced by the potential for overfitting [101]. The AIC was calculated using the following equation:

$$AIC_{F1,F2,S} = -2(log - likelihood) + 2(n_{par}) \tag{4.5}$$

Here, $AIC_{F1,F2,S}$ is the AIC value when feature $F2$ was added to a model that included the first feature, $F1$, and MRD when both features were calculated using software package

S. The log-likelihood reflects how well the model fits the data, while n_{par} is the number of parameters. The absolute value of the AIC is arbitrary, but feature combinations that result in smaller AIC values compared to models including one feature correspond to models of greater relative quality. In other words, the potential for overfitting due to the increased number of parameters in the regression is outweighed by the improved model fit reflected by the log-likelihood.

4.2.5 Individual ROI Pair Logistic Regression Modeling (M_{Ind})

Averaging the differences in feature values over all ROIs for each patient could potentially dampen the impact of ROI pairs that demonstrate larger texture differences and therefore could be more indicative of RP development. Therefore, an additional set of models was constructed using each ROI pair as a distinct analyzable unit when training the logistic regression model (Figure 4.3):

$$\Delta FV_{F,S,p} = FV_{F,S,p,r}^{post-RT} - FV_{F,S,p,r}^{pre-RT} \tag{4.6}$$

Similar to what is shown in Eq. 4.1, $FV_{F,S,p,r}^{post-RT}$ and $FV_{F,S,p,r}^{post-RT}$ are the values for feature F computed using software S in ROI r in the pre- and post-RT scans of patient p, respectively. For these models, $\Delta FV_{F,S,p,r}$ was included in the logistic regression for each feature along with the mean dose to the corresponding ROIs ($D_{p,r}$ in Eq. 4.2), resulting in a total of 4474 analyzable units instead of the 105 units used previously:

$$RP \sim D + \Delta FV_{F,S} \tag{4.7}$$

Models using a single feature (F) and individual ROI pairs are referred to as M_{Ind} models. ANOVA was used to determine whether the addition of each feature to logistic regression significantly improved model fit over using just the mean dose to the ROIs. During the testing process, both M_{Ind} and M_{Avg} models were assessed using the averages in feature value

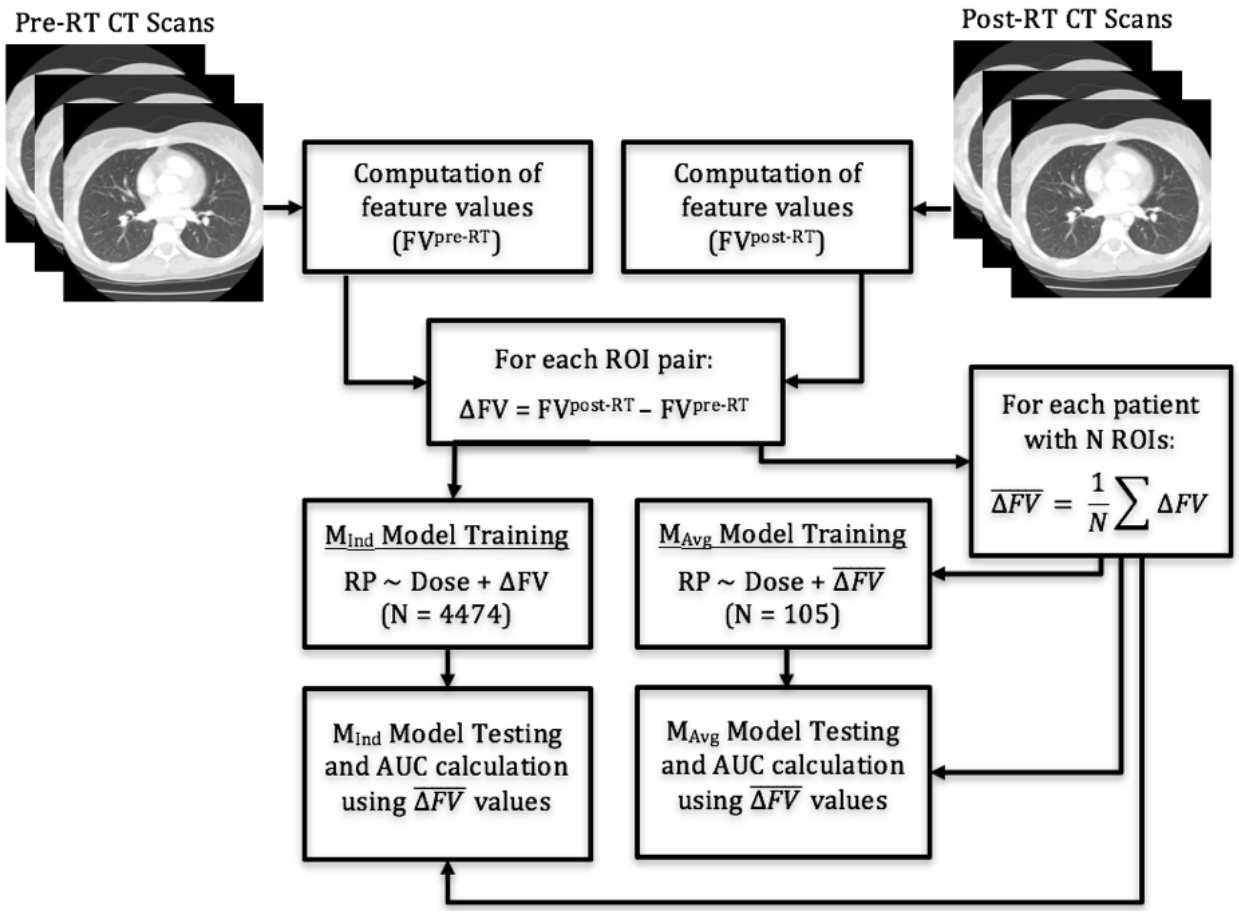


Figure 4.3: Flowchart depicting regression models trained using individual ROI pairs (M_{Ind} models) as well as averages over ROI pairs for each patient (M_{Avg} models). Both models are tested using $\overline{\Delta FV}_{F,S,p}$ values but were trained differently. (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

change over all ROIs for each patient ($\overline{\Delta FV}_{F,S,p}$ from Eq. 4.1), so that both models were assessed in the same way while also testing the M_{Ind} models using uncorrelated $\overline{\Delta FV}_{F,S,p}$ values. In other words, ROIs from the same patient cannot be used to test the classification models given that ROIs extracted from the same patient scans are likely to have feature values that are correlated with one another. Also, if each patient contributes a different number of ROIs, this would skew the classification ability of the model. It should be noted that for both models, patients were split into training and testing sets such that ROIs from the same patient were not used in both training and testing sets during the same sampling iteration. Half of the patients were sampled for training and all individual ROIs correspond-

ing to those patients were used for training while the remaining ROIs were averaged for each patient and used for testing, resulting in one value of $\overline{\Delta FV}_{F,S,p}$ for each feature for a given patient. Sampling, training, and testing were performed in the same way as described in the 'Single-Feature Logistic Regression Modeling' section. M_{Ind} models and M_{Avg} models were compared using Vuong's closeness test of non-nested model comparison [102].

4.3 Resuts

The patient- and treatment-specific variables shown in Table 4.1 were not significantly correlated with RP status.

4.3.1 Single-Feature Logistic Regression Modeling (M_{Avg})

Four first-order gray-level features and four second-order GLCM features were calculated using three different radiomics packages and using CT ROIs as input. For each feature, $\overline{\Delta FV}_{F,S,p}$ values were added to a logistic regression model including only MRD (M_{Avg} models) to determine whether the addition of the feature significantly improved the classification ability of that model (Table 4.3). Packages typically agreed on the features that were significantly associated with RP other than GLCM difference entropy and GLCM entropy, which only showed significant association for features extracted using Pyradiomics. Dosimetric parameters alone (MRD, V20, and MLD) were not found to be significantly associated with RP (p-values equal to 0.19, 0.24, and 0.10, respectively). Inclusion of any of the three dosimetric parameters into a regression model did not change which features were significantly associated with RP as shown in Table 4.3.

The mean AUC values along with the 95% confidence intervals for each of the eight features are shown in Figure 4.4.

Table 4.3: The p-values indicating whether the addition of a particular feature significantly improved classification ability over using the MRD alone. Each regression model was trained using averages in changes in feature values over all ROIs for each patient (M_{Avg}). (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

Features	A1	IBEX	Pyradiomics
Mean	<0.002	<0.002	<0.002
Minimum	0.593	0.133	0.593
Median	<0.002	<0.002	<0.002
Entropy	<0.002	<0.002	<0.002
GLCM Sum Average	<0.002	<0.002	<0.002
GLCM Sum Entropy	<0.002	<0.002	<0.002
GLCM Difference Entropy	0.008	0.010	<0.002
GLCM Entropy	0.222	0.947	<0.002

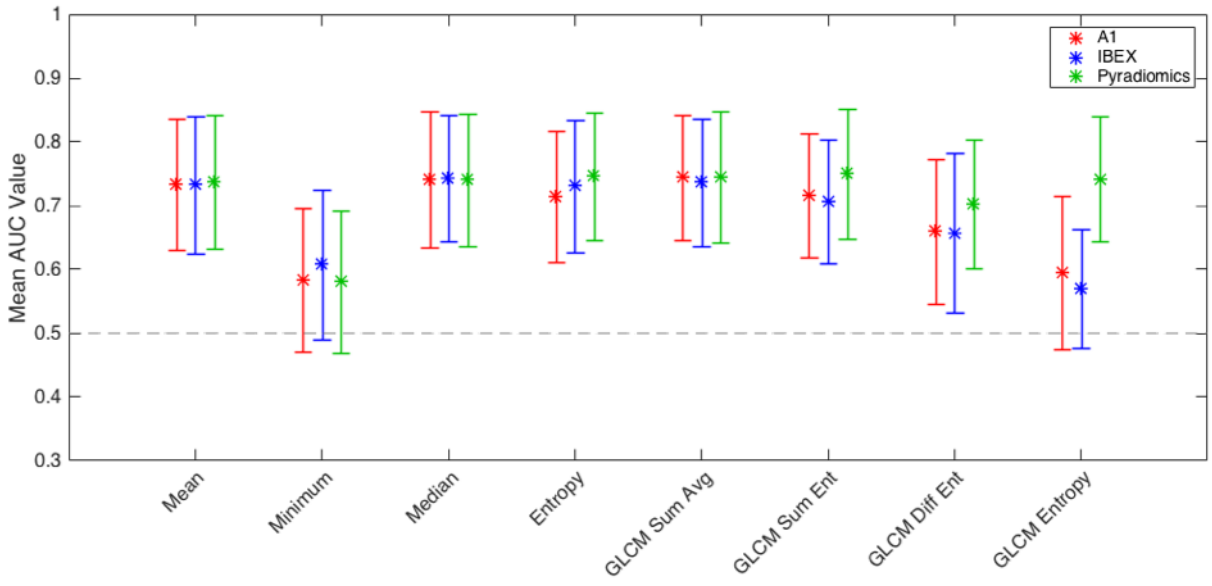


Figure 4.4: Mean AUC values along with the corresponding 95% confidence intervals for eight features used to train M_{Avg} models. (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

4.3.2 Multi-Feature Logistic Regression Modeling

Figure 4.5 illustrates feature combinations in green that significantly improved model fit over using the first feature and the MRD alone ($p<0.0009$). Cells in red illustrate features that did not improve model fit when added to the first feature in the model.

The addition of a second feature significantly improved model fit for most features for all three packages. Feature combinations that significantly improved model fit tended to agree among packages: of the 56 feature combinations for each package, 40 (71%) combinations resulted in significant improvement (green) in model fit for all three packages. Nine (16%) feature combinations did not result in significant improvement (red) in all three packages, and the remaining 7 (13%) combinations differed among packages. As an example, when GLCM sum entropy or GLCM difference entropy was added to a model already containing first-order entropy, the effect on the model fit differed among the three packages: there were multiple feature combinations that disagreed among packages when first-order entropy was the first feature included in the model. When adding GLCM difference entropy to a model using first-order entropy, the additional feature significantly improved model fit when features were calculated using Pyradiomics, but not for packages A1 or IBEX. When assessing model quality based on AIC, all feature combinations shown in green had AIC values that were lower (corresponding to higher model quality) than when just the first feature and MRD were used in the model.

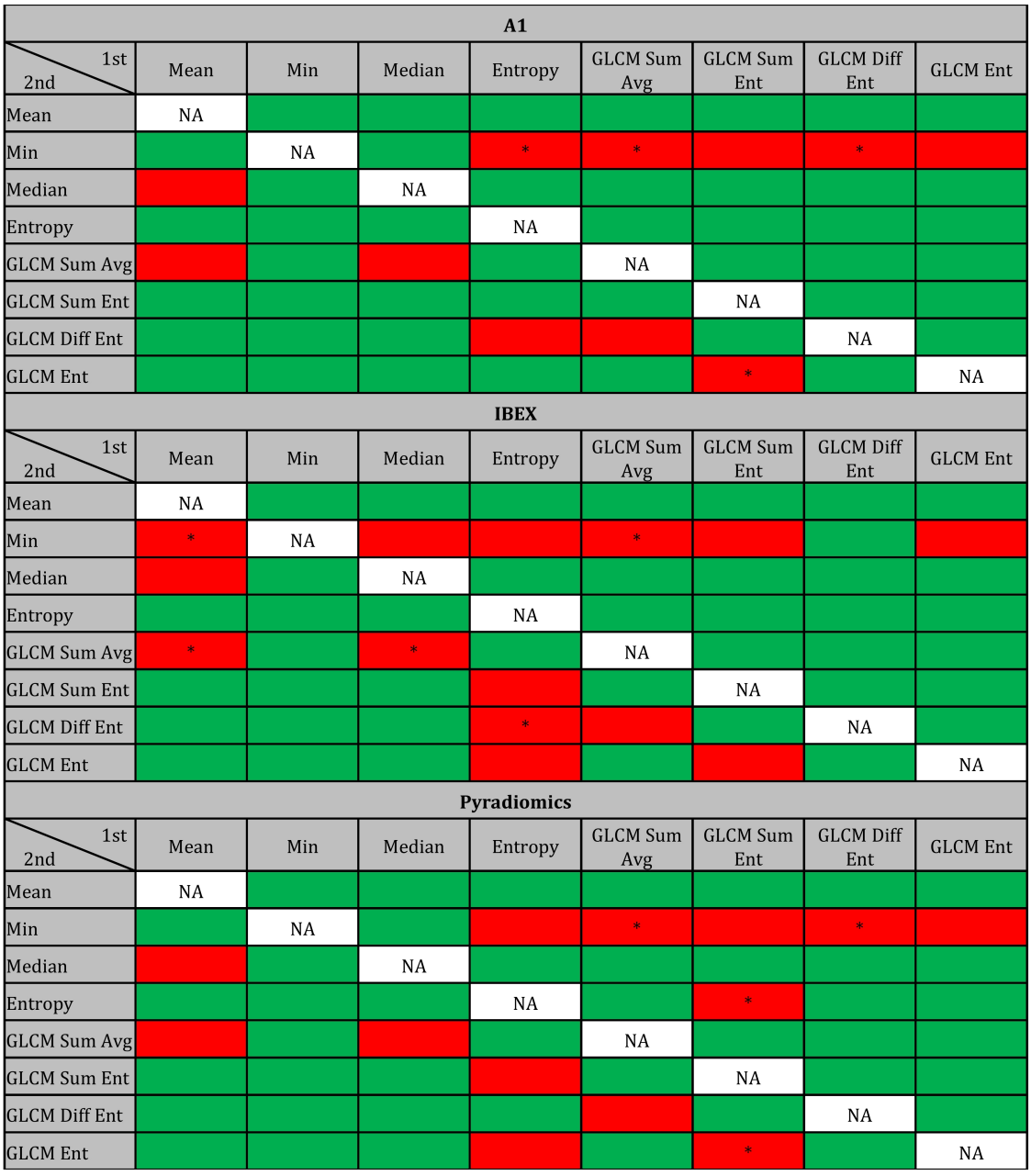


Figure 4.5: Green cells indicate the addition of a second feature in logistic regression that significantly improved model fit over using the first feature and MRD alone when features were calculated using package A1, IBEX, or Pyradiomics. Columns correspond to the first feature included in the model, and rows correspond to the second feature added to the model. Significance was assessed at the 0.05 level after correcting for the 56 different comparisons per package ($p<0.0009$). Cells labeled with an asterisk reflect feature combinations resulting in greater AIC values (lower model quality) than when only the first feature was included in the model. Each regression model was trained using averages of changes in feature values over all ROIs for each patient (M_{Avg}). (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

4.3.3 Individual ROI Pair Logistic Regression Modeling (M_{Ind})

When individual ROI pairs ($\Delta FV_{F,S,p,r}$) were used to train the logistic regression models instead of the average differences in feature values across ROIs for each patient ($\overline{\Delta FV}_{F,S,p}$), a greater number of features were significantly correlated with RP status as shown in Table 4.4. The bolded p-values are features that were not considered correlated with RP status for the M_{Avg} models.

