Computed Tomography Radiomic Features of Lung Nodules: Characterizing Feature Reproducibility Due to Variations in Image Acquisition and Reconstruction Parameters and Investigations into Mitigation Methods

Emaminejad, Nastaran

UNIVERSITY OF CALIFORNIA

Los Angeles

Computed Tomography Radiomic Features of Lung Nodules: Characterizing Feature Reproducibility Due to Variations in Image Acquisition and Reconstruction Parameters and Investigations into Mitigation Methods

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Physics and Biology in Medicine

by

Nastaran Emaminejad

2021

ABSTRACT OF THE DISSERTATION

CT Radiomic Features of Lung Nodules: Characterizing Feature Reproducibility Due to Variations in Image Acquisition and Reconstruction Parameters and Investigations into Mitigation Methods

by

Nastaran Emaminejad

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2021

Professor Michael McNitt-Gray, Co-Chair

Professor Matthew Sherman Brown, Co-Chair

Radiomic features are quantitative metrics calculated over regions of interest on medical images. Tumor-specific radiomic features can describe tumor characteristics such as shape, attenuation, and tissue heterogeneity. The promise of radiomics to link with tumor biology, treatment outcome, and pathology has been explored extensively. However, radiomics is not yet fully validated as a clinical biomarker. Two crucial steps in validation of radiomics are the assessment of its clinical utility and technical validity. Large multicenter trials are still required to ensure clinical utility of radiomics and the technical validity of radiomics has not been adequately addressed.

Radiomic features are quantitative metrics calculated over regions of interest on medical images. Tumor-specific radiomic features can describe tumor characteristics such as shape, attenuation,

and tissue heterogeneity. The promise of radiomics to link with tumor biology, treatment outcome, and pathology has been explored extensively. However, radiomics is not yet fully validated as a clinical biomarker. Two crucial steps in validating radiomics are the assessment of its clinical utility and technical validity. Large multicenter trials are still required to ensure the clinical utility of radiomics, and the technical validity of radiomics has not been adequately addressed.

Radiomics is data-driven and can get influenced by inconsistencies in image acquisition, image analysis, etc. While recent studies have demonstrated the susceptibility of radiomics to image acquisition, the reproducibility of CT radiomic features is not well established yet. Due to the unavailability of highly controlled datasets, previous efforts have been restricted to phantom data, limited patient cohorts representing narrow CT parameter ranges, or univariable analysis of a few CT parameters.

Furthermore, enforcement of harmonization strategies is needed to handle related inconsistencies. Thus far, only a few limited efforts have explored such strategies; however, harmonization of radiomics is not resolved yet, and continued research and evaluations are necessary.

This dissertation addressed the existing knowledge gap in understanding the variability of radiomic features and investigated potential strategies for harmonizing the radiomics approach. We investigated the effects of a wide range of CT acquisition and reconstruction parameters (dose, kernel, and slice thickness) on radiomic features in a realistic setting using clinical low-dose lung cancer screening cases. A computational pipeline was used that generated a unique and highly controlled dataset suitable for assessing the technical validity of radiomic features.

We performed univariable and multivariable exploration of reproducibility of well-known radiomic features. Only a few features were reproducible in response to variation of dose and kernel, and the majority of radiomic features were impacted by slice thickness. Multivariable analyses revealed interactions among CT parameters, suggesting that selecting specific combinations of CT parameters can adjust for (or worsen) the impact of CT condition variations.

We tested and compared two harmonization methods of Generative Adversarial Networks (GAN) and ComBat. A previously developed GAN model, Pix2Pix, was applied to sub-volumes surrounding lung nodules to transform lung nodule images at different CT conditions into harmonized images with radiomic features similar to a designated baseline CT condition. The ComBat method was applied separately to the radiomic feature data to estimate and adjust the deviations of radiomic features of non-baseline CT conditions to the baseline. The two mitigation techniques reduced radiomic feature variabilities at specific dose, kernel, and slice thickness ranges.

Our findings advise on the inclusion of a harmonization procedure in the radiomics approach to avoid facing technical challenges in multicenter studies. Harmonization can be achieved via careful radiomic feature selection based on reproducibility or by applying an effective mitigation technique. While further evaluation remains a future, we illustrated the possibility of alleviating some variabilities due to CT image acquisition variations. Hence, there is a potential for the inclusion of these techniques in harmonization procedures.

If validated, radiomics can be a valuable tool for clinical decision-making. Our explorations into the reproducibility and harmonization of radiomics contribute to enabling meaningful validation of radiomics.

The dissertation of Nastaran Emaminejad is Approved.

Grace Hyun Jung Kim

Denise R. Aberle

Michael McNitt-Gray, Co-Chair

Matthew Sherman Brown, Co-Chair

University of California, Los Angeles

2021

To my wonderful husband, Arsalan,

whose continued support, encouragements, and love enabled me to thrive in my doctoral program and balance my research with life.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest appreciation to my co-advisors, Drs. Michael McNitt-Gray and Matthew Brown, for their support and guidance through my dissertation research. I would not be the researcher I am today without your mentorship. I owe my success to you, and I am very grateful that you believed in me and gave me this opportunity to work with you.

Dr. McNitt-Gray patiently and convincingly showed me the light that steered my research and my education. He showed me how to learn from my mistakes and how to overcome challenges.

Dr. Brown taught me to think critically about my research. His guidance was instrumental in forming the research questions in my dissertation.

I would also like to express my gratitude toward the other members of my committee, Drs. Grace Kim and Denise Aberle. You have been my role models as successful female faculty members. Dr. Kim's feedback and assistance in our weekly meetings have always helped define a reliable experimental setup. Her role was essential in forming the statistical data analysis employed in this dissertation. Dr. Aberle has always inspired me through her insightful comments and critical questions. Her inputs and opinions helped to improve my dissertation research.

I was fortunate to have the opportunity to work with members of the Computer Vision and Imaging Biomarker (CVIB) team. I am very thankful to my colleagues Wasil Wahi-Anwar, Dr. John Hoffman, Dr. Anthony Hardy, Dr. Youngwon Choi, and Dr. Wenxi Yu, for their support. Wasil Wahi-Anwar and Dr. Hoffman developed part of the tools required for the analysis in this dissertation. Without their efforts and collaboration, it would not have been feasible to achieve the aims of this dissertation.

I would also like to acknowledge the helpful inputs that I have got from the other expert members of CVIB, Dr. Jonathan Goldin, Dr. William Hsu, and Dr. Fereidoun Abtin.

I am thankful for the contributions of Angela Sultan, Caroline Chou, Grace Lee, Dr. Anthony Hardy, Wasil Wahi-Anwar, Dr. William Hsu in image data collection.

Importantly, none of this would be possible without the support of the members of my family. My parents supported me unconditionally to continue my education. The sacrifices they have made for me are endless, and I am so fortunate to have them in my life. I am very grateful for my parents' and my sister's patience. My sister has cared for me and has been a constant source of love.

Last but not least, my husband, to whom this dissertation is dedicated, was by my side all the time through the hardships and brought light in the darkness. I am incredibly fortunate to have you in my life, and I am deeply thankful for supporting me with your love.

Vita

Education

University of Oklahoma, Norman OK

MSC in Electrical & Computer Engineering May 2015

Shiraz University, Shiraz, Iran

BSC in Electrical & Computer Engineering Feb. 2013

Research Experience

Center for Computer Vision and Imaging Biomarkers, UCLA Sep. 2015 – Present

Graduate Student Researcher

Research and Development, Genentech Aug. 2019 – Nov. 2019

Research Intern, Machine Learning in Quantitative Systems Pharmacology

Research and Development, Vertex Pharmaceuticals July 2018 – Aug. 2018

Research Intern, Machine Learning in Modeling & Informatics

Stephenson Cancer Center, OU Sep. 2013 – May 2015

Research Assistant, Computer-Aided Diagnosis Laboratory

Publications

- **N. Emaminejad**, M.W.Wahi-Anwar, G.H.J Kim, W. Hsu, M Brown, M. McNitt-Gray, "Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters", Med. Phys., March 2021. DOI: 10.1002/mp.14830
- M. McNitt-Gray, S. Napel, A. Jaggi, S.A. Mattonen, L. Hadjiiski, M. Muzi, D. Goldgof,Y. Balagurunathan, L.A. Pierce, P.E. Kinahan, E.F. Jones, A. Nguyen, A. Virkud, H.P. Chan, **N. Emaminejad,** et al., "Standardization in Quantitative Imaging: A Multicenter Comparison of Radiomic Features from Different Software Packages on Digital Reference Objects and Patient Datasets", Accepted, in press. Tomography (Special Issue on NCI's Quantitative Imaging Network).
- J. Hoffman, **N. Emaminejad**, *et al.*, "Design and Implementation of a High Throughput Pipeline for Reconstruction and Quantitative Analysis of CT Image Data", J. Medical Physics, 46(5): 2310-2322, May 2019, DOI: 10.1002/mp.13401

- **N. Emaminejad**, *et al*., "Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients," in IEEE Transactions on Biomedical Engineering, 63(5): 1034-1043, May 2016. DOI: 10.1109/TBME.2015.2477688.