Table 4.4: The p-values indicating whether the addition of a particular feature significantly improved classification ability over using the MRD alone when individual ROI pairs were used in the training of each model (M_{Ind}). During testing, average changes in feature values ($\overline{\Delta FV}_{F,S,p}$) values were used. Bolded p-values indicate features that were considered significantly correlated with RP for M_{Ind} models but not for M_{Avg} models. (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

Features	A1	IBEX	Pyradiomics
Mean	<0.002	<0.002	<0.002
Median	0.024	**<0.002**	0.024
Minimum	<0.002	<0.002	<0.002
Entropy	<0.002	<0.002	<0.002
GLCM Sum Average	<0.002	<0.002	<0.002
GLCM Sum Entropy	<0.002	<0.002	<0.002
GLCM Difference Entropy	**<0.002**	**<0.002**	<0.002
GLCM Entropy	**<0.002**	**<0.002**	<0.002

The mean AUC values and the corresponding 95% confidence intervals are shown in Figure 4.6. The same trends in mean AUC values were observed for M_{Ind} and M_{Avg} models, but the mean AUC values were slightly greater when trained using M_{Ind} models rather than M_{Avg} models for all packages and for all features other than first-order and GLCM entropy from A1. The models constructed with the two methods (i.e., M_{Avg} vs M_{Ind}), however, were not significantly distinguishable based on Vuong's closeness test.

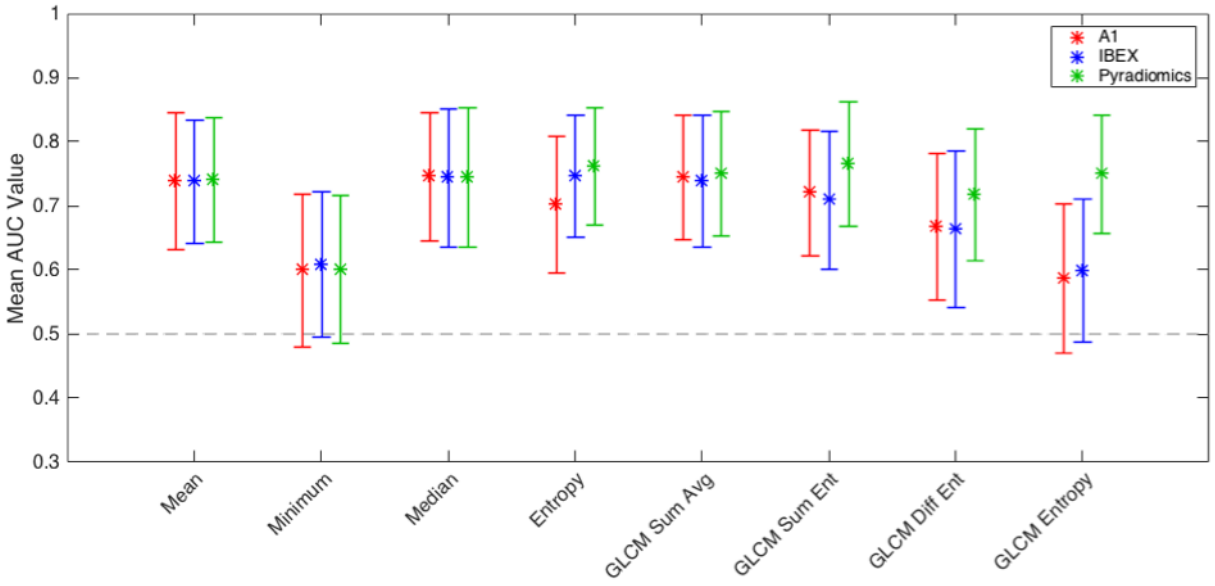


Figure 4.6: Mean AUC values along with the corresponding 95% confidence intervals for eight features when individual ROI pairs were used in model training. (Reprinted with permission Foy et al. J. Med. Imag. 7(1) 2020).

4.4 Discussion

This study demonstrated the variability in classification ability when radiomic features were computed with three radiomics software packages and applied to a clinically relevant classification task. The previous chapter localized this variability in software to discrepancies in a number of aspects in the feature calculation pipeline including differences in image importation and preprocessing, algorithm implementation, and feature-specific calculation parameters [29,37]. While many first-order features agreed among packages on whether the feature was correlated with RP status, two of the four GLCM features disagreed among packages (Table 4.3). A similar trend was shown in the mean AUC values plotted in Figure 4.4: for packages A1 and IBEX, GLCM entropy was not significantly greater than 0.5, but the mean AUC was significantly greater than 0.5 for Pyradiomics. The previous chapter illustrated how GLCM feature values are heavily dependent on the parameters used to construct the matrices prior to feature calculation, and the default GLCM parameters unique to

each package can alter the value of each GLCM feature [37]. When the GLCM parameters were made to be consistent across the three packages using the parameters shown in Table 3.4 (Gray Level Limits: [Min, Max]; Number of Gray levels: 64; Number of Directions: 8), all four GLCM features were significantly correlated with RP status for all three packages. Additionally, all four GLCM features had AUC values that were significantly greater than 0.5 for all three packages, indicating that the parameters used to calculate various radiomics features can greatly affect the agreement in classification ability among software packages. While modifying these parameters may result in greater agreement among radiomics packages, many radiomics-based studies do not report complete definitions of each feature or parameters used for their calculation. Consequently, package-specific default GLCM parameters are often used thus preventing independent investigators from accurately recalculating these features. When applied to the clinical task of classifying patients with RP, the effects of these discrepancies are still noticeable: mean AUC values agreed much more among packages for first-order features compared to GLCM features. Consequently, when investigators attempt to reproduce, validate, or advance radiomics studies reported in the literature using different radiomics software packages, they may identify a different set of correlated features or achieve different levels of classification ability. Therefore, to maximize the reproducibility and transparency of radiomics research, investigators should use open-source and freely available radiomics software to allow for this work to be validated.

A common method of improving classification performance for many radiomics-based classification schemes is to combine features during model construction if the dataset is large enough. In the current study, pairwise combinations of features were used in logistic regression, and the addition of a second feature significantly improved the model fit over using the first feature alone for the majority of combinations. While features that improved model fit over dose alone typically agreed among packages, feature combinations varied from package to package as shown in Figure 4.5. Additionally, feature combinations that significantly improved model fit differed more for GLCM features than for first-order features

because of the increased complexity of GLCM feature calculation and the potential for larger discrepancies in GLCM feature values computed with different software. Moreover, feature combinations that decreased the model quality based on AIC (Figure 4.5) showed very little agreement among packages, further illustrating that radiomics studies reporting promising features or combinations of features may not translate to other institutions using different radiomics packages.

When individual ROI pairs were used in model training (M_{Ind} models) instead of averages over ROI pairs (M_{Avg} models), more features were significantly correlated with RP status, and seven out of eight features agreed among packages. The discrepancies in correlated features between M_{Ind} and M_{Avg} models illustrate that differences in model implementation can affect the library of features capable of accurately classifying patients into different disease states. Despite the set of correlated features differing between the M_{Ind} and M_{Avg} models, the predictive ability of the two models was not significantly distinguishable based on Vuong's closeness test for any feature computed using any package; however, when classifying other diseases or using larger patient databases, M_{Ind} models may be more sensitive to changes in tissue structure and result in differences in classification ability.

This study also illustrates how variations in feature values extracted from the same image set does not necessarily imply variations in classification ability. The previous chapter found significant differences in feature values among radiomics packages for 11 out of 12 features extracted from head and neck CT scans and mammograms. In the present study, all features also reflected significant differences across packages, but the variability in feature values was much greater than the variability in AUC values when assessed using the coefficient of variation (COV). For example, when the mean of GLCM sum average is calculated across all ROIs for each package, the COV values among the three packages for the pre-and post-RT ROIs were 1.456 and 1.403, respectively. In comparison, the COV for the AUC values among the three packages was 0.006, indicating much less variability in AUC values than feature values across packages.

Despite the relatively large differences in the feature values themselves, when assessing which features were significantly correlated with RP (Table 4.3), six of the eight features agreed among packages. This is in part due to the inherent nature of delta-radiomics schemes: variations in radiomic feature values introduced by each software package due to differences in image preprocessing or algorithm implementation may be mitigated when assessing the changes in radiomic features over time. Therefore, only quantifying the differences in radiomic features may not be sufficient. Whether these differences translate into clinical practice when applied to a particular task is imperative to understanding the reproducibility or lack thereof of radiomics research.

This study contained a number of limitations in its methodology and areas that could be improved upon in future studies. First, only three radiomics packages were used in this investigation. Future studies could include additional radiomics packages or combinations of packages to allow for the evaluation of a greater number of features. This study was limited to only the eight first-order and GLCM features that were common among the three packages; however, additional feature categories, such as fractal, Fourier, or gray-level run-length matrix features may display variations in classification ability when computed with different software packages. Investigating these additional feature categories may illustrate additional areas of discrepancy among radiomics software, which will subsequently aid in the overarching aim of standardizing the radiomics workflow to make future research studies more reproducible and more translatable [43].

Additional studies could also investigate the variability in classification ability when applied to different diseases and imaging modalities. Previous studies have shown that the degree of variability in raw feature values among packages can differ depending on the tissues being analyzed and the modalities used to image these tissues [37]. Radiomics packages are often developed and designed to analyze a particular range of pixel values or particular disease or tissue type. When these packages are used to analyze images beyond those for which they were designed, the resulting feature values may be meaningless and not reflective

of the true texture. Therefore, it may be beneficial for future studies to analyze images outside of lung CT to determine whether classification ability differs when applied to a different clinical task.

4.5 Conclusion

This study investigated the variability in classification ability among three radiomics packages when distinguishing patients with and without radiation pneumonitis. When assessing which features were significantly associated with RP, first-order features reflected greater agreement among packages, whereas GLCM features reflected greater variation. When additional features were added to the logistic regression models, feature combinations that improved classification ability over using the first feature alone also differed among packages. The current study also illustrated that changes to the statistical modeling process may change the subset of features that reflect a significant correlation with a particular disease status. Initiatives have worked towards standardizing the radiomics workflow across institutions; however, the present findings indicate that additional effort must be put towards harmonizing radiomics research to achieve greater clinical implementation of their results.

CHAPTER 5
HARMONIZATION OF RADIOMICS SOFTWARE ON CLASSIFYING PATIENTS WITH RADIATION PNEUMONITIS

5.1 Introduction

Chapter 3 quantified the variability in radiomic features when different radiomic software packages were used to analyze the same medical images [37]. Chapter 4 illustrated that this variability may translate to differences in classification ability when applied to a clinical task using a delta-radiomics modeling architecture; however, if this variability can be mitigated using the harmonization methods investigated in Chapter 2, these methods may allow for greater translation of radiomics research across institutions [103]. Facilitating multi-center radiomics studies may allow for classification models to be more easily reproduced and validated and ultimately implemented into clinical practice.

The image processing methods used in Chapter 2 (i.e., histogram normalization, pixel size resampling, and Butterworth filtering) would not be useful when attempting to harmonize features calculated using different radiomics software packages given the same images are used as input for each of these packages. Any modification to the images themselves would propagate through the software packages, and the variability in the feature calculation process due to these packages would remain in the resultant feature values. Therefore, only harmonization methods applied to the feature values themselves would be viable (i.e., ComBat). This study focuses on ComBat harmonization and its potential to limit the variability in the feature calculation process while preserving the indicators of RP development.

Previous studies have shown that ComBat has potential to mitigate the batch effects that result from differences in repeated measurements in genomics and, more recently, radiomics studies. This method was initially designed to reduce the variability introduced in genomics

studies due to sampling measurements on different days, with different equipment and, often performed by different technicians. Johnson et al. [63] first used ComBat to align the stabilization of mRNA from human lung fibroblast samples after nitric oxide exposure. Samples were acquired at three different times resulting in slight biases in measurements because of slight differences in sampling conditions. When using a hierarchical clustering algorithm, measurements were clustered primarily by the time at which they were sampled prior to the use of ComBat. After ComBat harmonization was applied, samples were correctly clustered based on the stabilization of the mRNA with the variability due to the sampling day removed.

Additional studies have attempted to apply ComBat harmonization to radiomics research. Orlhac et al. [47] extracted various texture features and standardized uptake values (SUVs) from the pre-treatment positron emission tomography (PET) scans of patients with non-metastatic breast cancer acquired at different institutions. Before ComBat, several of the features reflected significant differences between institutions, while none of the features reflected significant differences after ComBat was applied. When assessing the influence of CT unit manufacturer on radiomic features extracted from scans of a phantom composed of ten different materials, ComBat harmonization reduced the number of features reflecting significant differences to zero for all phantom materials [64]. Robinson et al. [104] extracted radiomic features from full-field digital mammogram screenings when mammograms were acquired with different unit manufacturers, and they found that ComBat was capable of removing the variability introduced by differences in manufacturers while preserving the indicators of breast cancer risk. Finally, Whitney et al. [105] used ComBat to align radiomic features extracted from dynamic contrast-enhanced (DCE) MRI scans between US and Chinese databases while preserving the indicators of breast cancer malignancy. These studies have illustrated ComBat's ability to harmonize radiomics research; however, most of these studies did not apply the harmonized features to a clinical task, and no study as attempted to align classification ability between batches. Previous studies have attempted to use Com-

Bat to align feature values among batches to allow for the combination of databases acquired under different conditions, however, the dependence of the subsequent classification performance on ComBat harmonization has not be fully investigated. Therefore, the purpose of the study reported in this chapter was to assess ComBat's potential to remove the variability introduced by differences in radiomic software packages while preserving imaging biomarkers to harmonize the classification of radiation pneumonitis development.

5.2 Methods and Materials

5.2.1 Imaging Data

The same radiomic software packages and ROI pairs from the 105 esophageal cancer patients examined in the previous chapter (Chapter 4.2.1) were used for this study.

5.2.2 Feature Calculation

In the current study, ROI placement and mapping between pre- and post-RT ROI pairs were performed in the same manner as in Chapter 4.2.2. Packages A1, IBEX, and Pyradiomics were used to calculate the features robust to deformable registration and common among all three packages outlined in Table 4.2 using package-specific default GLCM parameters (Table 3.4).

5.2.3 Statistical Analysis and Modeling

Chapter 4 illustrated that the association of the changes in some radiomic features with RP development was dependent on the software package used to calculate these features [103]. To assess methods of mitigating this dependency (i.e., batch effect), the delta-radiomics models outlined in Chapter 4.2.3 (M_{Avg} models), were compared to three additional ComBat-based models. For each model, the changes in each feature between pre- and post-RT ROI pairs placed within the high-dose regions within the lung volume were calculated using the three

software packages (Equation 4.1). The changes in feature values between time points were averaged over all ROIs for each patient and combined with the mean dose across ROIs (MRD) in logistic regression to classify patients with and without RP. Patients were randomly sampled so half of the patients were used to train the model, while the remaining half of the patients were reserved for testing while maintaining the ratio of RP-positive to RP-negative patients for both training and testing sets. Receiver operating characteristic (ROC) curve analysis was used to quantify the classification ability of each feature using the AUC value. This sampling, training, and testing approach was repeated 1000 times, and a mean AUC value was calculated along with the 95% confidence intervals. Analysis of variance (ANOVA) was used with Chi-squared tests to determine whether the addition of a particular feature in logistic regression significantly improved the model fit over using the MRD alone. Significance was assessed at the $\alpha = 0.05$ level after correcting for multiple comparisons using Bonferroni ($p<0.002$). The agreement in feature values among packages was quantified using the intraclass correlation coefficient (ICC) initially described in Chapter 3.2.5. The agreement in mean AUC values among packages and across features was also quantified with a single ICC value for a particular modeling architecture. For modeling architectures where the variability in mean AUC values among packages was much less than the variability in mean AUC values across features, ICC values will be close to one indicating good agreement. Conversely, if the variability in mean AUC values among packages is relatively large, the ICC value will be closer to zero indicating poorer agreement [92,93].