Conference Presentations

- *N. Emaminejad, *et al.,* "Understanding reproducibility of radiomic features of lung nodules under heterogeneous CT acquisition and reconstruction conditions", AAPm2019, Oral Presentation.
- *Emaminejad N, *et al.,* "Robustness of Lung Nodule Classification by Radiomic Texture Feature across CT Acquisition and Reconstruction Parameters", AAPM 2018, Oral Presentation.
- *N. Emaminejad, et al. "The effects of variations in parameters and algorithm choices on calculated radiomics feature values: initial investigations and comparisons to feature variability across CT image acquisition conditions", Proc. SPIE Medical Imaging 2018: Computer-Aided Diagnosis; 105753W (2018) DOI: 10.1117/12.2293864, Poster Presentation.
- *N. Emaminejad, et al., "Assessing the Performance of CAD in Lung Nodule Detection from Low-Dose Lung Cancer Screening CT Exams Under Different Combinations of Radiation Dose Level, Slice Thickness, and Reconstruction Kernel", RSNA 2017, Oral Presentation.
- *N. Emaminejad, et al., "The effects of slice thickness and radiation dose level variations on computer-aided diagnosis (CAD) nodule detection performance in pediatric chest CT scans", Proc. SPIE, Medical Imaging 2017: Computer-Aided Diagnosis, DOI: 10.1117/12.2255000, Oral Presentation.
- *N. Emaminejad, et al., "Evaluation of CAD Nodule Detection Performance in Low Dose CT Lung Cancer Screening Across a Range of Dose Levels, Slice Thicknesses and Reconstruction Kernels", AAPM 2017, Oral Presentation.

Awards

Travel Grants:

Grace Hopper Celebration Scholarship	2020
Google Conference Scholarship	2019
Society of Women Engineers	2019
International Conference on Machine Learning (ICML)	2019
Computing Research Association for Women (CRAW)	2018
Microsoft Conference Scholarship	2018
Association of American Physicists in Medicine Expanding Horizon	2018

Chapter 1 Introduction

1.1 Radiomics

Quantitative medical imaging is a continuously expanding field in medical imaging research. Quantitative analyses on medical images are approaches that can extract information from the underlying properties of the tissue hidden from the human eye. Extraction of such information expands the utility of medical images from primarily a visualization tool to mineable quantitative data. The mineable data can be expected to provide complementary and augmented information compared to other disease-specific biomarkers, lab tests, biopsies, etc.[1]

Radiomics is an example of quantitative medical image techniques. Radiomic features are quantitative descriptors calculated over regions of interest on medical images. Tumor-specific radiomic features describe various tumor properties such as size, shape, tissue attenuation, and heterogeneity[2]. Research has demonstrated a link between radiomic features and tumor biology[3]. Several radiomic studies have shown the promise of CT radiomic features of lung tumors to provide decision support for diagnostic or prognostic tasks in lung cancer[4–6] or lung cancer screening patients [7–10]. For instance, these studies have demonstrated the potential of radiomic features to serve as a digital biomarker in phenotyping lung tumor tissue [1], describing its histopathological characteristics[11], serving as a computer-aided tool for the radiologist in cancer diagnosis as well as for oncologists[12] to predict treatment outcome and perform patient survival analysis.

1.2 Overview of the Existing Challenges in Radiomics Approach

A recent consensus statement[13] outlined the roadmap for three essential validations of technical, clinical, and cost-effectiveness for any imaging biomarker to become usable in the clinic. For radiomic features, technical validations should examine and assess whether these metrics can be measured precisely. Repeatability and reproducibility describe the precision of radiomic features. Repeatability identifies whether radiomic features of a subject are comparable when they are calculated repeatedly in short timeframes, using the same software and scan settings, etc. Reproducibility defines whether radiomic features of the same subject are comparable when they are acquired with different software, settings, etc.

On the other hand, clinical validations should evaluate whether the radiomic features are associated with biology or patient outcome. The clinical utility of radiomic features can be assessed by testing whether these metrics can improve health outcomes or provide useful diagnostic or prognostic information[13].

Despite the widespread research, the appropriateness of radiomic features to serve as clinical imaging biomarkers is still in its discovery phase[14]. Currently, there is a lack of rigorous radiomic-based multicenter clinical trials that thoroughly study the clinical utility and technical validity of radiomic features. Adoption of radiomics into trials has been slow, and this can be partly due to the uncertainties and complexities in its approach[15] and the lack of detailed guidelines or standardizations for radiomics study.

1.2.1 Complexities of the Radiomics Approach and Required Standardizations

Since radiomic features are data-driven, they are different than biospecimen-derived biomarkers. Radiomic features rely on various concepts such as image acquisition, imaging modality, image preprocessing, segmentation of the volume of interest (VOI), radiomic feature computation, software implementation of features, and statistical analysis technique. Specific guidelines and standards are vital for each of these concepts

In the context of lung nodule CT radiomic features, the Quantitative Imaging Biomarkers Alliance (QIBA)[16] has provided standardization for measuring the volume of lung nodules. The Image Biomarker Standardization Initiative (IBSI) by Zwanenburg *et al.*[17] has also provided radiomic feature nomenclature and definitions intending to enable standardization of radiomic feature computations. For example, in our previous study, we found that texture-based radiomic features showed a wide variation when different feature definition and computation approach was used[18,19]. Therefore, IBSI can be used to provide standards for the computation of radiomic features.

Nonetheless, there are no other guidelines that, for example, provide standards for scan protocols, image analysis, and segmentation software, or statistical analysis that are specific to radiomic feature calculations.

1.2.2 Translation to Multicenter Clinical Studies

To translate radiomics from the discovery phase, it is required to evaluate the clinical or biological validity of radiomic features in extensive collections of image datasets from multiple imaging centers to acquire large and diverse patient cohorts[15]. Using large and diverse datasets increases the power of the study in finding radiomic signatures or in building radiomic-based machine learning models. However, this can result in potential inconsistencies if datasets are not balanced or represent a wide range of image acquisitions or different populations and categories[15,20].

In the context of CT radiomic features, since different CT scan parameters can affect image quality differently, there is a risk that the radiomic feature quantification can differ between various datasets with varying scan protocols.

Meanwhile, radiomics is based on the hypothesis that the quantitative features extracted from images of tissues capture pathophysiological information, and therefore they can inform tumor phenotype or protein signatures[1,3]. However, if radiomic features are influenced by non-biological processes such as the variation of image quality due to variation of image acquisition protocols, their ability to provide accurate information about tumor phenotype may be negatively impacted, and radiomic features become unreliable.

1.2.3 Non-Biological Factors Causing Inconsistency in Radiomics approach

As was previously mentioned, the extraction of radiomic features and their analyses depends on various concepts. This increases the number of potential sources of variability. Other than data analysis, software, feature computation approaches, ROI segmentation algorithms, a predominant source of variability can be the image acquisition and the vendor-specific characteristics of the imaging modality.

In the context of CT radiomic features, dominant factors that can impact image quality or cause inconsistencies between images acquired at different scanners or different clinical sites include but are not limited to radiation dose level (which depends on several factors including the tube current) field of view, pitch, reconstruction algorithm (filtered back-projection versus iterative), reconstruction kernel, slice thickness, scanner calibration and scanner vendor.

Changes in some factors mentioned above can have a more noticeable impact on image quality than others and specific to this research, have an impact on the quantitative radiomic features calculated. For example, variation of radiation dose level results in variation Poisson noise due to photon counting statistics. Variation of reconstruction kernel determines the spatial frequency or the sharpness/smoothness of the image as well as image noise level. Sharp kernel settings result in higher contrast resolution and higher noise content, whereas a smooth reconstruction setting results in less noise and lower resolution. Different choices of slice thickness change the representation of anatomy in each slice, and the amount of volume averaged in each slice. This results in the variation of image noise; thick slice images have less noise and have more volume averaging than thin slices.

First-order radiomic features, such as mean, standard deviation, etc., describe the characteristics of the intensity histogram of the region of interest (ROI). Second-order radiomic features measure spatial heterogeneity of gray levels and the relationship of neighboring gray levels[21]. The variation of image quality, in terms of noise or spatial frequency, may impact the quantitative measurements acquired by the first-order or second-order features. Additionally, radiomic features that describe the size or shape of ROI may also be affected if the ROI segmentation algorithm is influenced by variation of CT scan parameters. For example, in a previous study[22], it was shown that dose reduction and slice thickness changes varied the automatic segmentation of lung nodules. Another prior study showed that dose reduction by 50% in chest CT scans resulted in a slight decrease in true positive detections by a computer-aided detection (CAD) tool[23].

Some radiomic features that, for example, measure noise in an image, such as standard deviation, clearly will be impacted by the variation of the three CT parameters of dose, kernel, and slice thickness. Yet, the dependence of several other radiomic features of patient tumors, especially the texture-based radiomic features, to the variation of CT parameters, such as dose, kernel, and slice thickness, is not well understood.