5.2.4 ComBat Harmonization

ComBat harmonization operates by estimating the effect of using different radiomics software packages by assuming that the value of each feature, f, measured in each ROI, r, with software package, s, can be expressed as

$$f_{r,s} = \alpha + \gamma_s + \delta_s \varepsilon_{r,s} \tag{5.1}$$

where α is the average value for feature $f_{r,s}$, γ_s is the additive software effect, and δ_s is the multiplicative software effect. The $\varepsilon_{r,s}$ term is an error term that follows a normal distribution with mean zero and a variance equal to the variance across features for each ROI. ComBat estimates the model parameters α, γ_s, and δ_s using a maximum likelihood approach and calculates the adjusted feature values using the estimates of these parameters, $\hat{\alpha}$, $\hat{\gamma_s}$, and $\hat{\delta_s}$:

$$f_{r,s}^{ComBat} = \frac{f_{r,s} - \hat{\alpha} - \hat{\gamma_s}}{\hat{\delta_s}} + \hat{\alpha} \tag{5.2}$$

ComBat harmonization uses Eq. 5.2 to remove the scaling and translational effects of the various batches (i.e., software packages), while attempting to preserve the quantitative descriptors of RP development.

In this study, three models were developed with ComBat implemented at different components of the feature calculation and modeling process (Figure 5.1). Each model labeled the package used to calculate a particular feature as the "batch effect" to be removed, and the RP status associated with that patient or ROI was labeled as the covariate to be preserved. The first model, refered to as the $M_{ComBat1}$ model, applies ComBat to the pre- and post-RT feature values separately and for each ROI before the differences in feature values were calculated (n = 4474 ROI pairs). For the pre-RT ROIs, all ROIs were assigned an RP status of zero corresponding to no RP development given none of the patients had pneumonitis prior to RT. For the post-RT features, ROIs were assigned the proper binary RP status associated with that patient after treatment.

In the original genomics studies, it was assumed that the variability introduced into one measurement was related to the variability introduced into additional measurements, so ComBat uses the variance across measurements or feature values for each patient or ROI to adjust each feature. Therefore, ComBat was applied separately to first-order features and for GLCM features for all models using ComBat.

For the second model ($M_{ComBat2}$ model), ComBat was applied to the differences in

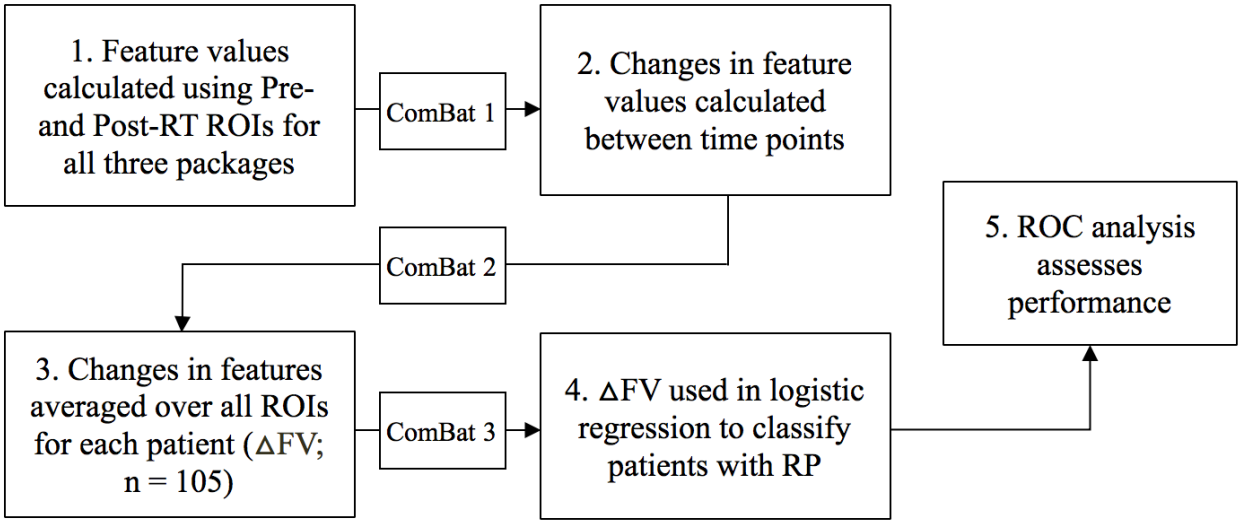


Figure 5.1: Flowchart illustrating the calculation of the feature values to be used to classify patients with RP. ComBat was applied to the feature values at one of various stages of this workflow to determine when it was most effective: For $M_{ComBat1}$ models, ComBat was applied to the pre- and post-RT features separately. For $M_{ComBat2}$ models, the differences in feature values were harmonized prior to averaging features among ROIs. Finally, for $M_{ComBat3}$ models, the differences in feature values between time points averaged over all ROIs for each patient were harmonized. M_{Avg} models were achieved when ComBat was not applied.

feature values between time points but before differences were averaged over all ROIs for each patient. These models applied the RP status from each patient to the corresponding ROI pairs. After ComBat, the modified changes in feature values were averaged over all ROIs for each patient and used in logistic regression.

For the third and final model ($M_{ComBat3}$ model), ComBat was applied to the features after the differences in these features were averaged over all ROIs for each patient (n = 105 patients), and the RP status for each patient was used as the covariate to be preserved.

For each model, repeated measures ANOVA was used before and after ComBat to assess significant differences in feature value distributions among the three packages after assessing normality using the Shapiro-Wilk test. ICC values were calculated to reflect the relative agreement in feature value distributions among packages. The Kolmogorov-Smirnov (K-S) test was used to quantify the separation of features corresponding to RP-positive and

RP-negative patients before and after ComBat. The K-S test statistic is calculated using the maximum differences between the cumulative distribution functions associated with the feature value distributions of patients with and without RP using the following equation:

$$D^* = max_x(|\hat{F}_1(x) - \hat{F}_2(x)|) \tag{5.3}$$

Here, $\hat{F}_1(x)$ is the proportion of $x1$ values less than or equal to x and $\hat{F}_2(x)$ is the proportion of $x2$ values less than or equal to x. Values of D^* closer to 1 correspond to greater discrepancy between RP-positive and RP-negative feature values, while values of D^* closer to zero corresponding to greater overlap in feature value distributions [106].

To determine the effect of the differences on the scale of the feature value distributions, models were constructed after the distributions were z-normalized to have an average value of zero and a standard deviation of one:

$$FV_{Z-Norm,f} = \frac{FV_f - \mu_f}{\sigma_f} \tag{5.4}$$

Here, f_i is the feature value distribution for feature f, and μ_f and σ_f are the average value and standard deviation of feature f. To assess the influence of the covariates to be preserved during ComBat (i.e., RP status), ComBat harmonization was also implemented without including RP status as the covariate to be preserved.

5.3 Results

5.3.1 Effect of Harmonization on Radiomic Features

The change in each of eight radiomic features between pre- and post-RT CT scans was calculated using three software packages. The change in each feature was used to construct four logistic regression models classifying patients with RP: one model used either the differences in feature values between time points (M_{Avg}), while the remaining three used ComBat

harmonization to align the changes in feature values at different portions of the modeling workflow (Figure 5.1). The p-values resulting from the repeated measures ANOVA used to assess significant differences in feature distributions among the three software packages are shown in Table 5.1 for each of the four modeling architectures. Table 5.1 also shows the ICC values reflecting the agreement in feature values across packages for all patients for each of the four modeling architectures.

Table 5.1: Results of the repeated measures ANOVA assessing significant differences in feature values among packages for each of the four modeling architectures. The ICC values are also shown assessing the relative agreement in feature values among the three packages.

	M_{Avg}		$M_{ComBat1}$		$M_{ComBat2}$		$M_{ComBat3}$	
Feature	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*	*ICC*
Mean	<0.006	1.000	0.927	1.000	<0.006	1.000	<0.006	0.999
Minimum	0.029	0.536	0.674	0.547	0.806	0.569	0.636	0.559
Median	0.243	1.000	0.800	1.000	<0.006	1.000	<0.006	0.999
Entropy	<0.006	0.858	0.461	0.854	0.585	0.868	0.048	0.872
GLCM Sum Avg	<0.006	0.433	0.947	0.296	0.708	0.207	0.830	0.325
GLCM Sum Ent	<0.006	0.799	0.316	0.856	0.374	0.858	0.966	0.860
GLCM Diff Ent	0.033	0.830	0.637	0.830	0.643	0.838	0.994	0.836
GLCM Entropy	<0.006	0.035	0.935	0.158	0.917	0.056	0.880	0.046

As shown in Table 5.1, no features reflected significant differences among packages for any of the three ComBat models other than first-order mean and median for the $M_{ComBat2}$ and $M_{ComBat3}$ models. Of these two features, only first-order mean reflected significant differences before ComBat was applied as shown under M_{Avg}. To illustrate the effect of ComBat on the features mean and median, boxplots illustrating the feature value distributions for mean and median before and after ComBat 3 was implemented are shown in Figure 5.2. Similar trends were found among the three ComBat-based methods, so ComBat 3 was chosen as an example.

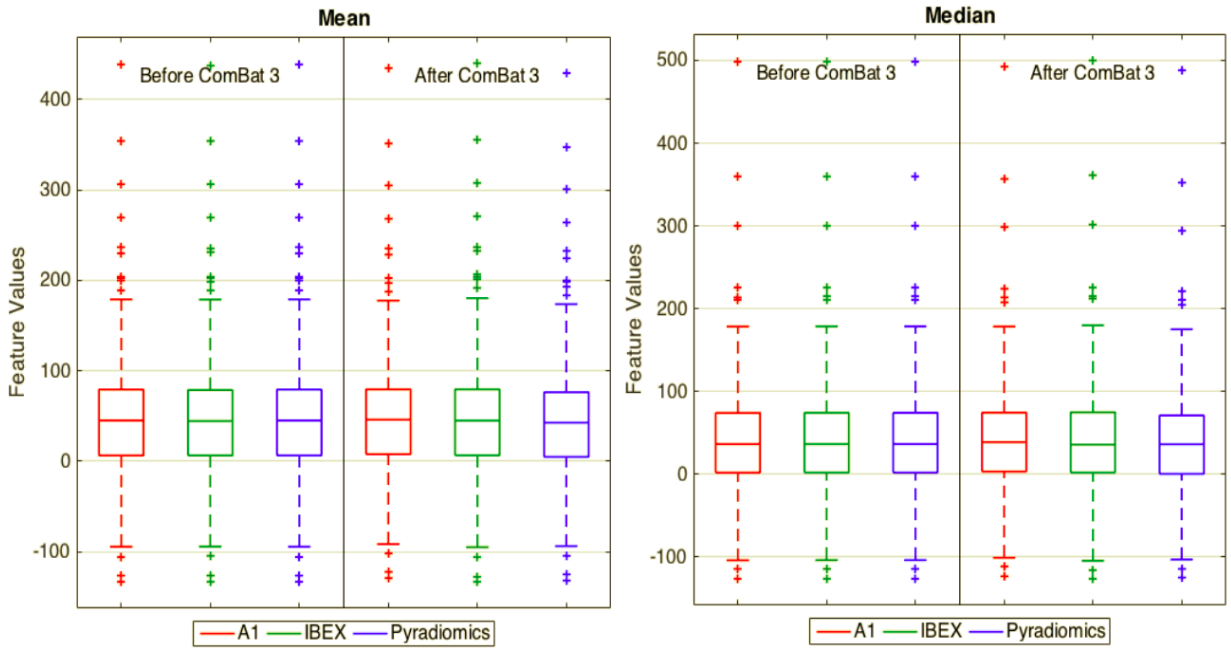


Figure 5.2: The distribution of feature values among the three radiomic software packages before and after ComBat 3 for first-order mean (left) and median (right).

The feature value distributions shown in Figure 5.2 for mean and median do not show any clear discrepancies among the three packages before or after ComBat 3 was applied, and the ICC values for these features were equal to or very close to 1 as shown in Table 5.1. To determine if a consistent bias exists between packages, the relative differences in first-order mean and median were computed for each pairwise combination of the three packages using the first package as the reference: A1 vs IBEX, A1 vs Pyradiomics, and IBEX vs Pyradiomics (Figure 5.3). The differences in feature values were normalized by the absolute value of the feature values from the reference package to negate the influence of the sign of the features from the reference package.

When the relative differences in feature values were calculated for the two features as shown in Figure 5.3, the differences are centered around zero prior to ComBat 3, indicating that any bias in these features between combinations of software packages was relatively small. After ComBat 3 was applied, however, first-order mean calculated with packages A1 was consistently greater than for Pyradiomics as indicated by the boxplot above "A1

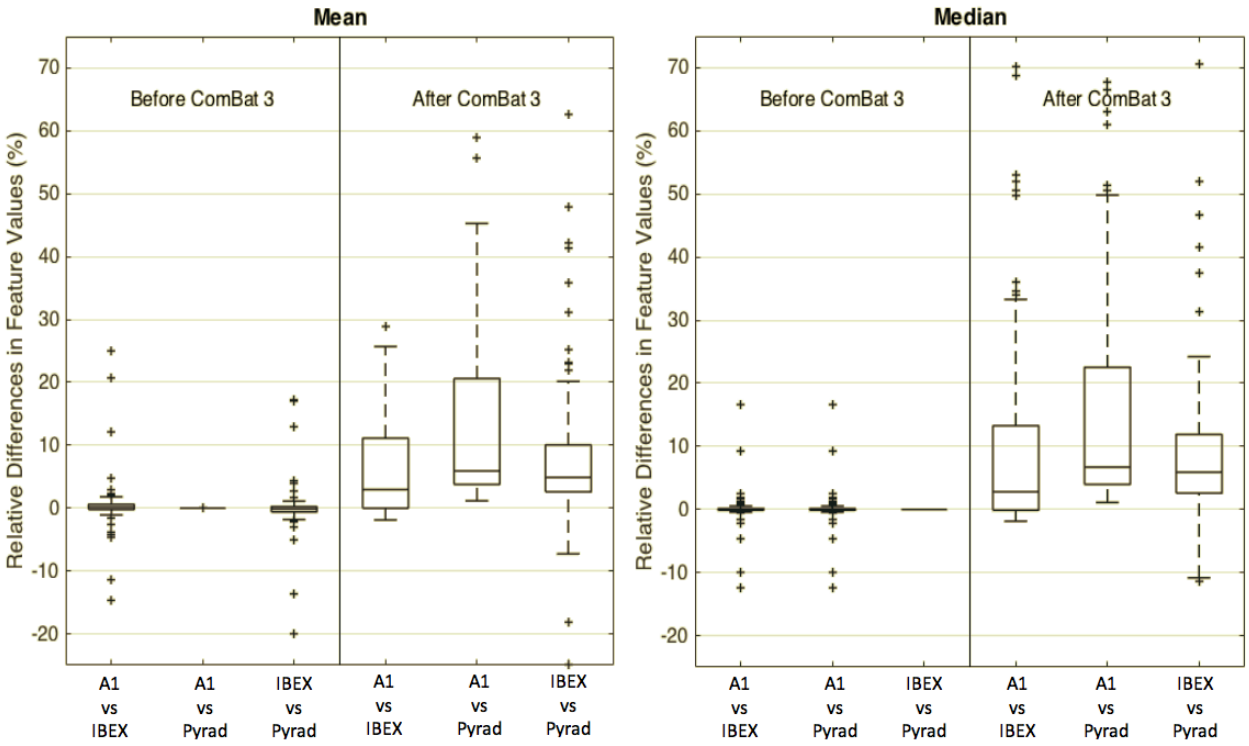


Figure 5.3: Boxplots illustrating the relative differences in feature values across patients between each pairwise combination of packages. Boxplots are shown for first-order mean (left) and median (right) before and after ComBat 3 was implemented. Package A1 was used as reference when compared to packages IBEX or Pyradiomics, while IBEX was used as reference when compared to Pyradiomics. Relative differences in feature values that are located entirely above or below zero indicate consistent biases in feature values between the two software packages in question. Axes were reduced to adequately visualize the distribution of relative differences, thus some outliers are not shown.

vs Pyrad" consisting of entirely positive values. The same trend was seen when comparing first-order median among packages. The ICC values for all ComBat models increased or were within 0.5% of the ICC values for the M_{Avg} models other than GLCM sum average.