1.2.4 Potential Radiomic Feature Inconsistencies Can Affect Its Reliability

As was previously mentioned, there has not been a standardized approach for radiomics in its discovery phase so far. For example, in routine clinical CT imaging, a wide variation of acquisition parameters and reconstruction parameters are being used; data collection for radiomic studies has not followed a predefined guideline to harmonize protocols between different clinical or research institutes. If radiomic features vary in response to variation of non-biological factors related to scan protocols, the findings of the discoveries that have assessed the clinical utility of radiomic features face a reliability concern.

For instance, consider a high-performing machine learning model that has been built using radiomic features to predict patient outcomes. Suppose the training dataset consists of images acquired with a consistent CT image acquisition and reconstruction protocols with no variation in any CT parameter. In that case, the model is expected to make reliable and accurate predictions for new image data acquired with the same set of CT parameters. However, if the radiomic features (i.e., the predictors of the model) vary in response to changes of CT parameters, the model may not generalize well to image data acquired with different CT parameters.

Similarly, another concerning scenario can happen when a predictive model has been built using radiomic features from training data with a heterogeneous CT protocol set. One example is when the model building procedure does not involve any steps that account for disparities between image protocols (such as harmonization and ensuring statistically representativity of data at different protocols, etc.). Another example is when the inter-protocol variation of radiomic features is more considerable than inter-patient radiomic feature variations. In these examples, the noise and uncertainty in the data may make it difficult to detect the important signal from actual biological variations (causing false negatives). In addition, the model’s performance may also become biased toward non-biological factors such that unimportant or irrelevant factors be taken as predictive factors for the model (causing false positives).

An example of a situation where a false positive discovery may happen is considering a multiple-time point study with baseline and follow-up images acquired at inconsistent protocols. If radiomic features are affected by such inconsistency, the actual biological variability may not be distinguished from the variability induced by acquisition variation.

1.2.5 Summary of Existing Challenges for Radiomics

In summary, the radiomics approach is complex, and there are several potential sources of variability that may result in inconsistency of radiomic feature measurements. In the context of CT radiomic features, different choices of CT parameters, such as dose, kernel, and slice thickness, may impact radiomic feature values.

The lack of technical validations and the lack of standardized scan protocols in retrospective or prospective studies may challenge the results of radiomic discoveries. Additionally, the generalizability of radiomic signatures to image datasets from different clinical centers faces uncertainty.

As a result, to translate radiomic features to clinic, single-center or multicenter technical validation of radiomics is a crucial step to ensure its reliability. The impact of several non-biological factors on radiomic features is not well understood yet. Therefore, technical validity shall be performed with the purpose to provide an understanding of precision of radiomic features, define the limitation of the approach, and identify strategies to reduce or handle corresponding risks and uncertainties.

1.3 Overview of the Literature: Validation of Robustness of CT Radiomic Features

Currently, there is a lack of sufficient knowledge regarding the impact of CT acquisition and reconstruction parameters on radiomic features of lung nodules in clinical patient datasets.

The current body of knowledge on the sources of uncertainty in CT radiomic features of lung nodules is limited to exploring radiomic robustness in phantom images or a few patient studies.

For example, a radiomics phantom, the Credence Cartridge Radiomics (CCR) phantom created by Mackin *et al.*[24], with two cartridges that they claim match the texture and intensity characteristics of lung tumors, has been used in various studies to investigate the robustness of radiomic features [25,26]. Shafiq-Ul-Hassan *et al.*[27] and Kim *et al.*[28], in their research on radiomic features of the CCR phantom, showed a significant impact of reconstruction kernel on most first-order and Gray Level Co-occurrence Matrix (GLCM) features. Other studies have used an anthropomorphic thoracic phantom[29] with vasculature and synthetic nodule inserts to investigate the robustness of radiomic features. These works have reported notable variation of radiomic features due to variation of slice thickness and reconstruction algorithm.[30,31]

Even though phantom studies provide a basic understanding of the impact of acquisition protocols on image quality in general, the impacts on radiomic features differ compared to patient datasets with nodules.[32] Phantom images do not provide a perfect representation of the complex shape and heterogenous composition patterns of lung tumors.

Furthermore, while the previously mentioned studies in the literature offer valuable insight, each work is focused on only one type of phantom data. Thus, the other limitation is that the disparities between the characteristics of the different types of phantoms with each other or with patient datasets and the potential impact of these disparities on generalization of findings to patient datasets is not assessed.

As a result, there is still a need for in-depth investigation of sources of variability in patient cohorts. However, due to the difficulty of acquiring patient images at a large variety of reconstruction or acquisition settings, only a few studies investigate the robustness of radiomic features in patient datasets. Some studies have focused on factors unrelated to CT acquisition parameters, such as the role of segmentation algorithms[33] or the impact of manual vs. automated tumor segmentation,[34] the variations between different radiomic software packages[19,35]. The findings of these studies indicate substantial variability of radiomic features and the need for dedicated standardization in using radiomics software and segmentation algorithms.

Balagurunathan *et al.*[36] assessed variability of radiomic features between repetitions of scans of thirty-two 32 NSCLC patients with the same scanner and CT protocol. This study found that only 30% of the radiomic features were repeatable. On the other hand, Hunter *et al.*[37] found that the robustness of a large number of radiomic features of test-retest CT scans of fifty-six NSCLC patients depended on the scanner. These findings demonstrate the need for studies that further assess the precision of radiomic features. Additionally, these findings suggest the necessity of incorporating an additional radiomic feature selection step that filters non-reproducible features that can be influenced by variation of CT parameters.

Some other studies have focused on the reproducibility of radiomic features of anatomical parts of the body other than the lung; for example, Midya *et al.*,[38] along with their phantom study, analyzed the effect of the variation of reconstruction kernel on radiomic features of the liver lesions in one human abdominal CT. Additionally, Meyer *et al.*[39] investigated the impact of reconstruction settings along with a variation of dose level in a patient cohort with liver lesions. The authors simulated different dose levels by adding noise to the raw data. Though, in this study, impact of

CT parameters, such as dose, was investigated in a univariate fashion and the interactions among different CT parameters were not described.

Currently, the patient studies of lung nodules are mostly limited to exploring the impact of only one or two parameters (generally reconstruction algorithm or slice thickness) on radiomic features of lung nodules. Little has been done to include the varying dose (or other parameters) in addition to these, as it has not been feasible for investigators to obtain multiple CT images of patients at various dose levels. Kim *et al.*[28] studied inter-reconstruction algorithm (FBP B50f kernel vs. iterative with strengths 3 and 5) along with intra- and inter-reader variability of 15 radiomic features of 42 pulmonary tumors. Zhao *et al.*[40] studied the impact of CT reconstruction algorithm (sharp, smooth) and slice thickness (1.25 mm, 2 mm, and 5 mm) on 89 radiomic features of 32 lung cancer patients. Both these studies reported significant differences induced by variation of the reconstruction algorithm. Among all lung nodule studies, to the best of our knowledge, only one study by Fave *et al.*[41] explored the impact of kV and mAs variation (applied on image data and not on the raw data) on radiomic features in patient datasets. However, this study was performed using cone-beam CT images, which is a different modality than helical CT imaging. Hence, the impact of dose variation on CT radiomic features of patient lung nodules has remained unaddressed to this end.

As a summary, while there are deviations between the robustness of radiomic features in different type of phantoms and patient datasets, there are no comparisons that thoroughly compares variability of radiomic features of different types of phantoms with radiomic features of patient datasets. On the other hand, only a few studies have analyzed the reproducibility of radiomic features in patient datasets.

Furthermore, the scope of available patient studies is limited in terms of the data points investigated and the range of parameters under investigation. The majority of available studies have performed univariable analyses that assess parameter impacts individually and one at a time without considering potential interactions between CT parameters. This demonstrates the need for further studies that examine the effect of multiple parameters simultaneously in radiomic features reproducibility within clinical datasets. Therefore, the technical validity of radiomic is still an open question in the radiomics roadmap.

1.4 Specific Aims

Motivated by the facts and limitations discussed above, the objective of this dissertation was to address the concern in the reliability of tumor-specific radiomic features by studying CT radiomic features of lung nodules and addressing the existing knowledge gap about the reproducibility of radiomic features. In this dissertation, the reproducibility of radiomic features was assessed by evaluating the effect of non-biological factors such as CT image acquisition and reconstruction settings on radiomic feature values calculated from lung nodules of a cohort of lung cancer screening patients. We also investigated the effects of CT image acquisition and reconstruction on a series of phantoms with different levels of complexity to better understand these effects in objects of known size, shape and composition.

Eventually, mitigation techniques were investigated and evaluated to test whether we can overcome the variability of radiomic features and ensure reproducibility of radiomic features.

This dissertation takes steps toward technical validation of radiomic features to serve as an image biomarker.

Firstly, the work contributes to understanding the reliability of radiomic features with regard to changes in CT technical parameter settings. We investigated combinations of wide variation of scan settings of dose, slice thickness, or reconstruction kernel via both univariable and multivariable assessments. Hence, we conducted a multi-faceted and simultaneous analysis of these CT parameters and captured their interactions in affecting radiomic feature values. Moreover, through this work, experimental methodologies were developed to evaluate the robustness of radiomic features properly. These methods will constitute a widely applicable platform for assessing the impact of CT image settings on radiomic features.

Secondly, we addressed the correctability of radiomic feature inconsistencies among CT images with different scan protocols by testing mitigation approaches. Therefore, our findings can assist in establishing reliable frameworks for conducting future multicenter radiomics trials acquired with heterogenous CT scan protocols.

This study's rationale is that understanding the reproducibility of radiomic features sheds light on possible non-biological factors that negatively impact radiomics outcome and intratumor measurements. Additionally, the findings of this work bring awareness on the importance of setting proper inclusion criteria for image settings in radiomics workflow and enable identification of mitigation techniques to ensure the robustness of radiomics. After addressing such challenges through rigorous testing, we can gain confidence in radiomic biomarkers to be implemented in the clinic for medical decision-making.

The objectives of this study were attained through the following specific aims (shown in *Figure 1-1* and Figure 1-2):

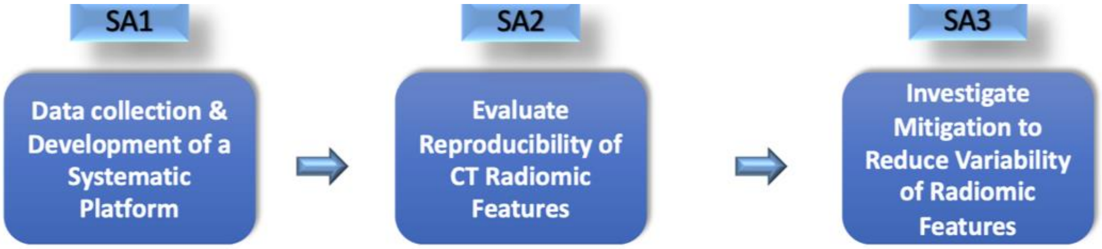

SA1
Data collection & Development of a Systematic Platform
SA2
Evaluate Reproducibility of CT Radiomic Features
SA3
Investigate Mitigation to Reduce Variability of Radiomic Features

Figure 1-1. An overview of the specific aims (SA) of this dissertation

1.4.1 Specific Aim 1 (SA1) Data collection, design, and development of a systematic framework for reproducibility analysis of radiomic features

This aim involved two main steps of data collection and development of an infrastructure to handle CT image data preparation and analyses to achieve the goals of this dissertation.

To assess the reproducibility of radiomic features in response to variation of CT imaging parameters, we collected CT scans of a cohort of lung cancer screening patients. We then generated an image dataset that consisted of multiple image datasets from the same subject but at a variety of clinically applicable CT image protocols. These protocols included a wide range of dose levels, reconstruction algorithm, slice thickness, and kernel settings. CT image generation and lung nodule segmentation were performed via the development of a reconstruction and simulation pipeline.

Subsequently, a framework was designed and developed to handle the required data analysis such as data cleaning, radiomic feature calculation, agreement analysis, statistical testing, data visualization, etc. The image simulation pipeline and feature calculation framework form a unique modular platform; this platform facilitates future experiments and various parallel large-scale statistical data analyses to evaluate the robustness of quantitative imaging metrics (e.g., radiomic features) across a wide range of technical factors (e.g., CT parameters).

1.4.2 Specific Aim 2 (SA2): Investigate the reproducibility of radiomic feature values in response to variation of CT image acquisition and reconstruction parameters

In this aim, the goal was to determine whether tumor-specific radiomic feature values vary or how they vary in response to changes in conditions unrelated to the underlying biological characteristic of tissue. To achieve this goal, we investigated the reproducibility of texture-based and intensity-based CT radiomic features in response to variation of technical conditions inherent to CT image acquisition and reconstruction parameters. The investigated CT parameters were radiation dose, reconstruction algorithm, kernel, and slice thickness.

We first evaluated reproducibility by measuring the agreement between radiomic feature values of lung nodules when they are calculated from the same subjects but on different images representing different CT conditions within a broad and clinically applicable range. Radiomic features that had high inter-condition agreement were considered as reproducible. We also explored the potential interactions between the effect of CT parameters on radiomic feature values by conducting multivariable analyses.

Furthermore, we examined the radiomic features of three different phantom datasets to understand whether the variability of phantom radiomic features and the similarities (or distinctions) between different phantoms or patient datasets can provide further insights into how the CT parameters impact radiomic features.

1.4.3 Specific Aim 3 (SA3) Investigate mitigation strategies in reducing the variability of radiomic features

In this aim, we evaluated the potential of strategies in mitigating the variability of radiomic features induced by CT image acquisition and reconstruction conditions. The idea is that the techniques that either apply image processing and transformation (e.g., via Generative Adversarial Networks) or adjust radiomic features can standardize radiomic features from different CT conditions to a common baseline.

To evaluate whether these techniques can harmonize radiomic features to a baseline condition, we will assess and compare the agreement of radiomic features obtained from baseline scan settings to those obtained from scan settings other than baseline before and after applying the mitigation techniques. Techniques that increase the inter-condition agreement of radiomic features are capable of improving the reproducibility of radiomic features. Therefore, such mitigation techniques are deemed to have a potential role in building reliable frameworks and harmonizing radiomic feature inconsistencies in future multicenter radiomic studies.

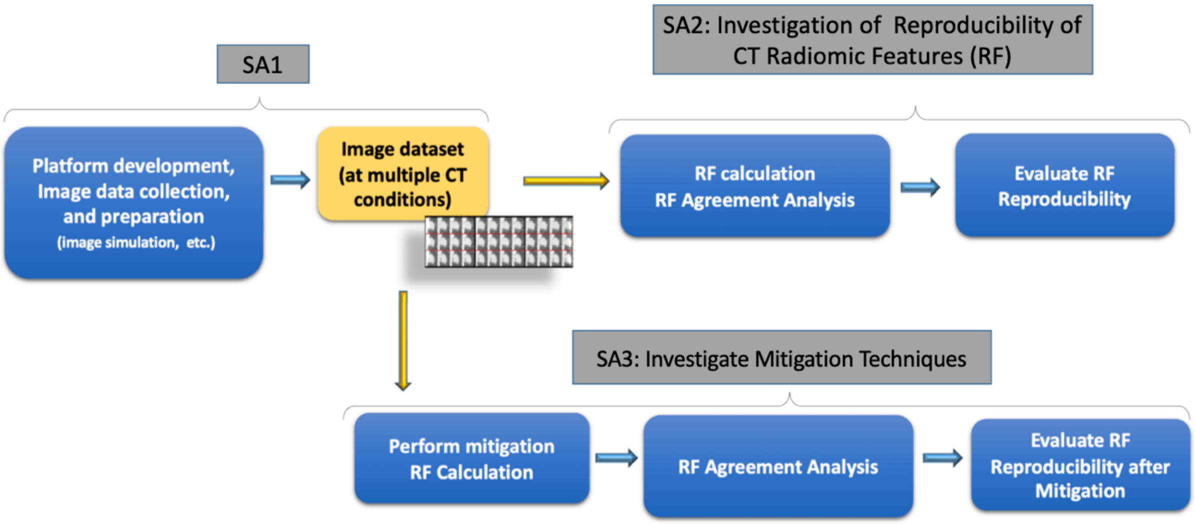

SA1
Platform development, Image data collection, and preparation (image simulation, etc.)
Image dataset (at multiple CT conditions)
SA2: Investigation of Reproducibility of CT Radiomic Features (RF)
RF calculation RF Agreement Analysis
Evaluate RF Reproducibility
SA3: Investigate Mitigation Techniques
Perform mitigation RF Calculation
RF Agreement Analysis
Evaluate RF Reproducibility after Mitigation

Figure 1-2. Summary of the studies in each Specific Aim of this dissertation

1.5 Dissertation Organization

This Dissertation is divided into two sections. Section 1 is dedicated to the description of the work done to fulfill Specific Aims 1 and 2. Section 2 is dedicated to the studies performed to achieve Specific Aim 3.

1.5.1 Section I of the dissertation

Chapter 2 provides information about the development of the infrastructure and the computational pipeline that enables systematic analyses as part of Specific Aim 1 of the dissertation.

Chapter 3 describes the collected patient cohorts utilized in this work as part of Specific Aim 1.

Chapter 4 discusses the investigations into understanding reproducibility of radiomic features and the impact of CT acquisition and reconstruction parameters according to Specific Aim 2.

Chapter 5 presents our study in assessing and comparing the variability of radiomic features of three different phantom datasets.

1.5.2 Section II of the dissertation

An introduction in this section will provide background about the mitigation techniques used in this study and reviews the available studies in the literature that have investigated harmonization of variability of radiomic features.

Chapter 6 discusses the investigations into application of Generative Adversarial Networks (GAN) in harmonization of radiomic features against variation of CT acquisition and reconstruction parameters.

Chapter 7 presents application of another method, ComBat, for harmonization of variability of radiomic features.

Chapter 8 examines and compares the mitigation performance of the two investigated techniques of ComBat and GAN.