5.3.2 Effect of Harmonization on Classification Ability

The changes in eight radiomic features were used along with the MRD to classify patients with and without RP. The p-values indicating which features significantly improved model fit over the MRD alone are shown in Table 5.2 for each of the four models.

Table 5.2: Results from the repeated measures ANOVA illustrating whether a feature from a particular package significantly improved model fit over the MRD by itself for each of the four logistic regression models. Asterisks reflect significant p-values (p<0.002), and cells highlighted in red indicate features that differed in significance for each ComBat-based model when compared to M_{Avg} models.

	M_{Avg}			$M_{ComBat1}$			$M_{ComBat2}$			$M_{ComBat3}$		
Feature	*A1*	*IBEX*	*Pyrad*	*A1*	*IBEX*	*Pyrad*	*A1*	*IBEX*	*Pyrad*	*A1*	*IBEX*	*Pyrad*
Mean	*	*	*	*	*	*	*	*	*	*	*	*
Minimum	0.734	0.092	0.734	0.795	0.207	0.794	0.714	0.038	0.714	0.717	0.036	0.718
Median	*	*	*	*	*	*	*	*	*	*	*	*
Entropy	*	*	*	0.006	*	*	0.003	*	*	*	*	*
GLCM Sum Avg	*	*	*	*	*	*	*	*	*	*	*	*
GLCM Sum Ent	*	*	*	0.008	0.009	*	0.006	0.007	*	0.003	0.003	*
GLCM Diff Ent	0.010	0.010	*	0.010	0.011	*	0.007	0.007	0.003	0.006	0.007	0.004
GLCM Entropy	0.318	0.958	*	*	*	*	*	*	*	*	*	*

The results reported in Table 5.2 show fewer features significantly improving the model fit after ComBat harmonization was applied ($M_{ComBat1}$, $M_{ComBat2}$, and $M_{ComBat3}$) than when ComBat was not used at all. Although, the p-values for all three ComBat-based methods approach significance for the GLCM features when using the relatively conservative Bonferroni correction.

The mean AUC values for each feature and package are shown in Figure 5.4 for the four modeling architectures. The ICC reflecting the relative agreement in mean AUC values among packages and across features is displayed above each plot. Based on the results reported in Figure 5.4, the agreement in the mean AUC values based on the ICC is the greatest for the original models without ComBat (M_{Avg}; ICC: 0.727). The ICC values for all three Combat models were less than that for the M_{Avg} models with clear differences in mean AUC values among packages, particularly for GLCM features.

The mean AUC values for GLCM sum average when calculated using Pyradiomics was nearly 1 for $M_{Combat1}$, while the mean AUC value for this feature was exactly 1 for $M_{ComBat2}$ and $M_{ComBat3}$. The mean AUC values for GLCM entropy were also very close

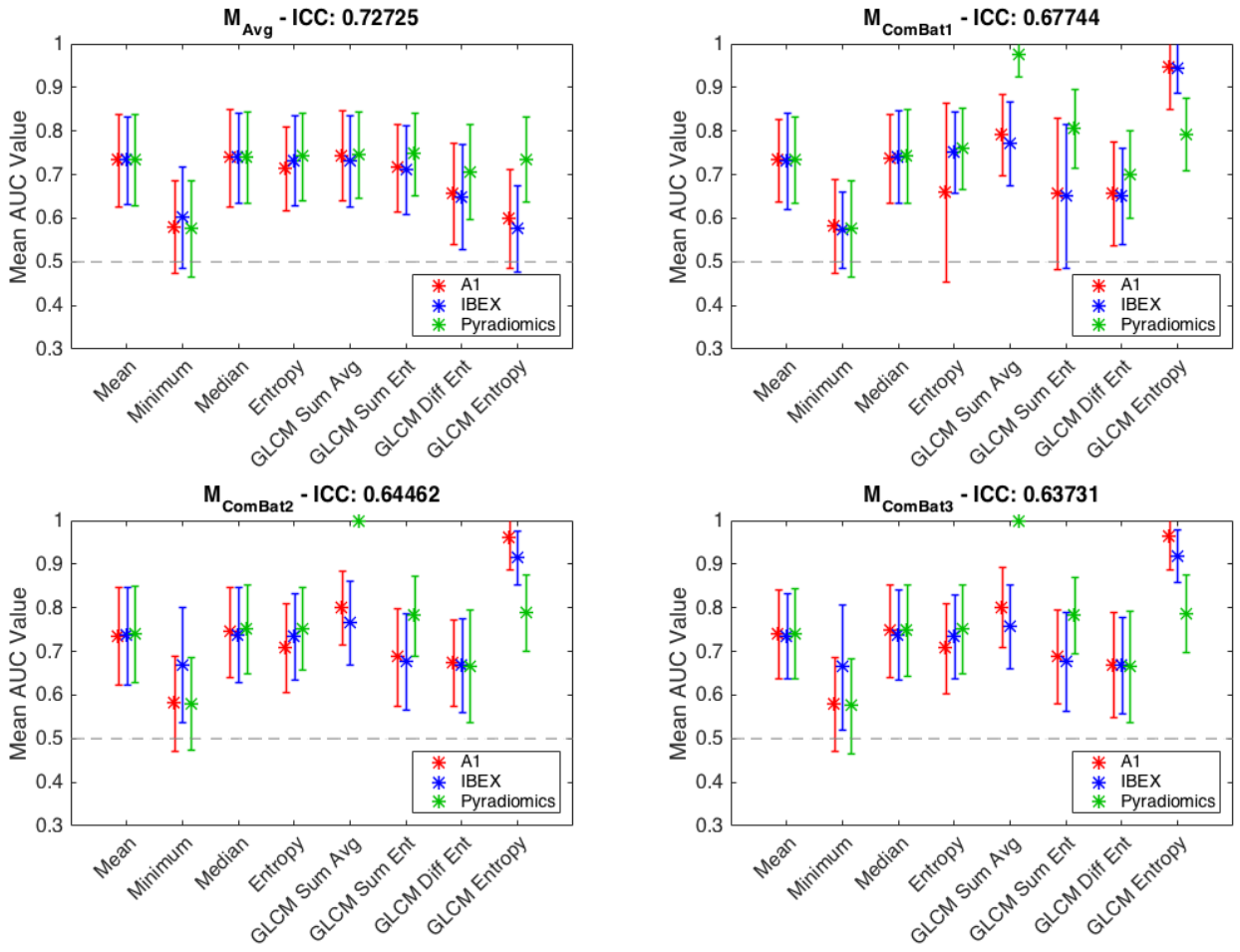


Figure 5.4: Plots illustrating the mean AUC values and corresponding 95% confidence intervals for each feature and software package when the four modeling architectures were used. The ICC value reflecting the agreement in the mean AUC values across features and among packages is reported above the corresponding plot.

to 1 for packages A1 and IBEX for all three ComBat models. To assess the effect of ComBat on these features, the feature value distributions for GLCM sum average and GLCM entropy for RP-positive and RP-negative patients before and after ComBat 3 are shown in Figure 5.5. All three ComBat-based modeling architectures typically resulted in similar findings, so ComBat 3 is used as an example when comparing the effects of ComBat to the non-ComBat-based methods.

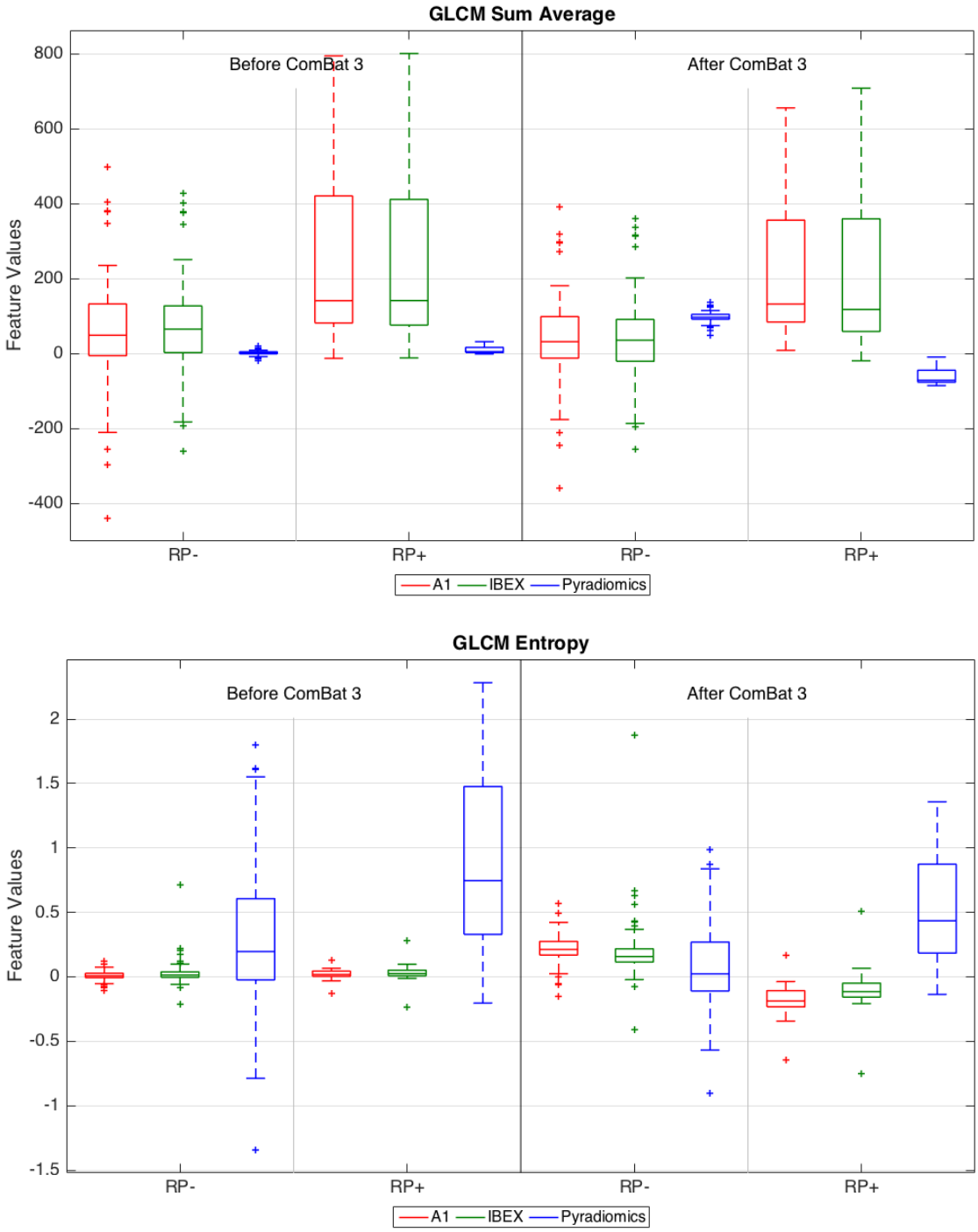


Figure 5.5: Feature value distributions for GLCM sum average (top) and GLCM entropy (bottom) with and without ComBat 3.

As shown in Figure 5.5, when GLCM sum average is calculated with Pyradiomics and when ComBat is not applied, the distributions for RP-positive and RP-negative patients

overlap but are much smaller in scale compared to the distributions for packages A1 and IBEX. When ComBat 3 is applied, the distributions for Pyradiomics' GLCM sum average for patients with and without RP are spread apart until there is no overlap between the two distributions resulting in an AUC value of exactly 1. A similar effect was seen for GLCM entropy because the feature value distributions for packages A1 and IBEX are much smaller than for Pyradiomics. When ComBat 3 is applied, the A1 and IBEX distributions between the two RP statuses spread apart resulting in an increased AUC value close to 1. To quantify the discrimination ability between RP statuses for each of the four models, the K-S test statistic, D^*, is reported in Table 5.3. Feature distributions that reflect greater discrimination between RP statuses after ComBat are highlighted in green (larger D^*), while distributions that reflect less discrimination are highlighted in red (smaller D^*).

Table 5.3: K-S test statistics, D^*, reflecting the degree of overlap in the feature value distributions between RP-positive and RP-negative patients for each of the four logistic regression models. Values of D^* closer to one indicate very little overlap in feature values, while values close to zero indicate a large degree of overlap. Among ComBat models, values of D^* greater than the M_{Avg} models indicating greater separation are highlighted in green, while values of D^* less than the M_{Avg} models are highlighted in red.

	M_{Avg}			$M_{ComBat1}$			$M_{ComBat2}$			$M_{ComBat3}$		
Feature	*A1*	*IBEX*	*Pyrad*	*A1*	*IBEX*	*Pyrad*	*A1*	*IBEX*	*Pyrad*	*A1*	*IBEX*	*Pyrad*
Mean	0.397	0.397	0.397	0.397	0.397	0.397	0.397	0.397	0.397	0.397	0.397	0.409
Minimum	0.315	0.329	0.315	0.465	0.462	0.465	0.291	0.726	0.291	0.303	0.726	0.303
Median	0.379	0.379	0.379	0.379	0.379	0.379	0.379	0.379	0.379	0.394	0.379	0.394
Entropy	0.374	0.418	0.432	0.271	0.468	0.491	0.326	0.444	0.479	0.362	0.421	0.479
GLCM Sum Avg	0.432	0.397	0.432	0.479	0.444	0.953	0.515	0.441	1.000	0.515	0.432	1.000
GLCM Sum Ent	0.403	0.429	0.418	0.288	0.265	0.526	0.291	0.288	0.526	0.338	0.300	0.491
GLCM Diff Ent	0.368	0.368	0.385	0.368	0.356	0.391	0.379	0.391	0.362	0.391	0.391	0.350
GLCM Entropy	0.235	0.259	0.409	0.879	0.926	0.562	0.950	0.926	0.585	0.903	0.832	0.538

The results reported in Table 5.3 indicate that ComBat can emphasize the discrimination between RP-positive and negative patients for several features, particularly GLCM features. For models Combat 1, 2, and 3, the K-S test statistic was maintained or increased for 17, 18, and 21 out of 24 feature-package combinations, respectively. Based on the results shown

in Figure 5.5, the differences in the range of feature value distributions among the three packages were suspected to result in the RP-positive and negative distributions becoming increasingly separated when ComBat was implemented. To assess the effect of differences in scale among batches, the feature values from package A1 were multiplied by 1000, so that the feature value distributions were much broader than those for IBEX and Pyradiomics. ComBat 3 was applied to harmonize the scaled feature values from A1 and for the unscaled features from IBEX and Pyradiomics, and the mean AUC values were recalculated. The mean AUC values for the M_{Avg}, $M_{ComBat3}$, and $M_{ComBat3}$ with the features from A1 multiplied by 1000 are shown in Figure 5.6.

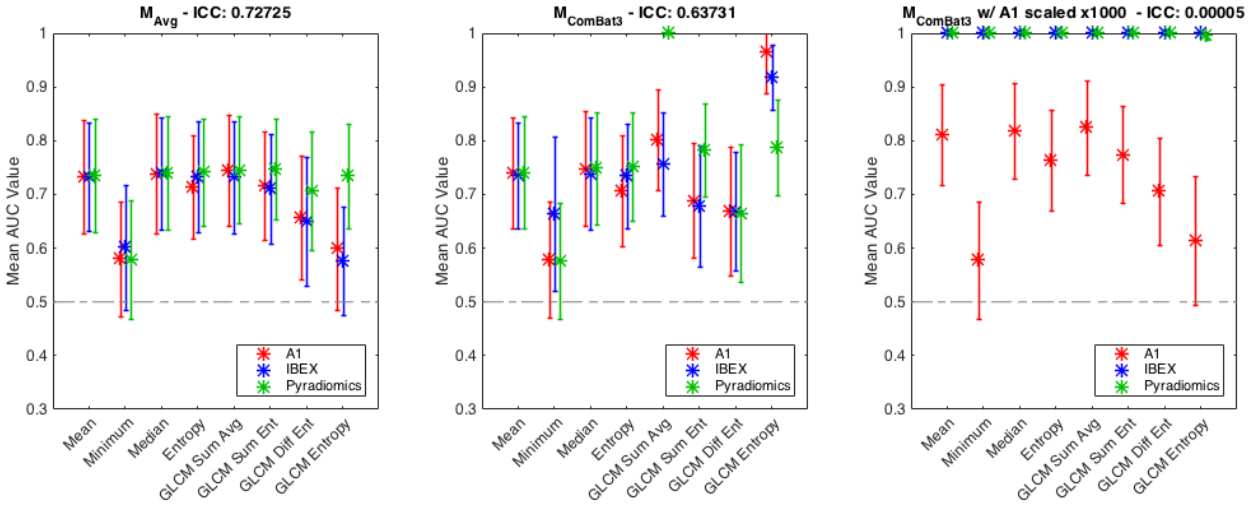


Figure 5.6: Mean AUC values and 95% confidence intervals before ComBat 3 (left), when ComBat 3 is applied (middle), and when ComBat 3 is applied after the features from package A1 are scaled by a factor of 1000.