Chapter 9 summarizes the work presented in this dissertation and presents discussion, conclusion, and potential future directions.

Section I

Chapter 2 Development of a Platform for Image Data Analysis[1]

2.1 Introduction

In this chapter, we will describe the computational platform developed in this dissertation as part of Specific Aim 1 that allows us to perform systematic analyses described below. This platform complements two existing in-house image simulation and nodule segmentation modules. The existing in-house modules with the developed modules in this dissertation form a high-throughput computational pipeline[42] that enables creating and analyzing a large, high-quality dataset of clinical CT images. The key characteristic of this pipeline is generating highly-controlled CT image data at a wide variety of image acquisition and reconstruction parameters and providing automation in the analysis of large datasets. CT image data generation is achieved via raw data simulation and reconstruction. The automation involves organizing a large volume of data and quantitative image analysis such as nodule detection, radiomic feature calculation, etc.

Currently, there exists no such framework that accomplishes these tasks together. On the other hand, large clinical datasets are required for assessing the clinical relevance and generalizability of quantitative medical imaging approaches[13]. Examples of such approaches are computer vision and image processing techniques that perform segmentation[43], quantitative imaging biomarkers[4,5,

[1] The content of this chapter is based on: Hoffman J, Emaminejad N, Wahi-Anwar M, et al. Design and Implementation of a High Throughput Pipeline for Reconstruction and Quantitative Analysis of CT Image Data. *Med Phys*. Published online 2019. doi:10.1002/mp.13401

or machine learning models that aid in disease diagnosis[45]. Additionally, a lack of automation can limit and challenge large-scale quantitative medical imaging studies. Therefore, developing a systematic framework, such as the one developed in this work, is essential to allow the establishment of large-scale quantitative medical imaging studies with less hurdle.

While the pipeline generates large-scale image datasets, it minimizes the time required for generating such image data. Furthermore, the pipeline enables future quantitative experiments (e.g., statistical data analysis) through its modular framework. A potential application for this pipeline is assessing the technical validity (i.e., robustness) of quantitative imaging biomarkers (i.e., radiomic features), which is the focus of this dissertation.

2.2 Reconstruction and Data Analysis Pipeline Design

The pipeline comprises two main blocks: 1) raw data simulation and reconstruction, 2) image data analysis. Figure 2-1 shows the high-level overview of the pipeline. In the following subsections, we will describe each component. In Figure 2-1, all the modules within the 'Image Data Analysis component' except for the 'Detection & Segmentation' module were developed in this dissertation which will be described in sections 2.2.2.2 and 2.2.2.3.

2.2.1 Raw data Simulation and Reconstruction

2.2.1.1 Raw Data Simulation

The pipeline can receive raw projection data as input, and it can simulate reduced-dose raw data. The dose reduction simulation is done by leveraging a realistic noise model[33] to add a calibrated amount of noise to the projection data; this approach was previously described and used in previous studies[46,47]. Young *et al.*[46,47] developed the current low dose simulation tool that applies the methods described by Zabic *et al.*,[48] on our own multidetector- row CT scanner equipped with tube current modulation (TCM). First, the photon fluence and bowtie filter shape of this scanner were estimated by acquiring air scans. Then, calibrated levels of noise (sampled from altered Poisson distribution) were added to the original raw projection data to simulate specific amounts of dose reduction. Because all of our scans used TCM and the quality reference mAs was the same for all patients, the dose reduction was expressed as a percentage dose reduction from the original. In the implementation, this reduction is modeled as linear scaling of the TCM function (which is recorded in the raw projection data of the scanner) to achieve the desired dose level for each patient scan. Young et al.[46] validated the low dose simulation tool by comparing simulations with anthropomorphic chest/lung phantom scans both qualitatively and quantitatively (via mean and standard deviation of Hounsfield- unit values).

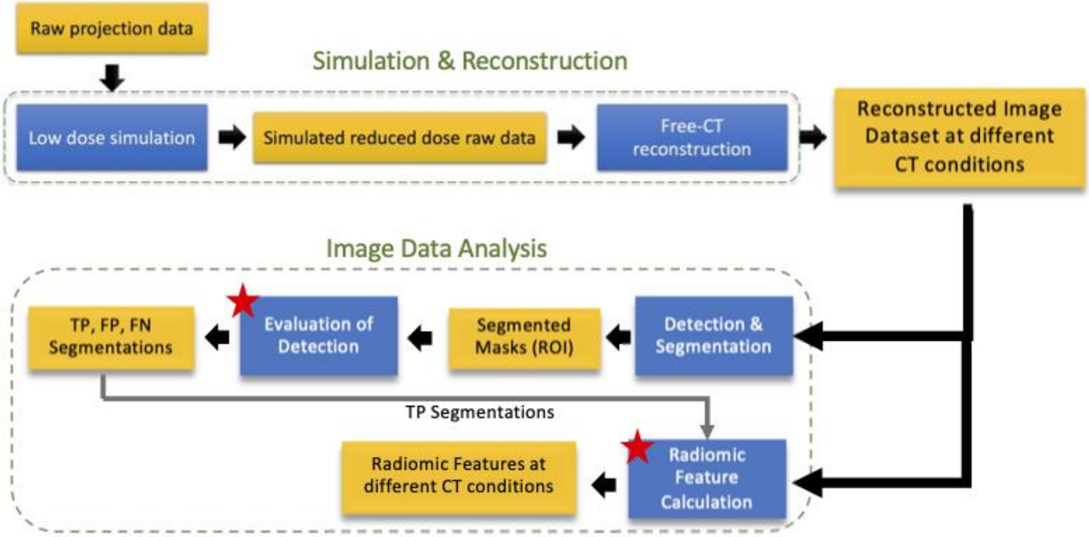


Figure 2-1. The two main sections of the pipeline: 1) simulation & reconstruction, 2) image data analysis. The different modules in each section are shown in blue. The red star shows the modules that were developed in this dissertation. The radiomic feature calculation module is a wrapper that adopts a previously available software in the pipeline platform. TP: true positive, FP: false positive, FN: false negative detections.

2.2.1.2 Image Reconstruction

In the image reconstruction block of the pipeline, a Free- CT tool, developed by Hoffman *et al.*[49], performs reconstruction of raw projection data using weighted Filtered Back Projected (wFBP) algorithm for helical multi-detector CT. The Free_CT tool employs an implementation of the wFBP algorithm described in the study by Stierstorfer *et al.*[50]. Furthermore, reconstructions are carried out through a GPU queuing framework. The details of GPU usage optimization and implementation are provided in the documentation for the software[51].

The Free_CT tool enables image reconstruction at three different kernel settings of smooth, medium, and sharp that resemble Siemens B20, Siemens B45, and Siemens B70. Boedeker *et al.*[52] plotted the modulation transfer function (MTF) for Siemens wFBP reconstruction kernels in the range of B10–B80 in their Figures 2 and 3 that presents how the contrast changes at different spatial frequencies as a result of the application of various kernels. Hoffman *et al.*[53] have also plotted the three free- CT kernels' profiles in the current study in their Figure 1.

The reconstruction module also allows for reconstructing raw data at three different slice thicknesses of 2mm, 1mm, and 0.6mm.

Figure 2-2 shows an example of a nodule region with Free_CT reconstructions at the three different kernels, three slice thicknesses, and four different dose levels of 100%, 50%, 25%, 10% of the original dose level. The nodule shown in Figure 2-2 was initially scanned with a low-dose screening protocol ($CTDI_{vol}$ ≅ 2mGy). Therefore, for example, the 10% dose level represents scans with $CTDI_{vol}$ ≅ 0.2mGy. The different combinations of CT parameters resulted in 36 unique CT conditions, as shown in Figure 2-2. This figure demonstrates how the appearance of the nodule tissue and the noise change as CT parameters change. Figure 2-3 and Figure 2-4 show an axial slice of a subject at images with different kernels, slice thicknesses, and dose levels.

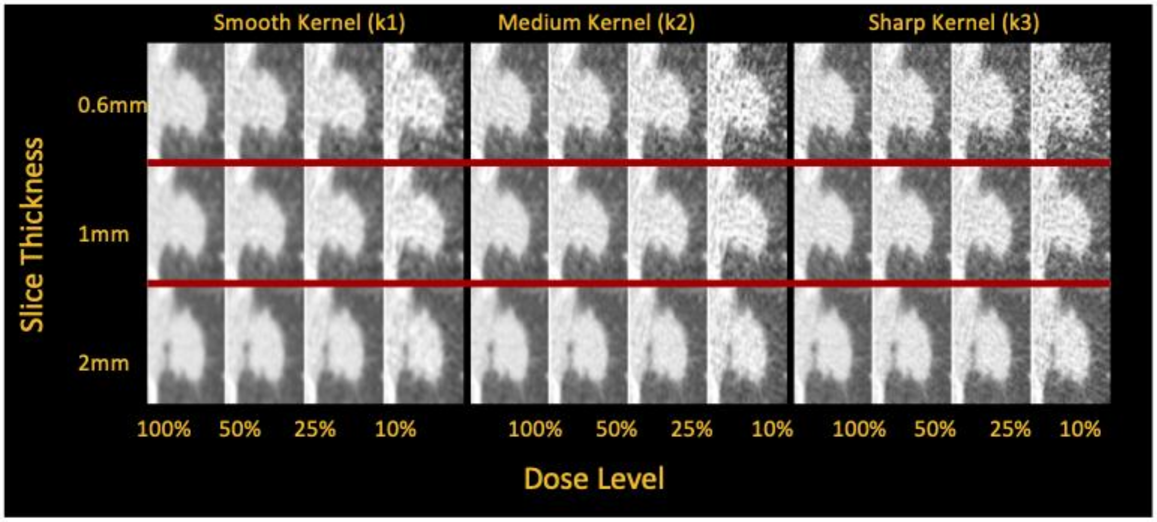


Figure 2-2. Example of a nodule at 36 different CT conditions with three different kernels, four different dose levels, and three different slice thicknesses

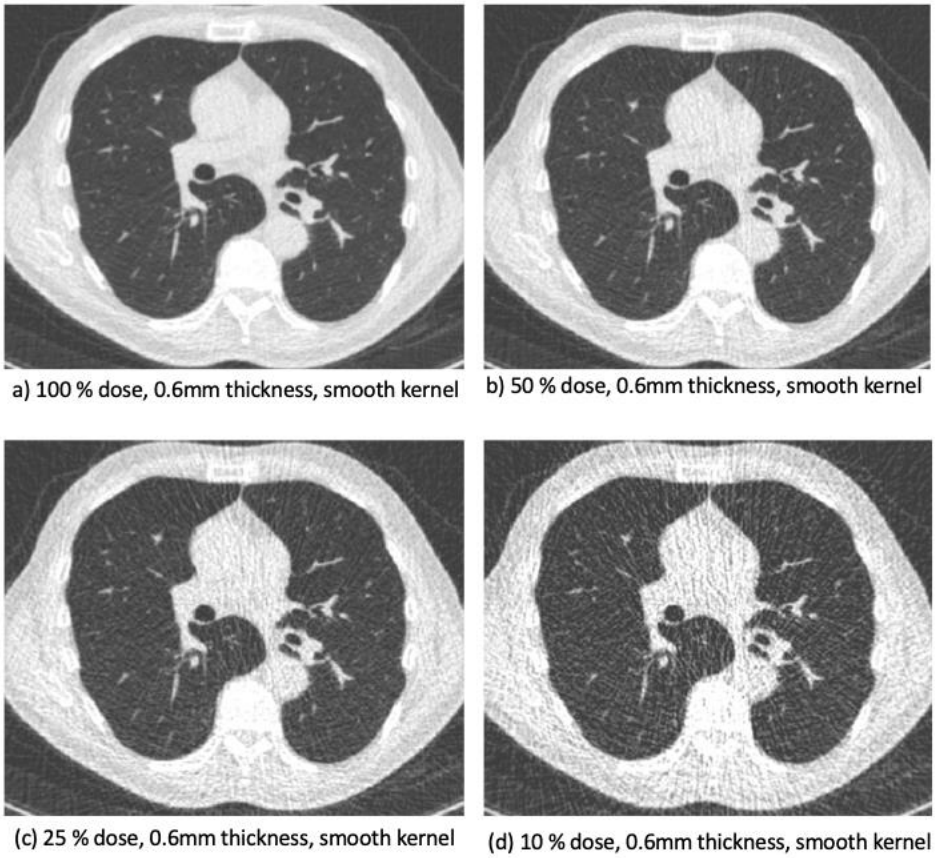
2.2.2 Image Data Analysis

The image data analysis block of the pipeline currently consists of a series of modules that apply computer vision tasks (e.g., image processing, denoising, nodule segmentation, etc.) or perform quantitative data analysis (e.g., radiomic feature calculation). However, these modules can be replaced with or followed by other modules to carry out various tasks (e.g., lung segmentation, etc.) and analysis (e.g., statistical analysis, etc.).

2.2.2.1 Detection & Segmentation Module

The nodule detection module was built from an in- house Computer- Aided Detection (CAD) tool[43] that performs automatic lung nodule detection and segmentation.

Before nodule segmentation, the anonymized reconstructed images are translated into an internal image format (.hr2) readable by the CAD software. The reformatted images are the input to the CAD tool. The output contains a list of segmented regions of interest (ROI) for some anatomical parts (such as lung, trachea, etc.) and nodule candidates. The segmented ROIs are provided in form of a binary mask that have pixel values of one in the pixels that belong to the segmented object in the input image. The CAD software runs in parallel on either of the operating systems of Windows or Linux.



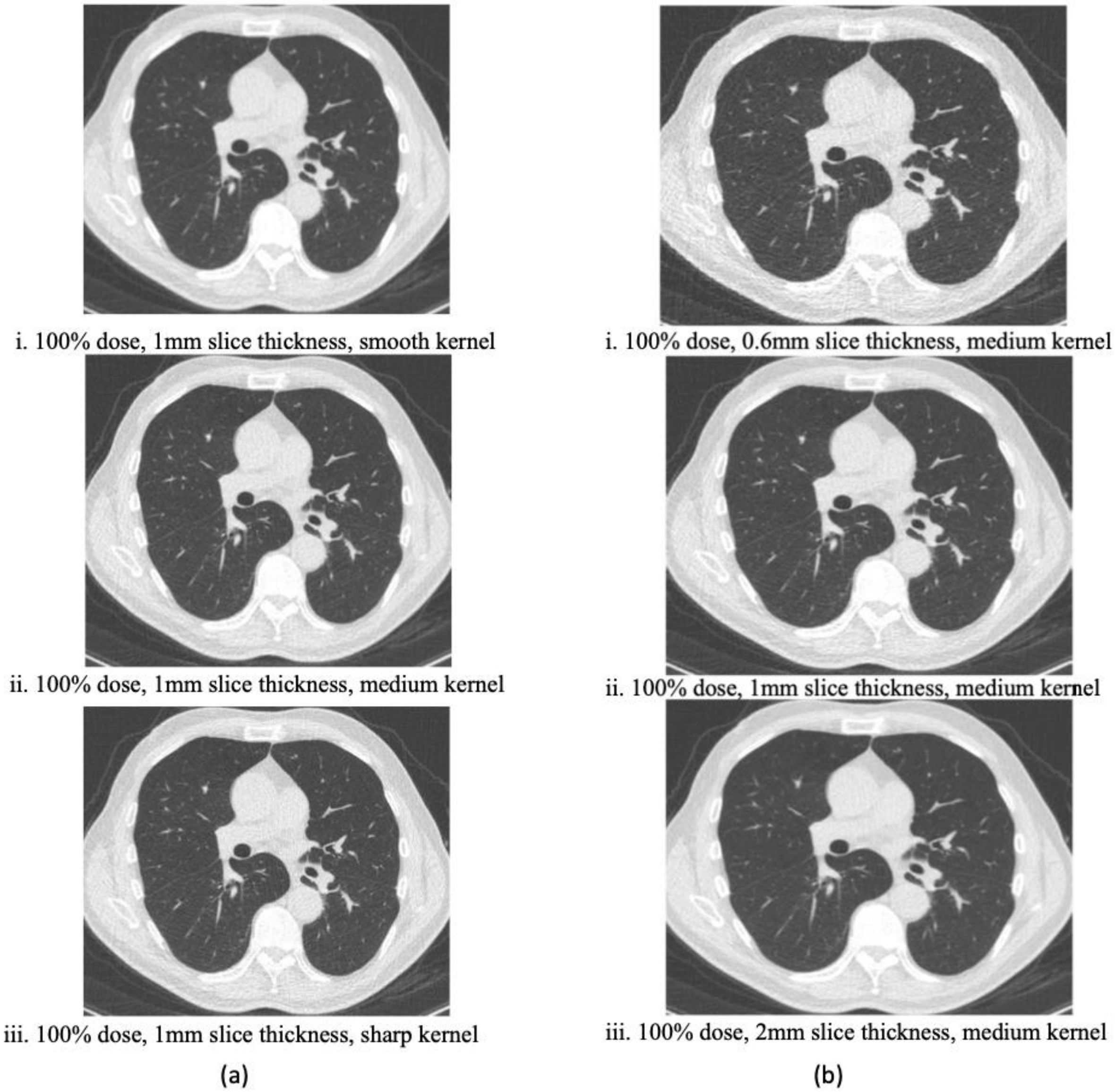
a) 100 % dose, 0.6mm thickness, smooth kernel

b) 50 % dose, 0.6mm thickness, smooth kernel

(c) 25 % dose, 0.6mm thickness, smooth kernel

(d) 10 % dose, 0.6mm thickness, smooth kernel

Figure 2-3. Lung CT images at 0.6mm slice thickness, smooth kernel and four different dose levels: a) 100%, b) 50%, c) 25%, d) 10%.



i. 100% dose, 1mm slice thickness, smooth kernel

ii. 100% dose, 1mm slice thickness, medium kernel

iii. 100% dose, 1mm slice thickness, sharp kernel

i. 100% dose, 0.6mm slice thickness, medium kernel

ii. 100% dose, 1mm slice thickness, medium kernel

iii. 100% dose, 2mm slice thickness, medium kernel

Figure 2-4. a) Lung CT images at 100% dose, 1mm slice thickness and three different kernels: i) smooth (k1), ii) medium (k2), iii) sharp (k3). b) Lung CT images at 100% dose, medium kernel and three different slice thicknesses: i) 0.6mm, ii) 1mm, iii) 2mm

2.2.2.2 Module for Evaluation of Nodule Detection

The evaluation module developed in this dissertation consists of two primary steps. The module is firstly responsible for gathering and managing lung nodule annotations (i.e., information regarding location, size, and center of nodules). Secondly, the module is responsible for evaluating the CAD detection results and identifying true positive (TP), false positive (FP), and false-negative (FN) detections.