Based on the mean AUC values shown in Figure 5.6, the effectiveness of ComBat harmonization and its ability to align the feature values while preserving indicators of RP development are contingent on how the ranges of the feature distributions relate among packages. The mean AUC values for all features calculated using IBEX and Pyradiomics are exactly 1, and the mean AUC values for all features calculated with A1 increased aside from first-order minimum and GLCM entropy. Therefore, to remove differences in the scaling of feature distributions, feature values were z-normalized to have a mean of 0 and a standard deviation

of 1. The mean AUC values are shown in Figure 5.7 when ComBat 3 was used alone, when the feature values were z-normalized alone, and when z-normalization was applied prior to Combat 3.

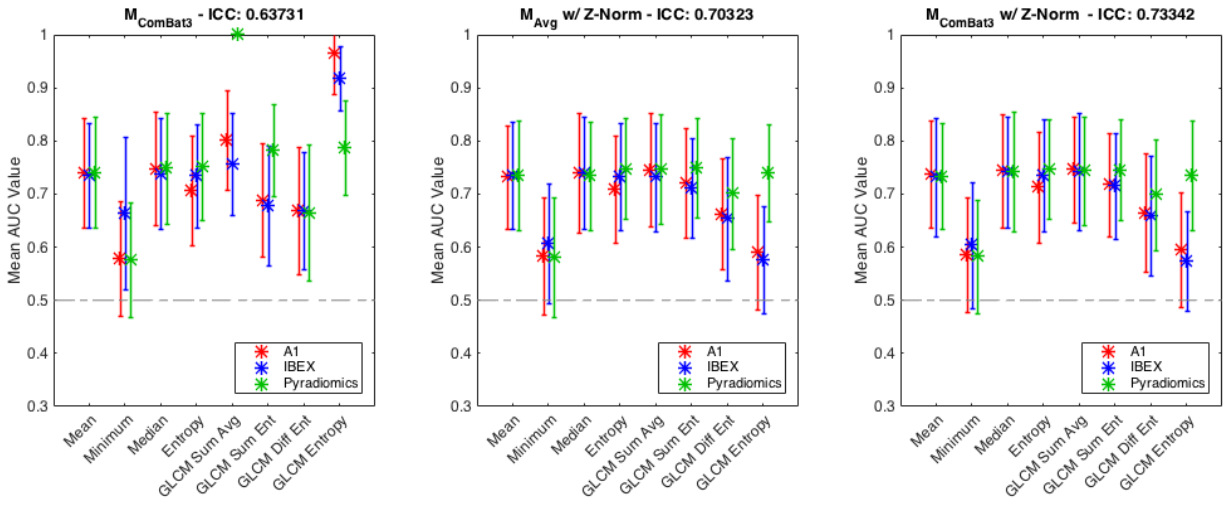


Figure 5.7: The mean AUC values when ComBat 3 is used alone (left), when z-normalization is used alone (middle), and when the feature distributions are z-normalized prior to ComBat 3 (right).

Figure 5.7 shows that when the feature distributions are normalized prior to ComBat 3, the scaling effects that ComBat is susceptible to are removed because the artificial increase in AUC values no longer persist; however, the mean AUC and ICC values when ComBat and z-normalization are used together remain similar to when z-normalization is used alone (ICC: 0.733 and 0.703, respectively). Because z-normalization does not affect the relative position of the feature values corresponding to RP-positive and RP-negative patients within the feature value distributions, the classification performance is the same between M_{Avg} models with and without normalization with small differences attributed to the random sampling of the patients for each training iteration. Additionally, the mean AUC and ICC values are also quite similar to when ComBat is not used as shown in Figure 5.4 under M_{Avg} (ICC: 0.727); however, the ICC was the highest for all models when combining z-normalization with ComBat.

To assess the influence of the covariates and the ability of ComBat to preserve the in-

dicators of these covariates, ComBat was used while RP status was not defined. The mean AUC values for models with and without the control of RP status are shown in Figure 5.8.

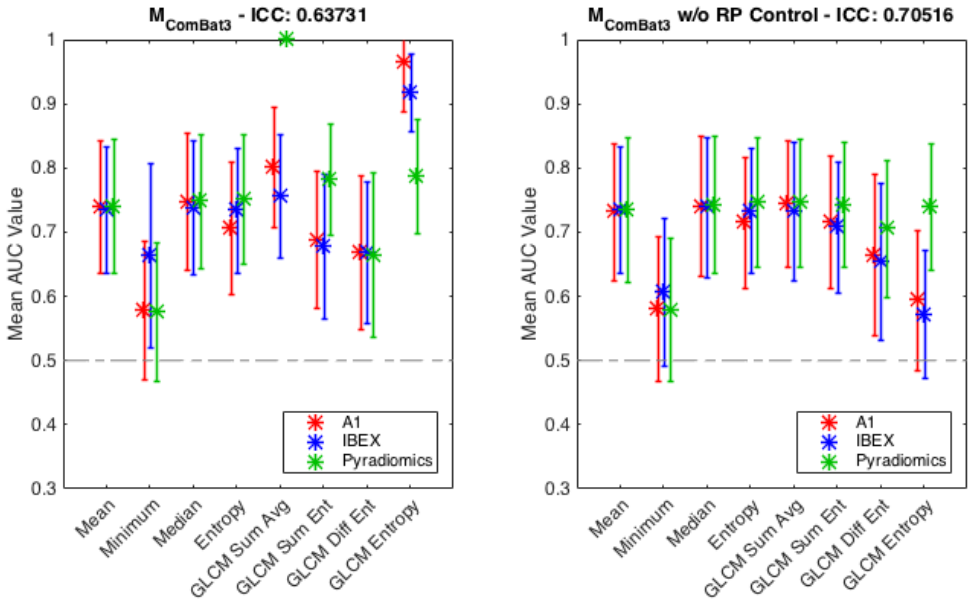


Figure 5.8: Mean AUC values when using ComBat 3 when RP status was (left) and was not (right) used as a controlled covariate.

When RP status was not controlled for as shown on the right side of Figure 5.8, the agreement in the mean AUC value reflected by the ICC increased, but the ICC and the mean AUC values do not greatly differ from when ComBat was not used at all. The ICC value for the M_{Avg} models was 0.727 versus 0.705 when ComBat 3 was applied without RP status preserved. These results indicate that ComBat harmonization only affects the classification ability when covariates are preserved and when the scales of the feature value distributions differ among batches, but ComBat does not harmonize the classification ability among these batches.

5.4 Discussion

The previous chapter illustrated that, when radiomic feature values are calculated using different software packages, the classification of patients with RP may differ based on the

packages used to calculate these features. In the present chapter, various delta-radiomics modeling architectures incorporating ComBat harmonization were developed to mitigate the differences in the feature values across the software packages. The ultimate goal of this study was to assess the potential of ComBat to align the classification ability across packages for each feature used to classify patients with and without RP. Four logistic regression modeling architectures were developed: one using the differences in feature values between pre- and post-RT CT scans as described in the previous chapter (M_{Avg}) and three that applied ComBat harmonization at different points in the feature calculation and modeling workflow. Based on the results shown in Table 5.1, the three ComBat-based models resulted in none of the features reflecting significant differences among software other than first-order mean and median for ComBat 2 and 3; however, the ICC values for these two features were all very close to one for all four models. Because ComBat uses the variability across features for each ROI or patient to estimate the batch effect applied to each feature, features that initially show very high agreement prior to ComBat may be affected by the variability in the remaining features. This variability among features introduces slight biases in features that would otherwise not reflect significant differences such as first-order median (Figure 5.2 and 5.3), although features that initially reflected significant differences and lower agreement prior to ComBat benefitted greatly from this harmonization method as shown in Table 5.1. Among the four GLCM features, three reflected significant differences prior to ComBat, but none of these features reflected significant differences when any of the three ComBat-based methods were implemented. Previous studies have also reported that ComBat harmonization is capable of removing the batch effects and aligning the feature values between batches, but no study to date has shown that the same classification ability is achieved after ComBat harmonization [47,63,64,104,105].

When each feature was added to a logistic regression model containing the MRD, whether a particular feature significantly improved the model fit varied between models that did and did not use ComBat as shown in Table 5.2. Among the three ComBat-based models,

only first-order entropy from package A1 and GLCM difference entropy from Pyradiomics differed, indicating that the point in which ComBat is applied to the feature calculation and modeling workflow does not greatly affect the resultant features. When comparing models before and after ComBat, several features differed among packages on whether the addition of a particular feature significantly improved model fit over using the MRD alone (Table 5.2). Features that were significant before ComBat, such as first-order entropy and GLCM sum entropy, were not significant after ComBat was applied. Other features, such as GLCM entropy, did not improve the model fit prior to ComBat but did for all three ComBat models for packages A1 and Pyradiomics. Because of the sometimes large adjustments made to the feature values with ComBat harmonization, it is not surprising that there was a different set of features associated with RP. On the other hand, the number of features associated with RP agreed among packages for 5/8 features for ComBat 1, 6/8 features for ComBat 2, and 7/8 features for ComBat 3. Models that did not use ComBat harmonization (i.e., M_{Avg}) models agreed for 6/8 features. This increase in agreement across features indicates that ComBat does possess some potential for aligning the features associated with RP.

Features found to be associated with RP as shown in Table 5.2 often translated into which features had a mean AUC value that was significantly greater than 0.5 as shown in Figure 5.4; however, the differences in mean AUC values are plainly visible for the three ComBat-based models. The ICC values for the ComBat models are much less than that for the M_{Avg} models with some mean AUC values approaching or equal to 1. The cause of this variability in classification ability is shown in Figure 5.5: because of the differences in the scale of the feature value distributions among the three software packages, ComBat overemphasis the effect of the RP status covariate and spreads the smaller distributions (e.g., GLCM sum average from Pyradiomics) apart between RP-positive and RP-negative patients (Figure 5.6). This spread results in little to no overlap in the feature value distributions between RP-positive and negative patients, and subsequently the AUC value is very close to 1 for these features. Whitney et al. [105] reported similar findings when attempting

to mitigate the differences in US and Chinese breast DCE-MRI scans: after ComBat was implemented, the AUC value associated with the Chinese database was nearly 1 because of the comparatively smaller distribution in feature values.

When features were normalized to remove the effects of the scaling during the ComBat harmonization process, the agreement in mean AUC values greatly increased compared to when z-normalization was not used (Figure 5.7); however, the agreement in mean AUC values when z-normalization and ComBat were used in combination did not greatly differ from when ComBat was not used at all for the M_{Avg} models. The similarities between the M_{Avg} and $M_{ComBat3}$ models when z-normalization was used indicates that the differences in feature values among radiomic software may be too large and inconsistent across features for ComBat to accurately estimate the batch affect while preserving the indicators of RP status. Normalizing the feature distributions prior to ComBat harmonization may maintain the classification ability for each of the batches, but these results indicate that the classification ability associated with each database will align. Therefore, future studies incorporating ComBat into their investigations should normalize feature values prior to ComBat, so the classification performance is not artificially altered due to differences in the scale of feature value distributions.

To illustrate the influence of the covariates during ComBat, ComBat 3 was used to align the feature values without RP status assigned as a covariate, and the resultant AUC values are shown in Figure 5.7. When RP status was not preserved, the agreement among features greatly increases, and the mean AUC and ICC values converge with those of the M_{Avg} models further indicating that aligning the classification ability among software may be a task outside the scope of what ComBat is designed to accommodate. As previously mentioned, Whitney et al. [105] was capable of aligning the feature values extracted from a US and Chinese database of DCE-MRI breast scans, so they did not reflect significant differences, but the classification of benign and malignant breast lesions significantly improved. Robinsons et al. [104] reported similar findings: when features extracted from mammograms

acquired with different unit manufacturers were harmonized using ComBat, features were no longer dependent on unit manufacturer, but the classification of patient breast cancer risk significantly differed compared to when ComBat was not used. These studies support the notion that ComBat successfully removes the batch effects from features, so they do not reflect significant differences; however, classification performance is not necessarily preserved for any of these studies.

The limitations of ComBat and its application to the present study stem from a number of components of the feature calculation workflow. ComBat was initially developed to mitigate the sometimes random effects introduced by sampling measurements acquired under different circumstances. The additive and multiplicative batch effect estimators, γ and δ, are assumed to abide by a normal and inverse gamma distribution, respectively. In the present study, however, different packages do not necessarily introduce biases of this form for each feature. Because of the systematic differences between packages, consistent biases are introduced into some features, particularly first-order features as shown in Chapter 3. Although, when a delta-radiomics workflow is used, the changes in feature values between time points are quantified, and additive biases resulting from different radiomics software are canceled out. When higher-order features are considered, the variability in feature distributions among packages do not necessarily translate across features. For example, differences in GLCM sum average when calculated with different packages may not relate to the differences in GLCM entropy when calculated with the same packages. The assumptions ComBat makes during the harmonization process further enforce the notion that ComBat may not be capable of mitigating the differences in radiomic software packages while properly preserving RP status.

Previous radiomics studies have shown ComBat to successfully align feature values across batches while preserving the desired covariates [47,63,64,104,105]. These studies, however, attempted to mitigate differences in imaging unit manufacturer and PET scans acquired at different institutions to allow for the combination of databases acquired under different conditions. While these are issues that need to be addressed, they may align with the

assumptions ComBat makes during harmonization, and they also have a comparatively small effect on the resultant features relative to the effects of calculating radiomic features with different software packages. As shown in Chapter 3, using different radiomic software can affect the feature values extracted from the same images by several orders of magnitude, particularly for higher-order features. Therefore, when different radiomics software are used to process medical images, proper application of delta radiomics and assessing the change in feature values rather than the feature values themselves may be the best option to maximize the reproducibility of radiomic research. For more subtle and consistent batch effects, such as those investigated using the cadaveric liver in Chapter 2, ComBat harmonization may be more appropriate; however, additional research should be conducted to determine the limitations at which ComBat is no longer applicable.

5.5 Conclusion

This chapter attempted to use various logistic regression modeling architectures using ComBat harmonization to mitigate the differences in the radiomic feature calculation process with the overall aim of aligning the classification ability among software packages. When ComBat was implemented, fewer features reflected significant differences among the software packages used to calculate them. When these harmonized features were applied to the clinical task of classifying patients with RP, the agreement in mean AUC value was greatly reduced. These differences were found to be a result of ComBat's susceptibility to differences in the scales of the feature value distributions among packages. When z-normalization was applied prior to ComBat, the mean AUC and ICC values increased to approach those obtained when ComBat was not utilized indicating that aligning the classification ability among software packages may be outside the scope of what ComBat is designed to accommodate. While ComBat has shown great potential in removing the batch effects introduced by differences in repeated measurements in other radiomics studies, additional research should be conducted to assess the strengths and limitations of these harmonization methods.

CHAPTER 6

CONCLUSIONS

Throughout this dissertation, the variability in different components of the radiomics workflow were quantified, and the effect of this variability on a clinically relevant classification task was assessed. To achieve this goal, the dependence of radiomic features on a number of CT image acquisition and reconstruction parameters was investigated when scans were acquired of a cadaveric liver, and methods of mitigating these differences were assessed (Chapter 2). In addition, we calculated radiomic features using medical images from various imaging modalities and tissue types using five radiomics software packages and showed that feature values can differ by up to several orders of magnitude among packages (Chapter 3). Extrapolating on the results from Chapter 3, we used a subset of these software packages to calculate radiomic features from serial thoracic CT scans of patients undergoing radiation therapy and attempted to classify patients with and without radiation pneumonitis when using the same statistical methods. We found that classification ability differed among the software packages used to calculate the radiomic features, particularly for higher-order features (Chapter 4). Finally, we assessed the potential of a harmonization method outlined in Chapter 2 to align the classification ability among radiomic software (Chapter 5). Our results demonstrated that appropriately implementing a delta-radiomics modeling structure minimizes the variability due to differences in the radiomics workflow. However, additional work must be done to advance the overall goal of making radiomics research more easily reproduced and validated.

Understanding the effect of each component of the radiomics workflow on the resultant feature values and ultimately on the application of these features as imaging biomarkers is essential to the translation of radiomics research into clinical practice. The variability in radiomic features due to differences in image acquisition and reconstruction parameters is well documented; however, such studies focused primarily on phantom images or ensembles of patient images and assessing means of mitigating these differences had not been thoroughly

investigated [2,30,36,37,41,53,56,74-78]. As shown in Chapter 2, some imaging parameters such as field of view and image reconstruction plane altered the resultant feature values by more than 100% when these features were extracted from the CT scans of a single cadaveric liver. Other imaging parameters such as tube voltage or iDose reconstruction level resulted in relatively small biases in the subsequent feature values. Understanding which features are robust to each image acquisition and reconstruction parameter may allow for future investigators to better assess the allowances that can be made when curating databases of patient images for radiomics research. Otherwise, researchers may also implement any of the harmonization methods analyzed in Chapter 2, particularly ComBat harmonization. While ComBat resulted in complete harmonization in all features regardless of the imaging parameter analyzed, the remaining harmonization methods reflected some potential to mitigate the effects of different imaging parameters. For example, pixel size resampling and Butterworth filtering combined removed the dependence of radiomic features on iDose reconstruction level, while histogram normalization limited the influence of differences in slice thickness on these features.

In addition to image acquisition and reconstruction parameters, this dissertation also quantified the effect of variations during the calculation of radiomic features [37]. Previous studies have illustrated the need for greater standardization in the computation of radiomic features; however, the dependence of radiomic features on the feature calculation process and the algorithms used to calculate these features had not been adequately quantified [38-44,79,80]. In Chapter 3, feature values were extracted from mammograms, head and neck CT scans, and serial breast MRI scans using a number of radiomic software packages, and the resultant feature values were compared across packages. The results indicated that differences in the feature calculation process can alter the feature values by several orders of magnitude due to differences in image importation and preprocessing, algorithm implementation, and feature-specific parameters (e.g., GLCM parameters) [37]. Properly implementing a delta-radiomics workflow and comparing features extracted from the same patient images

acquired at different time points as was done with the breast MRI database may limit the differences in radiomics software. Because of the large dependency of radiomic features on the feature calculation process, understanding the implications of using different radiomics software packages is vital to the reproducibility of radiomics research. While this dissertation does not aim to demonstrate that one package is objectively better than another for all purposes, it does illustrate that some packages may be more appropriate to analyze images of a particular modality or tissue type based on how the packages were designed. Based on these results, researchers should think critically about which package to use in their own research. Researchers should use open-source or freely-available packages so that their work can be reproduced and validated. If freely-available software cannot be used, details of the feature calculation and algorithm implementation should be made available. If researchers intend on reproducing a particular study, they should understand the consequences of using a different software package than what was used in the initial investigation.

While understanding the dependence of radiomic features on various components of the radiomics workflow is important, whether these dependencies affect classification ability when applied to a particular clinical task should also be assessed. Therefore, Chapter 4 of this dissertation aimed to analyze the differences in classification ability when various radiomics software packages were used to classify patients with radiation pneumonitis [103]. In this chapter, changes in feature values between time points were quantified using each software package, and the dependence of the classification ability for each feature on the software package used to calculate these features was assessed. While first-order features reflected relatively strong agreement in classification ability among packages, higher-order features began to differ among packages even though a delta-radiomics approach was used. The deviation in classification ability among higher-order features indicates that using different software to calculate radiomic features may limit the reproducibility of radiomics research.

To assess the potential of removing the variability in radiomics software that results in differences in classification performance, a number of mathematical modeling techniques

were analyzed in Chapter 5. In this chapter, the feature values resulting from the delta-radiomics scheme outlined in Chapter 4 were adjusted using ComBat harmonization applied at various components of the feature calculation and modeling workflow. Previous investigations have shown that ComBat harmonization can sufficiently remove the batch effects resulting from medical images acquired on different days, with different imaging protocols, or using different imaging unit manufacturers [47,63,64,104,105]. When ComBat was employed in this dissertation research, however, no increase in agreement based on the classification ability among software packages was realized. Because of the relatively large dependence that radiomic features have on the software used to calculate them, we hypothesized that the task of removing the differences in software packages to align classification ability may be outside the scope of what ComBat is designed to accommodate. Properly implementing a delta-radiomics workflow alone may adequately remove some of the effects of differing software packages.

While the harmonization methods analyzed in this dissertation did not sufficiently remove the differences in radiomics software packages for this classification task, Chapter 5 does illustrate a theme that has been highlighted throughout this dissertation work: each aspect of radiomics workflow must be properly assessed for appropriateness and scope when conducting radiomics research. Although previous radiomics investigations have reported promising results when detecting various diseases or segmenting different tissues, future researchers cannot assume that these results will translate across institutions if details of the methods are altered. Chapter 2 outlined the lack of reproducibility in radiomics research due to differences in imaging parameters, while Chapters 3 and 4 reported the same lack of reproducibility due to differences in the feature calculation process. In Chapter 5, we found that ComBat harmonization is not appropriate for every harmonization task despite previously reported publications finding ComBat to be highly successful for some tasks. Therefore, researchers implementing radiomics into their studies should properly assess whether their patient databases, radiomics software packages, and modeling methods are appropriate for

the clinical task at hand.

Future Work

This dissertation work provides an avenue for a number of future studies to further aid in the harmonization of radiomics research. Chapter 4 of this work reported that using different radiomics software packages to process the same medical images can impact the classification ability of the resultant features when applied to a clinical task. Using different software packages, however, can alter the feature values by several orders of magnitude as shown in Chapter 3. Future work could investigate if more subtle changes such as those introduced by differences in image acquisition and reconstruction parameters have as large of an impact on the classification ability of each feature. When assessing the dependence of radiomic features extracted from CT scans of a cadaveric liver (Chapter 2), only changing the field of view or image reconstruction plane resulted in a median relative change in feature value of greater than 5%. The remaining imaging parameters were affected by a much smaller degree: reducing the CTDIvol, changing the iDose reconstruction level, and changing the convolution kernel resulted in median relative changes less than 1%. Because the biases resulting from differences in imaging parameters are relatively small, there may be no difference in classification ability, but additional work should be done to assess this claim.

Chapter 5 determined that ComBat harmonization is not appropriate for all applications and is limited based on the harmonization task at hand. Additional research should be conducted to determine the subset of clinical applications that would benefit from this method along with the additional harmonization methods outlined in Chapter 2. Whitney et al. [105] found that ComBat harmonization successfully removes the batch effects introduced when DCE-MRI scans are acquired in different countries while preserving the indicators of breast cancer risk; however, when applied to the more challenging task of removing the differences in radiomics software packages to align classification performance as was done in

Chapter 5, complete harmonization was not achieved. To fully understand the capabilities of ComBat harmonization, future work could determine if ComBat in addition to other harmonization techniques could properly align the classification ability among radiomics research. Radiomics research has often used the AUC value or other metrics (e.g., Sørensen-Dice coefficient) to quantify the success and accuracy of their modeling methods. If the methods of a radiomics investigation are duplicated, and a different AUC value is obtained, one could not necessarily state that the investigation has been successfully reproduced. Therefore, it may not be enough to harmonize radiomics features such that they do not reflect significant differences, but additional research must be conducted to standardize all components of the radiomics workflow. Furthermore, the potential of the remaining harmonization methods used in Chapter 2 (histogram normalization, resampling, filtering, and resampling and filtering combined) should be investigated when these methods are applied to a classification task. Because these methods do not attempt to preserve the imaging biomarkers that would be used to classify patients or detect diseases, the utility of these harmonization methods may end at aligning feature values without clinical application. Without the preservation of the quantitative indicators of disease development or tissue structure, some harmonization methods may reduce the differences in feature values between disease states and consequently limit the classification performance when these harmonization methods are used.

Finally, convolutional neural networks (CNNs) have become much more popular in recent years as well, but machine learning research using CNNs is often plagued by the same variability as radiomics research [107-109]. Therefore, a parallel investigation should be conducted comparing the effect of different CNN architectures on classification ability and how this variability in classification across architectures compares to that reported in this dissertation. Understanding the variability in machine learning research using radiomics or CNNs may aid in allowing for this work to be more easily translated across institutions.

This dissertation demonstrated the variability that can be introduced into radiomics research due to differences in each component of the radiomics workflow. Fully understand-

ing the implications of altering the image acquisition and reconstruction of medical images or changing the feature calculation process will aid investigators in determining methods appropriate for their radiomics research. While a delta-radiomics workflow has shown to mitigate some of the dependency that radiomic features have on these processes, additional harmonization methods such as ComBat can further align radiomic features. However, these harmonization methods could not completely harmonize all aspects of radiomics research, and further work must be conducted to achieve more reproducible radiomics studies. Assessing methods to make radiomics research more translatable across institutions may allow for this research to be more easily implemented into clinical practice and aid medical professionals in making potentially life-saving decisions.

APPENDIX

7.1 First-Order Gray-Level Histogram Features

First-order gray-level histogram features quantify the aspects of the pixel value histogram. The equations for the 22 first-order features used are shown below, where G is an $n \times m$ image matrix region with arbitrary range of pixel values. $G_{k\%}$ is the pixel value in which $k\%$ of pixels have a value less than $G_{k\%}$. q is a one-dimensional vector of length l where $q(i)$ is the relative frequency of a particular pixel value, $i \in \{1, \ldots, l\}$, in G.

$$Mean\colon \mu_G = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} G(i,j) \tag{7.1}$$

$$Median\colon median(G) \tag{7.2}$$

$$Mode\colon mode(G) \tag{7.3}$$

$$Maximum\colon max(G) \tag{7.4}$$

$$Minimum\colon min(G) \tag{7.5}$$

$$Mean\,Absolute\,Deviation\colon \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} |G(i,j) - \mu_G| \tag{7.6}$$

$$\begin{aligned} Median\,Absolute\,Deviation\colon & median(|G(i,j) - median(G)|), \\ & i \in \{i, \ldots, n\}\,and\,j \in \{1, \ldots, m\} \end{aligned} \tag{7.7}$$

$$Geometric\,Mean\colon exp\Big[\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} ln[G(i,j)]\Big] \tag{7.8}$$

$$Range\colon max(G) - min(G) \tag{7.9}$$

$$Interquartile\,Range\colon G_{75\%} - G_{25\%} \tag{7.10}$$

$$Standard\,Deviation\colon \sigma_G = \sqrt{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}[G(i,j) - \mu_G]^2} \tag{7.11}$$

$$Skewness\colon \frac{\sqrt{nm(nm-1)}}{nm-2} \times \frac{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}[G(i,j)-\mu_G]^3}{(\sigma_G)^3} \tag{7.12}$$

$$\begin{aligned} Kurtosis\colon & \frac{nm-1}{(nm-2)(nm-3)}[(nm+1)k-3(nm-1)]+3 \\ & where\, k = \frac{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}[G(i,j)-\mu_G]^4}{(\sigma_G)^2} \end{aligned} \tag{7.13}$$

$$Energy\colon \sum_{i=1}^{l} q(i) \tag{7.14}$$

$$Entropy\colon -\sum_{i=1}^{l} q(i) log_2[q(i)] \tag{7.15}$$

$$\begin{aligned} Binned\ Entropy\colon & Entropy\ calculated\ after\ interpolated\ G\ into\ 256 \\ & gray\text{-}level\ bins \end{aligned} \tag{7.16}$$

$$5\%\, Quantile\colon G_{5\%} \tag{7.17}$$

$$30\%\, Quantile\colon G_{30\%} \tag{7.18}$$

$$60\%\, Quantile\colon G_{60\%} \tag{7.19}$$

$$90\%\, Quantile\colon G_{90\%} \tag{7.20}$$

$$70\%\, Balance\colon \frac{G_{70\%}-\mu_G}{\mu_G - G_{30\%}} \tag{7.21}$$

$$95\%\, Balance\colon \frac{G_{95\%}-\mu_G}{\mu_G - G_{5\%}} \tag{7.22}$$

7.2 Gray-Level Co-OccurrenceMatrix (GLCM) Features

GLCM features are calculated after constructing the joint-probability matrix which quantifies how often a pixel with a value a appears some distance d from another pixel with value b at some angle θ (Figure 7.1). In this work, the distance between the neighboring pixels in question is 1 (only immediately adjacent pixels are considered). In order to have

directionally-invariant GLCM features, GLCM features are calculated using angles $\theta = 0^o$, 45^o, 90^o, and 135^o and then averaged over all four angles. The gray-level limit can also be modified to contain the full range of pixel values in an image region [max(G), min(G)] or a predetermined range can be used. The number of gray levels, or binning, can also be modified prior to GLCM construction such that the GLCM will have a specified number of indices. Here, we use the gray-level limits of [-1500, 1500] and the number of gray levels was 3001 by default. The GLCM is normalized by the sum of all values in the GLCM resulting in the joint probability distribution, which is therefore independent of the number of pixels in the image region.

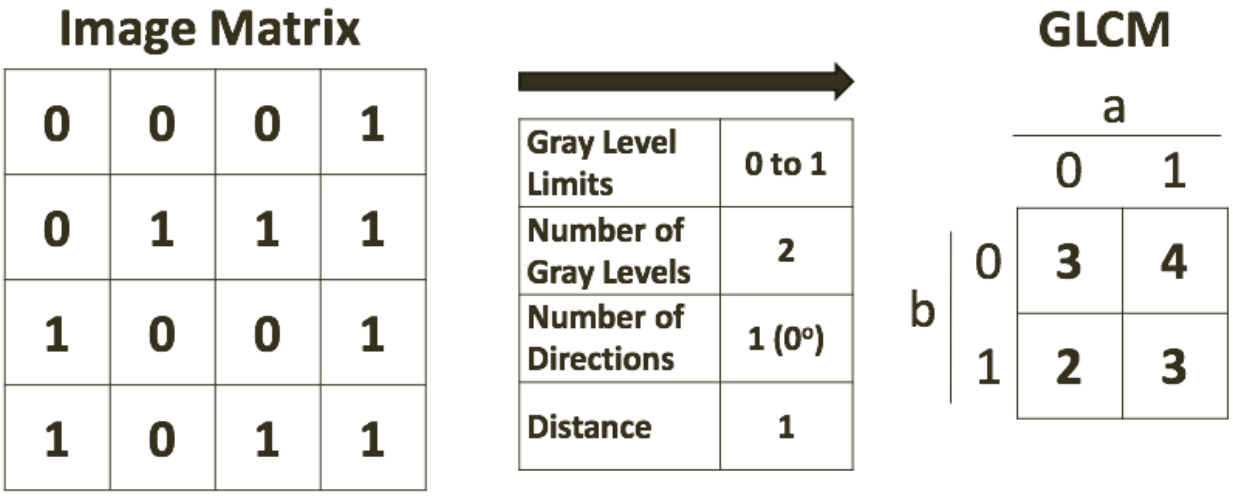


Figure 7.1: The gray-level co-occurrence matrix constructed from a binary image matrix. The construction of the GLCM is dependent on a number of parameters including the gray level limits, the number of gray levels, and the direction and distance between the two pixels in question.