The evaluation module consists of compiled programs and scripts in Python language. This module and the detection module can run independently from the simulation and reconstruction block of the pipeline. Therefore, they are readily available to run on image data other than the pipeline's reconstructions (e.g., images directly from a scanner acquired with various CT parameters). Furthermore, the evaluation module, similar to the detection module, can be utilized in Windows and Linux operating systems.

A database of nodule annotations has been acquired and archived in advance. To create this database, trained lab technicians imported annotations of each subject into the Quantitative Imaging Workstation (QIWS) according to clinical reports from the radiologist's interpretations. The annotations include information about anonymized subject ID, the number of reported nodules by the radiologist, location of two perpendicular and axial diameters of the nodule, and the lung-RADS[54] category (if available) or the nodule composition. The procedure in which the technicians acquire images and import this information into QIWS does not depend on the reconstruction pipeline or the reconstructed images. The technicians annotate the images by reading the original image data reconstructed by the CT scanner. Therefore, the reader markings are based on the original scans and not based on the pipeline reconstructed images.

Since in this pipeline, for each subject, we have image data at several CT conditions (e.g., 36 conditions), each image will be assigned with an anonymized ID that determines the subject identity and the image condition. Furthermore, we have created a directory convention so that image data and annotations, segmentation results, etc. be saved in directories that can be easily retrieved. Each unique image of a subject acquired at a unique combination of CT parameters (dose, kernel, slice thickness) has its corresponding folder.

In the first step, we use the evaluation module to run a script that exports the anonymized annotation data from the internal QIWS database. The script can either export the information for one specific case or a set of cases with their IDs provided in a spreadsheet. The annotation data will be written in YAML (http://yaml.org) files, a "data serialization" format for reading and writing data and can be used by Python, in the corresponding directory for each case.

The second step in this module is performed after annotations are exported. This step can operate under two modes, single-case mode or multiple-cases modes. A specific configuration file (in YAML format) must be submitted to launch the evaluation step and run a series of scripts and programs for each operation mode. The Python scripts are shared between the two modes except that it is necessary to run an additional script for the multiple-cases mode. This additional script for multiple-cases mode retrieves and accumulates information (directories of image data and annotations) in a spreadsheet (referred to as the series list) to be readily available.

The configuration file for the multiple-cases mode requires information about the location of the series list, location of a job file, location of a spreadsheet to write the final evaluation results. In the single-case mode, the configuration file requires information about file directories (image data location and the location to report the results) and the corresponding CT condition.

The multiple-cases mode operation allows for both a batch-mode run and a parallel run. This is achieved by generating a job file that lists all the required python jobs in a simple text file. Each job runs the detection evaluation for one subject at one condition. For example, for ten subjects with images at ten different CT conditions, there will be a list of 100 detection evaluation jobs. Each job includes the name and the directory of a wrapper program and information about the location of the image data. Therefore, the submission of each job launches the wrapper program and provides the information necessary for the wrapper to run. The wrapper program will then perform a series of processes and call other required programs to evaluate the CAD detections.

The job file can either be directly submitted to the system when the configuration file is called or be submitted later by the user. We use Condor to run and manage the parallel jobs. Condor is a framework that supports high-throughput computing on an extensive collection of computing resources[55]. Since Condor handles the computing resource requests according to the job requirements and logs information regarding the job (e.g., whether it finished with or without errors), it serves as a helpful framework for our pipeline.

The main components of the detection evaluation in the wrapper program are as follows: 1) loading image data and corresponding annotations, 2) performing required pre-processing such as registration of annotations to the reconstructed images, 3) finding overlaps between CAD segmentations and the annotations, 4) identifying TP, FP, and FN CAD segmentations, 5) writing results of CAD segmentations in YAML files in the corresponding folder determined in the configuration file or the series list, 6) creating and storing visualization of TP or FP CAD segmentations, 7) (if in multiple-cases mode) running a batch-based analysis and calculation of mean or median sensitivity and FP numbers, 8) reporting and storing results (TP and FP detections) in spreadsheets.

As mentioned in step 2, a registration process may be needed between the reconstructed images and the annotated data. We found a few structural misalignments between the pipeline reconstructed images and the scanner reconstructed images for some subjects in our dataset. This problem was resolved via a simple B-spline image registration[56]. As mentioned previously, since the nodule's location is annotated according to the scanner images and is independent of the pipeline reconstructed images, we needed to register the annotations to the pipeline reconstructed images to avoid misalignment issues.

In assessing overlaps between CAD segmentation masks and the true nodules, for each nodule, we evaluated whether the nodule's center or the points on the nodule's axial diameter, acquired from the annotations, overlap with the CAD segmented mask. If there was an overlap identified between a CAD segmentation and the true nodules, then the segmented ROI will be reported as a TP, and otherwise as a FP.

2.2.2.3 Module for Radiomic Feature Calculation

Once the module that evaluates CAD nodule detections finishes its job, it will create a spreadsheet containing a list of TP detections. The spreadsheet can then be used to acquire a list of cases with their location and information (nodule ID, size, etc.) about the TP segmented masks. The radiomic feature calculation module involves the following steps: 1) submitting a configuration file, 2) generating a job file, 3) calculation radiomic features in multiple-cases mode, 4) reporting calculated radiomic features in a spreadsheet.

First, via submission of a configuration file (in YAML format), the list of jobs required to calculate the radiomic features will be written. Then the job file can be submitted to Condor to run feature calculation of multiple cases in parallel, or it can be forwarded to the command line to run the feature calculation jobs in a batch mode.

The configuration file contains information about the spreadsheet that has information about TP detections, location for saving the job file, directory for saving the calculated features (in the form of a spreadsheet), information regarding specific settings for radiomic feature calculation (e.g., number of discretized gray levels, etc.), and information regarding the required wrapper program (feature calculation wrapper) that runs the feature calculation. The wrapper script consists of specific functions that pre-process images and segmented nodules and call the radiomic feature calculation functions. If for different image datasets and in various studies, different image pre-processing is desired, or additional radiomic features are of interest, a different wrapper shall be created in advance and be referred to in the configuration file.

The job file consists of a list of jobs for each subject at each CT condition. So, for example, for ten subjects with images at ten different CT conditions, there will be a list of 100 jobs. Each job identifies the wrapper (set in the configuration file), the path to the location of a subject's image and the location of the TP segmented ROI, and the output (a spreadsheet) set in the configuration file. When the job file is submitted to Condor, each job runs the wrapper independently. The wrapper performs required image pre-processing (if any), calls functions for the needed radiomic features, and finally writes the calculated measurements in the output spreadsheet.

The functions that the wrapper calls for radiomic feature calculation are based on an in-house quantitative imaging library (QIA) and Pyradiomics radiomic feature calculation software[57]. The radiomic features available in the QIA library include first-order radiomic features, such as mean, standard deviation, median, kurtosis, skewness, and a few second-order texture features of Gray Level Co-occurrence Matrix (GLCM). The radiomic features that are available from the Pyradiomics software are all the features in the following categories: GLCM, gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), neighboring gray tone difference matrix

(NGTDM), Gray level dependence matrix (GLDM), as well as first-order wavelet features. The complete list of the radiomic features and their mathematical definition is provided in the software's documentation (https://pyradiomics.readthedocs.io/en/latest/features.html). In addition, the user can set the radiomic feature calculation parameter settings to default or to desired values. These settings either specify the computation method (e.g., 2D vs. 3D) or determine factors such as preprocessing (e.g., quantized gray level numbers).

2.3 Conclusion

To achieve the goals of this dissertation, image data analysis modules were developed as part of an in-house computational pipeline. This dissertation's image data analysis modules form an automatic framework that manages and analyzes large sets of CT image data that represent a wide range of acquisition and reconstruction conditions. This framework tracks and evaluates the automatic lung nodule detections by a computer-aided detection software, manages ground truth nodule annotations, and a database consisted of information about image data, patient nodules (such as size, composition, etc.), and information about nodule detections and segmentations (TP, mask diameter, etc.). Furthermore, this framework performs post-segmentation analysis by calculation of radiomic features from the segmented regions.