GLCM feature definitions are derived from Haralick et al. [14] and outlined in full in the *Handbook of Computer Vision and Applications* [15]. In these equations, for a GLCM P with n rows and columns derived from an image G with N_G distinct gray levels, $p(i, j)$ is the $(i, j)^{th}$ index of the joint probability matrix,

$$p = \frac{P}{\sum_{i=1}^{n}\sum_{j=1}^{n} P(i,j)} \tag{7.23}$$

$$p_x(i) = \sum_{j=1}^{n} P(i,j) \; and \; p_y(j) = \sum_{i=1}^{n} P(i,j) \tag{7.24}$$

$$\mu_x = \sum_{i=1}^{n} i p_x(i) \; and \; \mu_y = \sum_{j=1}^{n} j p_y(j) \tag{7.25}$$

$$p_{x+y}(k) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(i,j) \; for \; (i+j) = k; \; k \in \{2,3,\ldots,2N_G\} \tag{7.26}$$

$$p_{x-y}(k) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(i,j) \; for \; |i-j| = k; \; k \in \{0,1,\ldots,N_G-1\} \tag{7.27}$$

The 14 texture features calculated using these matrices:

$$Autocorrelation\colon \sum_{i=1}^{n}\sum_{j=1}^{n} ijp(i,j) \tag{7.28}$$

$$Inertia\colon \sum_{i=1}^{n}\sum_{j=1}^{n} [i-j]^2 p(i,j) \tag{7.29}$$

$$Inverse\ Difference\ Moment\colon \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{p(i,j)}{1+(i-j)^2} \tag{7.30}$$

$$Energy\colon \sum_{i=1}^{n}\sum_{j=1}^{n} [p(i,j)]^2 \tag{7.31}$$

$$Entropy\colon \sum_{i=1}^{n}\sum_{j=1}^{n} p(i,j) ln[p(i,j)] \tag{7.32}$$

$$Contrast\colon \sum_{k=0}^{N_g-1} k^2 p_{x-y}(k) \tag{7.33}$$

$$Sum\ of\ Squares\ Variance\colon \sum_{i=1}^{n}\sum_{j=1}^{n} [i-\mu_x]^2 p(i,j) \tag{7.34}$$

$$Sum\ Average\colon \sum_{k=2}^{2N_g} kp_{x+y}(k) \tag{7.35}$$

$$Sum\ Variance\colon \sum_{k=2}^{2N_g} (k - [SumEntropy])^2 p_{x+y}(k) \tag{7.36}$$

$$Sum\ Entropy\colon -\sum_{k=2}^{2N_g} p_{x+y}(k) ln[p_{x+y}(k)] \tag{7.37}$$

$$Difference\ Average\colon \sum_{k=0}^{N_g-1} kp_{x-y}(k) \tag{7.38}$$

$$Difference\ Variance\colon \sum_{k=0}^{N_g-1} (k - [DifferenceEntropy])^2 p_{x-y}(k) \tag{7.39}$$

$$Difference\ Entropy\colon -\sum_{k=0}^{N_g-1} p_{x-y}(k) ln[p_{x-y}(k)] \tag{7.40}$$

$$Absolute\ Value\colon \sum_{i=1}^{n}\sum_{j=1}^{n} |i-j| p(i,j) \tag{7.41}$$

7.3 Fractal Features

Fractal analysis and fractal features characterize the self-similarity of an image at different scales with the aims of quantifying the detail of a region [16]. This self-similarity is often characterized by the fractal dimension which compares the number of self-similar pieces contained within the image with the magnification factor required to scale the self-similar pieces to obtain the original image. For example, a square can be broken up into 4 self-similar pieces with a magnification factor of two. The fractal dimension is subsequently quantified using the equation:

$$Dimension = \frac{log(\#\,Self-Similar\,Pieces)}{log(Magnification\,Factor)} = \frac{log(4)}{log(2)} = 2 \tag{7.42}$$

A square therefore has a fractal dimension of 2, a cube a fractal dimension of 3, and so on. The fractal dimension quantifies how "complicated" a particular image is and how many points generally fall within that image set. For gray-scale images such as those used in medical imaging, the number of self-similar pieces and the magnification factor are more complicated to compute, and pixel values are often conceptualized as a the "heights" of the pixels in a third dimension normal to the imaging plane. The heights of these columns and how self-similar they are can then be characterized by a number of methods including the blanket method, Brownian motion method, and the box-counting method.

7.3.1 Blanket Method

The blanket method projects two "blankets" above and below the three-dimensional conceptualization of an image region with the pixel intensities corresponding to the height in the third dimension [17]. The intensity values corresponding to the two blankets are separated by no more than some distance, ρ. The surface area of the region describing the complexity of the image is quantified by calculating the area between the two sheets which fully encapsulate the image region. The surface area, $S(\rho)$, of the region can then be estimated and related to the fractal dimension, D, using the following equation:

$$ln(S(\rho)) \propto (3 - D)ln(\rho) \tag{7.43}$$

7.3.2 Brownian Motion Method

The Brownian motion method uses the variability in neighboring pixel values as random walks with the randomness of these walks characterized by a fractal Brownian motion model [18]. The texture of the image region is assessed by the variability of the walks between directly adjacent pixels and pixels separated by a specified distance. The average differences in pixel intensities separated by all distances from d_1 to d_n are calculated resulting in what is known as the normalized multiscale intensity difference (NMSID) vector. Similar to the

Blanket method, the NMSID is used to derive the fractal dimension using the following equation:

$$ln(NMSID) = (3 - D)ln(NPN) + k \tag{7.44}$$

Here, NPN is the number of pairs of pixels separated by each distance r_i, and k is an arbitrary constant.

7.3.3 Box-Counting Method

The box-counting method also calculates the surface area, $S(\varepsilon)$, of an image region similarly to the Blanket method; however, the box-counting method calculates $S(\varepsilon)$ at various resolutions, which can be used to quantify the fractal dimension at different levels of image discretization [19,20]. The fractal dimension D is calculated by

$$ln(S(\varepsilon)) = (2 - D)ln(\varepsilon) + k \tag{7.45}$$

where ε is the resampling factor for each image resolution, and k is an arbitrary constant. The fractal dimension can then be calculated over various resampling factors to characterize the fine and course image detail (i.e., small values of ε for high-resolution region detail and large ε for low-resolution detail). In the present study, for an image region with size $n \times m$ pixels, ε_{max} was determined to be a fourth of the size of the dimensions of the image region ($\varepsilon_{max} = floor[min(m, n)/4]$). Then, the image region is divided into all possible resampling factors (ε_i spans all factors of ε_{max}) such that each resampling factor subdivides the region into equal parts. For example, for an image with dimensions of 32×32 pixels, resampling factors of $\varepsilon_1 = 32 \times 32$ pixels, $\varepsilon_2 = 16 \times 16$ pixels, $\varepsilon_4 = 8 \times 8$ pixels, and $\varepsilon_{max} = 4 \times 4$ pixels. For an image with dimensions of 40×40 pixels, ε has all values {1,2,5,8,10} and so on. In this study, images that did not have at least three discrete integer resampling factors, bilinear interpolation was used to obtain resampling factors that were close to $\varepsilon =$

{1,2,4,8} depending on the dimensions of the image region. In this way, both coarse and fine dimensions could be calculated. For example, an image with dimensions of 34×34 pixels would have integer resampling factors of $\varepsilon = 1$ and $\varepsilon = 2$ corresponding to pixel dimensions of 34×34 pixels and 17×17 pixels, respectively. To obtain the two additional resampling factors, the image was bilinearly interpolated to have dimensions of 8×8 pixels ($\varepsilon = 4.25$) and 4×4 pixels ($\varepsilon = 8.5$).

The surface area $S(\varepsilon)$ is calculated for each resampling factor, and the total fractal dimension over all resampling factors is calculated by fitting the $S(\varepsilon)$ and ε to Eq. 7.45. The fine and coarse fractal dimensions were determined by calculating a cutoff threshold b by fitting $S(\varepsilon)$ and ε to to the following quadratic equation:

$$ln(S(\varepsilon)) = k_0 + k_1 ln(\varepsilon) + k_2[ln(\varepsilon)]^2 \tag{7.46}$$

The course and fine fractal dimensions was calculated using the threshold:

$$b = \frac{1}{2}\frac{D - k_1}{k_2} \tag{7.47}$$

where the coarse dimension was quantified for all $S(\varepsilon)$ and ε where $\varepsilon > b$, and the fine dimension was quantified for all $S(\varepsilon)$ and ε where $\varepsilon < b$. This methodology resulted in three measures of the fractal dimension: fine and coarse fractal dimensions as well as the fractal dimension at all resampling factors.

7.4 Fourier Features

Fourier analysis uses a rotationally-invariant Fourier transform to decompose an image region into the sum of sine and cosine waves of varying amplitude, frequency, and phase reflecting the relative fluctuations of pixel values in that region [21,22]. The Fourier transform of a discrete two-dimensional image is

$$F(u,v) = \int\int f(x,y)e^{-j2\pi(ux+vy)}dxdy \tag{7.48}$$

where u and v are the spatial frequency components of the image in the x and y direction, respectively. The transformed image reflects the spatial frequency of pixel values in the various spatial directions. Images that contain fine detail and closely spaced pixels of different values will have a larger amount of high-frequency data contained in the corresponding Fourier transform, while low-contrast "smooth" textural patterns will be reflected in the low-frequency portions of the Fourier transform. In this study, a total of 17 Fourier features were calculated. The various components of the Fourier transform can be characterized to describe the original image such as the root-mean square (RMS) and first moment of the noise power spectrum (FMP).

$$RMS = \sqrt{\int\int |F(u,v)|^2 dudv} \tag{7.49}$$

$$FMP = \frac{\int\int \sqrt{u^2+v^2}|F(u,v)|^2 dudv}{\int\int |F(u,v)|^2 dudv} \tag{7.50}$$

In addition to these two features, the energy of the rotationally-invariant transformed image was calculated in different frequency components to fully characterize the spatial frequencies in each direction. The transformed image was subdivided into four regions using three concentric circles centered on (u=0,v=0) corresponding to the low-, moderately low-, moderately high-, and high-frequency regions. The energy is then calculated using all four subsections combined, the high- and moderately high-frequency sections combined, the low- and moderately low-frequency sections combined, and then each section separately, resulting in seven rotationally-invariant Fourier features.

The remaining eight rotationally dependent Fourier features were calculated by dividing the transformed image into eight equal sectors emanating from (u=0,v=0). Each sector has an angel of 45^o at the center, and from each of these sectors, the energy is calculated.

7.5 Laws' Filter Features

Laws' filter features emphasize region microstructure by convolving image regions with combinations of 2-dimensional filter vectors [23]. In this study, five vectors were used to construct 25 two-dimensional image filters:

Level: L5 = [1 4 6 4 1]

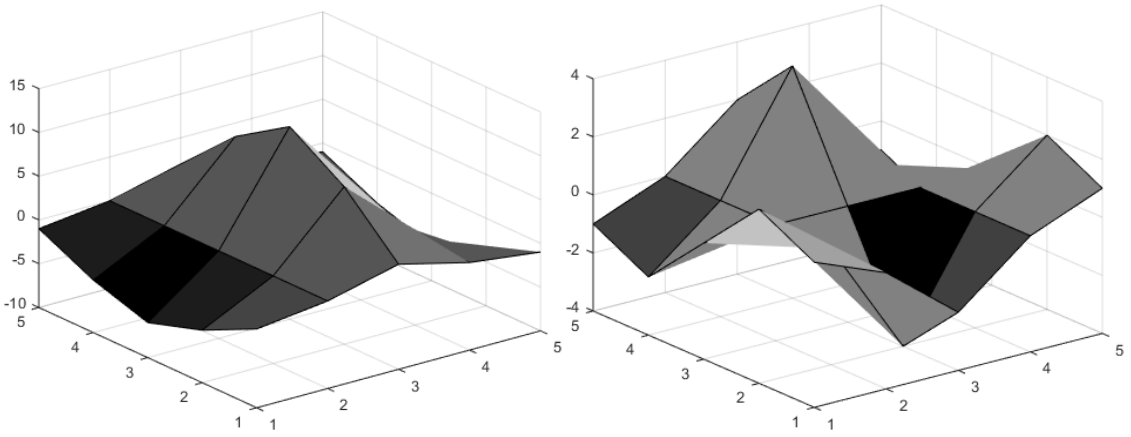
Edge: E5 = [-1 -2 0 2 1]

Spot: S5 = [-1 0 2 0 -1]

Wave: W5 = [-1 2 0 -2 1]

Ripple: R5 = [1 -4 6 -4 1]

Each pairwise combination of these five filters were used to construct the image filters which were convolved with the image region, with two examples of these two dimensional filters shown in Figure 7.2.



(a) Level-Edge (b) Edge-Surface

Figure 7.2: Two of the 2-dimensional Laws' filters: level-edge and edge-surface, which are convolved with the image region to emphasize region microstructure.

To obtain rotationally invariant filters, regions were convolved with complementary filters (e.g., Level-Edge and Edge-Level) and summed. The summed filtered images were then normalized by the Level-Level convolution of the region, resulting in a total of 14 filtered

images. From each of these filtered images, six first-order features were calculated: mean [Eq. 7.1], maximum [Eq. 7.4], minimum [Eq. 7.5], standard deviation [Eq. 7.11], energy [Eq. 7.14], and binned entropy [Eq. 7.16] resulting in a total of 84 Laws' filter features.

REFERENCES

[1] T. M. Elsheikh, S. L. Asa, J. K. Chan, R. A. DeLellis, C. S. Heffess, V. A. LiVolsi, and B. M. Wenig. Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. *Am J Clin Patho*, 130(5):736-44, 2008.

[2] R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: Images Are More Than Pictures, They Are Data. *Radiology*, 278(2):563-577, 2016.

[3] R. Rakheja, H. Chandarana, L. DeMello, K. Jackson, C. Geppert, D. Faul, C. Glielmi, and K. Friedman. Correlation Between Standard Uptake Value and Apparent Diffusion Coefficient of Neoplastic Lesions Evaluated With Whole-Body Simultaneous Hybrid PET/MRI. *Am J Roentgenol*, 201(5):1115-9, 2013.

[4] J. Ma, J. Chen, C. Zhu, J. Chen, and Y. Feng. Irradiation of the Chest Wall and Regional Nodes as an Integrated Volume with IMRT for Breast Cancer: Acute Toxicity from a Pilot Study. *Int J Rad Oncol Biol Phys*, 78(3):S250, 2010.

[5] G. S. Lodwick, T. E. Keats, and J. P. Dorst. The Coding of Roentgen Images for Computer Analysis as Applied to Lung Cancer. *Radiology*, 81(2):185-200, 1963.

[6] H. C. Becker, N. J. Nettleton, P. H. Meyers, J. W. Sweeney, and C. M. Nice. Digital computer determination of a medical diagnostic index directly from chest x-ray images. *IEEE Trans Biomed Eng*, 11:67-72, 1964.

[7] P. H. Meyers, C. M. Nice, H. C. Becker, W. J. Nettleton, J .W. Sweeney, and G. R. Mechstroth. Automated computer analysis of radiographic images. *Radiology*, 83:1029-34, 1964.

[8] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89:211-215, 1967.

[9] C. Kimme, B. J. O?Laughlin, and J. Sklansky. Automatic Detection of Suspicious Abnormalities in Breast Radiographs. *Academic Press*, New York, 1975, pp. 427-447.

[10] J. Toriwaki, Y. Suenaga, T. Negoro, and T. Fukumura. Pattern recognition of chest x-ray images. *Comput Graph Image Process*, 2:252-271, 1973.

[11] R. Kruger, W. Thompson, and A. Turner. Computer diagnosis of pneumoconiosis. *IEEE Trans Syst Man Cybern*, SMC-4:40-50, 1974.

[12] W. Spiesberger. Mammogram inspection by computer. *IEEE Trans Biome Eng*, 2:213-219, 1979.

[13] M. L. Giger, H. P. Chan, and J. Boone. Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Med Phys*, 35(12):5799-820, 2008.

[14] R. M. Haralick, S. Shanmugam, and I. Sinstein. Textural features for image classification. *IEEE Trans Syst Man Cybern*, SMC-3(6):610-621, 1973.

[15] T. Wagner. Texture Analysis. In B Jahne, H Haussecker, and P Geissler, editors, *Handbook of Computer Vision and Applications*, vol 2, chapter 12, pages 275-308. 1999.

[16] C.-C. Chen, J. S. DaPointe, and M. D. Fox. Fractal feature analysis and classification in medical imaging. *IEEE Trans Med Imaging*, 8(2):133-142, 1989.

[17] S. Peleg, J. Naor, R. Hartley, and D. Avnir. Multiple resolution texture analysis and classification. *IEEE Trans Pattern and Mach Intell*, PAMI-6(4):518-523, 2004.

[18] B. B. Mandelbrot, and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Rev*, 10(4) :422-437, 1968.

[19] W. F. Sensakovic, A. Starkey, and S. G. Armato III. Two-dimensional extrapolation methods for texture analysis on CT scans. *Med Phys*, 34(9):3465-72, 2007.

[20] R. Creutzburg, and E. Ivanov. Fast algorithm for computing fractal dimensions of image segments. In *Recent Issues in Pattern Analysis and Recognition*, pages 42-51. Springer, 1989.

[21] S. Katsuragawa, K. Doi, and H. MacMahon. Image feature analysis and computer-aided diagnosis in digital radiography: Detection and characterization of interstitial lung disease in digital chest radiographs. *Med Phys*, 15(3):311-9, 1988.

[22] W. K. Pratt. Image Feature Extraction. In *Digital Image Processing: PIKS Scientific Inside*, chapter 16, pages 535-577. John Wiley & Sons, Inc, 4th edition, 2007.

[23] K. I. Laws. Texture Image Segmentation, USCIPI Technical Report No. 940, University of Southern California, 1980.

[24] S. S. Yip, and H. J. Aerts. Applications and limitations of radiomics. *Phys Med Biol*, 61(13): R150-R166, 2016.

[25] H. J. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossman, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*, 5:4006, 2014.

[26] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, et al. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur Radiol*, 25(10):2840-50, 2015.

[27] S. G. Sapate, A. Mahajan, S. N. Talbar, N. Sable, S. Desai, M. Thakur. Radiomics based detection and characterization of suspicious lesions on full field digital mammograms. *Comp Methods Prog Biomed*, 163:1-20, 2018.

[28] D. Truhn, S. Schrading, C. Haarburger, H. Schneider, D. Merhof, and C. Kuhl. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology*, 290(2):290-297, 2019.

[29] M. Valliéres, A. Zwanenburg, B. Badic, C. C. Le Rest, D. Visvikis, and M. Hatt. Responsible Radiomics Research for Faster Clinical Translation. *J Nucl Med*, 59(2):189-193, 2018.

[30] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones, and L. Court. Measuring computed tomography scanner variability of radiomic features. *Invest Radiol*, 50(11):757-65, 2015.

[31] A. R. Cunliffe, H. A. Al-Hallaq, Z. E. Labby, C. A. Pelizzari, C. Straus, W. F. Sensakovic, M. Ludwig, S. G. Armato III. Lung texture in serial thoracic CT scans: Assessment of change introduced by image registration. *Med Phys*, 39(8):4679-4690, 2012.

[32] B. Zhao, Y. Tan, W. Y. Tsai, L. H. Schwartz, L. Lu. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. *Trans Oncol*, 7(1):88-93, 2014.

[33] C. Caramella, A. Allorant, F. Orlhac, F. Bidault, B. Asselain, S. Ammari, P. Jaranowski, A. Moussier, C. Balleyguier, N. Lassau, et al. Can we trust the calculation of texture indices of CT images? A phantom study. *Med Phys*, 45(4):1529-1536, 2018.

[34] R. Berenguer, M. D. R. Pastor-Juan, J. Canales-Vzquez, M. Castro-Garcia, M. V. Villas, F. M. Legorburo. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*, 288(2):407-15, 2018.

[35] D. Mackin, R. Ger, C. Dodge, X. Fave, P. C. Chi, L. Zhang, J. Yang, S. Bache, C. Dodge, A. K. Jones, et al. Effect of tube current on computed tomography radiomics features. *Scientific Reports*, 8(1):2354, 2018.

[36] K. R. Mendel, H. Li, L. Lan, C. M. Cahill, V. Rael, H. Abe, and M. L. Giger. Quantitative texture analysis: robustness of radiomics across two digital mammography manufacturers? systems. *J Med Imag*, 5(1):011002, 2017.

[37] J. J. Foy, K. R. Robinson, H. Li, M. L. Giger, H. A. Al-Hallaq, and S. G. Armato III. Variation in algorithm implementation across radiomics software. *J Med Imaging*, 5(4):044505, 2018.

[38] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Le Rest, and D. Visvikis. Characterization of PET/CT images using texture analysis: the past, the present? and future? *Eur J Nucl Med Mol Imaging*, 44(1):151-65, 2017.

[39] M. Sollini, L. Cozzi, L. Antunovic, A. Chiti, and M. Kirienko. PET radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. *Sci Rep*, 7(1):358, 2017.

[40] M. J. Nyflot, F. Yang, D. Byrd, S. R. Bowen, G. A. Sandison, and P. E. Kinahan. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies needs for standards. *J Med Imaging*, 2(4):041002, 2015.

[41] R. T. Leijenaar, G. Nalbantov, S. Carvalho, W. J. van Elmpt, E. G. Troost, R. Boellaard, H. J. Aerts, R. J. Gillies, and P. Lambin. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep*, 5(5):11075, 2015.

[42] A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E. G. C. Troost, C. Richter, and S. Lck. Assessing robustness of radiomic features by image perturbation. *Sci Reports*, 9(1):614, 2019.

[43] A. Zwanenburg, S. Leger, M. Valliéres, and S. Löck . Image biomarker standardisation initiative. arXiv1612.07003, 2016.

[44] A. Zwanenburg. EP-1677: multicentre initiative for standardisation of image biomarkers [abstract]. *Radiother Oncol*, 123(suppl), S914-S915, 2017.

[45] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Bolten, L. B. da Silva Santos, P. E. Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*, 3:160018, 2016.

[46] D. Mackin, X. Fave, L. Zhang, J. Yang, A. K. Jones, C. S. Ng, and L. E. Court. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS One*, 12(9):e0178524, 2017.

[47] F. Orlhac, S. Boughdad, C. Philippe, H. Salla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat. A post-reconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med*, 59(8):1321-8, 2018.

[48] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, R. Boellard, A. Dekker, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*, 48(4):441-6, 2012.

[49] Z. Obermeyer, and E. J. Emanuel. Predicting the future: big data, machine learning, and clinical medicine. *N Engl J Med*, 375:1216-9, 2016.

[50] M. G. Wallis, M. T. Walsh, and J. R. Lee. A review of false negative mammography in a symptomatic population. *Clin Radiol*, 44(1):13-5, 1991.

[51] R. E. Bird, T. W. Wallace, and B. C. Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184(3):613-7, 1992.

[52] H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Z. Wu, and H. MacMahon. Improvement in radiologists? detection of clustered microcalsifications on mammograms. The potential of computer-aided diagnosis. *Invest Radiol*, 25(10):1102-10, 1990.

[53] A. Cunliffe, S. G. Armato III, R. Castillo, N. Pham, T. Guerrero, and H. A. Al-Hallaq. Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-based Features With Radiation Therapy Dose and Radiation Pneumonitis Development. *Int J Radiat Oncol Biol Phys*, 91(5):1048-56, 2015.

[54] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp*, 2(1):36, 2018.

[55] A. Traverso, L. Wee, A. Dekker, and R. Gillies. Repeatability and Reproducibility of Radiomics Features: A Systemic Review. *Int J Radiat Oncol Biol Phys*, S0360-3016(18):1-16, 2018.

[56] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*, 49(7):1012-6, 2010.

[57] I. Shiri, A. Rahmim, P. Ghaffarian, P. Geramifar, H. Abdollahi, and A. Bitarafan-Rajabi. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-center phantom and patient studies. *Eur Radiol*, 27(11):4498-509, 2017.

[58] C. P. Loizou, M. Pantziaris, C. S. Pattichis, and I. Seimenis. Brain MR Image Normalization in Texture Analysis of Multiple Sclerosis. *J Biomed Graph Comp*, 3(1):20-34 2012.

[59] K. J. Rothman. No Adjustments Are Needed for Multiple Comparisons. *Epidemiology*, 1(1):43-46, 1990.

[60] M. Shafiq-Ul-Hassan, K. Latifi, G. Zhang, G. Ullah, R. Gillies, and E. Moros. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep*, 8:10545, 2018.

[61] S. Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536-41, 1930.

[62] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital Image Processing Using MATLAB* (2nd Ed.): Gatesmark Publishing; 2009.

[63] W. E. Johnson, and C. Li. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118-27, 2007.

[64] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat. Validation of a Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology*, 291(1):53-9, 2019.

[65] M. L. Giger, K. Doi, and H. MacMahon. Computerized detection of lung nodules in digital chest radiographs. *Proc SPIE*. 767:384-6, 1987.

[66] M. L. Giger, K. Doi, and H. MacMahon. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. *Med Phys*, 15(2):158-66, 1988.

[67] H. P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, and P. M. Jokich. Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography. *Med Phys*, 14(4):538-48, 1987.

[68] K. Doi, H.-P. Chan, M. L. Giger, inventor; University of Chicago, assignee. Method and system for enhancement and detection of abnormal anatomic regions in a digital image, US patent 4,907,156. March 6, 1990.

[69] H.-P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Oruga, Y. Z. Wu, and H. MacMahon. Improvement in radiologists? detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. *Invest Radiol*, 25(10):1102-10, 1990.

[70] K. Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput Med Imaging Graph*, 31(4-5):198-211, 2007.

[71] H. Yoshida, Y. Masutani, P. MacEneaney, D. Rubin, and A. H. Dachman. Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study. *Radiology*, 222(2):327-36, 2002.

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

[73] J. M. Wardlaw, and P. M. White. The detection and management of unruptured intracranial aneurysms. *Brain*, 123(3):205-21, 2000.

[74] J. A. Oliver, M. Budzevich, G. G. Zhang, T. J. Dilling, K. Latifi, and E. G. Moros. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. *Transl Oncol*, 8(6):524-34, 2015.

[75] L. A. Hunter, S. Krafft, F. Stingo, H. Choi, M. K. Martel, S. F. Kry, and L. E. Court. High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. *Med Phys*, 40(12):121916, 2013.

[76] N. M. Cheng, Y. H. Fang, and T. C. Yen. The promise and limits of PET texture analysis. *Ann Nucl Med*, 27(9):867-9, 2013.

[77] M. Shafiq-Ul-Hassan, G. G. Zhang, K. Latifi, G. Ullah, D. C. Hunt, Y. Balagurunathan, M. A. Abdalah, M. B. Schabath, D. G. Goldorf, D. Mackin, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*, 44(3):1050-62, 2017.

[78] A. R. Cunliffe, S. G. Armato, X. M. Fei, R. E. Tuohy, and H. A. Al-Hallaq. Lung texture in serial thoracic CT scans: Registration-based methods to compare anatomically matched regions. *Med Phys*, 40(6):061906, 2013.

[79] M. Hatt, M. Valliéres, D. Visvikis, and A. Zwanenburg. IBSI: an international community of radiomics standardization initiative [abstract]. *J Nucl Med*, 59(suppl):287, 2018.

[80] J. Kalpathy-Cramer, A. Mamomov, B. Zhao, L. Lu, D. Cherezov, S. Napel, S. Echegaray, D. Rubin, M. McNitt-Gray, P. Lo, et al. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomography*, 2(4):430-7, 2016.

[81] S. G. Armato, N. P. Gruszauskas, H. MacMahon, M. D. Torno, R. M. Engelmann, A. Starkey, C. L. Pudela, J. S. Marino, F. Santiago, P. J. Chang, et al. Research Imaging in an Academic Medical Center. *Acad Radiol*, 19(6):762-71, 2012.

[82] H. Li, M. L. Giger, O. I. Olopade, A. Margolis, L. Lan, and M. R. Chinander. Computerized Texture Analysis of Mammographic Parenchymal Patterns of Digitized Mammograms. *Acad Radiol*, 12(7):863-73, 2005.

[83] H. Li, M. L. Giger, L. Lan, J. Bancroft-Brown, A. MacMahon, M. Mussman, O. I. Olopade, and C. Sennett. Computerized Analysis of Mammographic Parenchymal Patterns on a Large Clinical Dataset of Full-Field Digital Mammograms: Robustness Study with Two High-Risk Datasets. *J Digit Imag*, 25(5):591-8, 2012.

[84] H. Li, M. L. Giger, L. Lan, J. Janardanan, and C. A. Sennett. Comparative analysis of image-based phenotypes of mammographic density and parenchymal patterns in distinguishing between BRCA1/2 cases, unilateral cancer cases, and controls. *J Med Imag*, 1(3):031009, 2014.

[85] P. Szczypinski, and M. Strzelecki. (2005) ?MaZda User?s Manual,? Lodz, Poland: The Institute of Electronics, Technical University of Lodz, Poland.

[86] M. Strzelecki, P. Szczypinski, A. Materka, and A. Klepaczko. A software tool for automatic classification and segmentation of 2D/3D medical images. *Nuc Instrum & Meth Phys Res*, 702(21):137-40, 2013.

[87] P. Szczypinski, M. Strzelecki, A. Materka, and A. Klepaczko. MaZda-A Software package for image texture analysis. *Comp Meth Prog Biomed*, 94(1):66-76, 2009.

[88] P. Szczypinski, M. Strzelecki, and A. Materka. MaZda - a Software for Texture Analysis. *Proc Of ISITC 2007*, Republic of Korea, 245-9, 2007.

[89] L. Zhang, D. V. Fried, X. J. Fave, L. A. Hunter, J. Yang, and L. E. Court. IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics. *Med Phys*, 42(3):1341-53, 2015.

[90] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Robin, S. Pieper, and H. J. W. L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104-e107, 2017.

[91] J. C. Felipe, A. J. M. Traina, and C. Traina. Retrieval by content of medical images using texture for tissue identification. *Proc 16th IEEE CBMS*, 2003.

[92] K. O. McGraw, and S. P. Wong. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychol Methods*, 1(1):30-46, 1996.

[93] T. K. Koo, and M. Y. Li. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr Med*, 15(2):155-63, 2016.

[94] R. M. Haralick, and L. G. Shapiro. In *Computer and Robot Vision: Vol. 1.* Addison-Wesley, p 460, 1992.

[95] W. B. A. Karaa, and N. Dey. Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes. *IGI Global*, 2016.

[96] G. J. Anthony, A. Cunliffe, R. Castillo, N. Pham, T. Guerrero, S. G. Armato III, and H. A. Al-Hallaq. Incorporation of pre-therapy 18F-FDG uptake data with CT texture features into a radiomics model for radiation pneumonitis diagnosis. *Med Phys*, 44(7):3686-94, 2017.

[97] U.S. Department of Health and Human Services, National Institutes of Health, and National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE) Version 4.0. URL *http://evs.nci.nih.gov/ftpl/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf*, page 174, 2009.

[98] G. C. Sharp, N. Kandasamy, H. Singh, and M. Folkert. GPU-based streaming architectures for fast cone-beam CT image reconstruction and demons deformable registration. *Phys Med Biol*, 52(19):5771-83, 2007.

[99] G. Rodrigues, M. Lock, D. D'Souza, E. Yu, and J. Van Dyk. Prediction of radiation pneumonitis by dose - volume histogram parameters in lung cancer - a systematic review. *Radiother Oncol*, 71(2):127-38, 2004.

[100] E. D. York, A. Jackson, K. E. Rosenzweig, S. A. Merrick, D. Gabrys, E. S. Venkatraman, C. M. Burman, S. A. Leibel, and C. C. Ling. Dose-volume factors contributing to the incidence of radiation pneumonitis in non-small-cell lung cancer patients treated with three-dimensional conformal radiation therapy. *Int J Radiat Oncol Biol Phys*, 54(2):329-39, 2002.

[101] H. Akaike. This Week's Citation Classic. *Current Contents Engineering, Technology, and Applied Sciences*, 12(51):42, 1981.

[102] Q. H. Vuong. Likelihood Ratio Tests for Model Selection and non-nested Hypothesis. *Econometrica*, 57(2):307-33, 1989.

[103] J. J. Foy, S. G. Armato III, and H. A. Al-Hallaq. Effects of variability in radiomics software packages on classifying patients with radiation pneumonitis. *J Med Imaging*, 7(1):014504, 2020.

[104] K. Robinson, H. Li, L. Lan, D. Schacht, and M. L. Giger. Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM. *Med Phys*, 46(5):2145-56, 2019.

[105] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger. Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. *J Med Imaging*, 7(1):012707, 2020.

[106] F. Lampariello. On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison. *J Quant Cell Science*, 39(3):179-88, 2000.

[107] P. Fan, T. Zhao, and L. Su. Deep learning the high variability and randomness inside multimode fibres. *Optics Express*, 27(15):20241-58, 2019.

[108] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks [abstract]. *CoRR*. abs/1312.6199:1-10, 2019.

[109] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the Recent Architectures of Deep Convolutional Neural Networks. Artificial Intelligence Review. *Springer Nat*, [online]:1-68, 2019. Available: https://arxiv.org/abs/1901.06032.