The availability of an image dataset that represents a wide variation of CT acquisition and reconstruction conditions is a challenge. However, the in-house computational pipeline overcomes such limitations by generation and automatic handling of thousands of CT images required to represent the desired range of image acquisition and reconstruction modes. Therefore, the automated and modular analysis framework developed in this dissertation accelerates quantitative CT image analysis.

Chapter 3 Imaging Data and Patient Cohorts

3.1 Patient Population

Under an IRB approved protocol, an in-house data archive composed of DICOM image data and raw CT projection data has been collected from approximately 1400 lung cancer screening cases. These data were collected from patients undergoing a clinically indicated low dose screening scan within the UCLA Health System.

Patients were scanned using low dose protocols using the following multidetector Siemens CT scanners: Definition, Definition AS64, Force, Sensation 64, and Definition Flash (all from Siemens Healthineers, Forchheim, Germany). For each scan performed, image data was reconstructed using 1mm slice thickness and 1mm spacing between images.

All cases had reports from radiologist's clinical interpretations which included information relating to Lung-RADS[54] categories. Trained imaging research lab technicians used the information in the structured radiologist report and annotated the two axial perpendicular diameters of all the nodules with a diameter $\geq 4mm$ on the 1mm reconstruction.

For the data used in the studies in this dissertation, only the raw data from the Definition, Definition AS64 and Definition Flash scanners were included because it was a pre-requisite for the reconstruction pipeline that was describe in Chapter 2. These scans were acquired with the key acquisition and reconstruction parameters of 120 kV, Quality reference mAs of 25, collimation of 64 × 0.6 mm (using the z- flying focal spot) for the Definition AS and 128 x 0.6mm (using z-flying focal spot) for the Flash, 0.5s rotation time, pitch of 1.0, reconstructed slice thickness of 1.0 mm, and B30 reconstruction kernel with CareDOSE 4D and Care kV.

For each patient, only the largest representative nodule was included in the study. Another inclusion criterion for the scans was the presence of a radiologist-reported nodule with a diameter $\geq 4mm$.

For each of the patient studies presented in this dissertation, several different patient cohorts were constructed selected from the in-house data archive described above. The selection of the cohorts for each study was in accordance with the availability of cases at the time of the studies, the pre-requisites of the approach, or computational considerations. In the following sub-sections, we will provide the details of each of the cohorts used for the patient studies in Chapters 4 and 6-8. The dataset used for the phantom study will be described thoroughly in Chapter 5.

3.1.1 Patient Cohort 1:

Initially, for testing the robustness of lung nodule radiomic features, a set of 89 lung cancer screening subjects that met the inclusion criteria and had available annotations at the time of the study were selected. All these cases were scanned with the Definition AS 64 (Siemens Healthineers, Forchheim, Germany).

The robustness study will be discussed in Chapter 4. The study's goal in Chapter 4 was to assess the reproducibility of radiomic features in response to variation of dose, kernel, and slice thickness. The in- house high- throughput pipeline[42] processed the raw CT projection data to create a series of simulated raw data at reduced- dose levels and reconstruct the raw data using the wFBP algorithm via the free- CT tool[53] as described in Chapter 2. The resulting unique image dataset consisted of 36 different conditions representing a wide range of dose levels, reconstruction kernels, and slice thicknesses, as shown in Table 3-1. An example of images at 36 CT conditions was already provided in Chapter 2.

Table 3-1. Patient cohort 1 scans: CT parameters of image dataset generated by the reconstruction pipeline

	Dose Level	**Slice Thickness**	**Reconstruction Kernel**[b]
CT parameter ranges	100%[a], 50%, 25%, 10%	2mm, 1mm,0.6mm	Smooth (k1), Medium (k2), Sharp (k3)

[a] 100% dose level represents the standard lung cancer screening dose with CTDIvol ≅2mGy
[b] Smooth, medium, and sharp kernels correlate to wFBP settings of B20, B45, and B60, respectively for Siemens

The in-house pipeline was not designed for an iterative reconstruction algorithm. Therefore, to acquire scans of subjects with an iterative algorithm, the simulated low dose raw data was taken back to the scanner, and images were reconstructed with the CT parameters in Table 3-1 using the scanner reconstruction software's SAFIRE ("Sinogram Affirmed Iterative Reconstruction," Siemens Healthineers, Forchheim, Germany) algorithm. The SAFIRE sharpness settings of I26, I44, I50 were used to represent smooth, medium, and sharp settings with strength 3.

Another inclusion criterion for the study in Chapter 4 was the availability of nodule segmentation via the automatic Computer-Aided Detection (CAD) tool for at least three conditions with three different slice thicknesses in Table 3-1. Therefore, all the 89 subjects had at least three CAD segmentations with three different slice thicknesses. In addition, the availability of segmented masks at each of the three different slice thicknesses allowed for using one thickness-specific mask for all the 12 conditions at each of the three slice thicknesses.

3.1.2 Patient Cohorts 2 and 3

Chapters 6 and 7 are dedicated to investigating potential mitigation techniques to improve the reproducibility of radiomic features in response to variation of dose, wFBP slice thickness, and wFBP kernel.

Chapter 6 will discuss harmonizing radiomic features through training and applying a deep learning-based Generative Adversarial Network (GAN) model. We will further discuss that in training the GAN model, we were not limited to the availability of segmented volumes of interest

(VOI). Therefore, unlike the study in Chapter 4, there were no segmentation-based inclusion criteria for this study.

Chapter 7 will describe our other investigations into the mitigation of radiomic feature variability via the ComBat technique. The ComBat technique is directly applied to the radiomic feature data extracted from (VOI) segmented by CAD. Therefore, similar to patient cohort 1, this study had an additional inclusion criterion for the availability of segmentation masks.

At the time of these studies, based on radiologist clinical interpretation of these cases, the trained technician had annotated a total of 385 unique cases with at least one lung nodule with a diameter $\geq 4mm.$

The patients in this cohort were scanned either with Definition, Definition AS64 or Definition Flash scanners. Since the reconstruction pipeline tool was not evaluated for reconstructing raw CT data acquired with the Definition Flash scanner, the raw data from this scanner was reconstructed with the Siemens Healthineers ReconCT reconstruction software.

To minimize computational time due to the time-consuming process of image reconstruction, images were reconstructed at fewer CT conditions than cohort 1. Therefore, in addition to the condition representing the reference screening protocol (i.e., 100% dose, medium kernel, 1mm thickness), raw data was reconstructed at another seven non-reference conditions described in Table 3-2. In selecting the seven non-reference conditions, we aimed to include variability of radiomic features due to changes in each of the three CT parameters in question. Thus, the non-reference CT conditions were selected such that each condition was only different in one CT parameter compared to the reference CT condition. Table 3-2 shows the variation in reconstruction kernel (for two conditions), dose level (for three conditions), and slice thickness (for two conditions).

From these cases, 134 subjects had available CAD segmentation masks for at least three image conditions with the three different slice thicknesses shown in Table 3-2. Therefore, these 134 subjects (with their images at a total of eight CT conditions) were considered the ComBat study data in Chapter 7 (cohort 2).

As was previously mentioned, no segmentation-related restriction was enforced for the data used in the GAN study. Hence, the 385 subjects (with their images at a total of eight CT conditions) were considered the dataset for Chapter 6 (cohort 3).

The patient cohorts 1-3 were overlapped with each other. Figure 3-1 shows the subject-based division of the patients. This figure demonstrates how the three cohorts overlapped in terms of the subjects. Figure 3-2 shows the range of CT conditions included in cohorts 2 and 3.

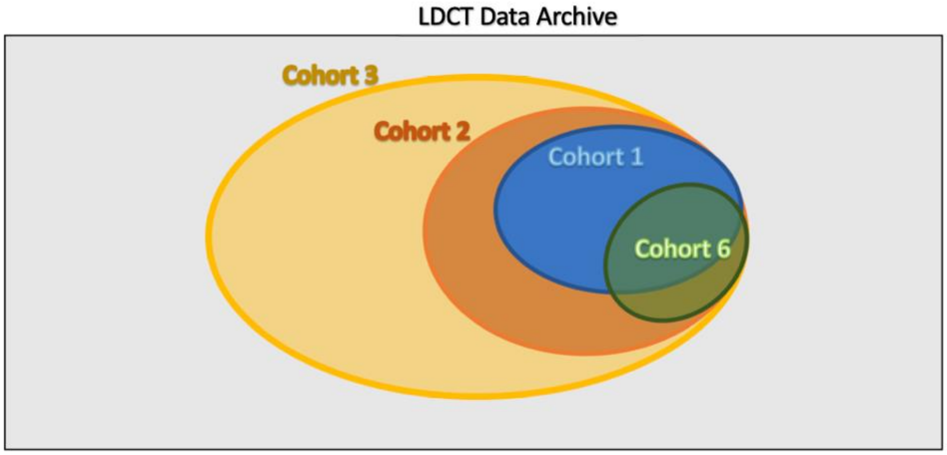


Figure 3-1. Subject-based division of low-dose CT screening cases into four patient cohorts used in this dissertation. Cohort 1: Chapter 4, cohort 2: Chapter 7, cohort 3: Chapter 6, cohort 6: Chapters 6 and 8

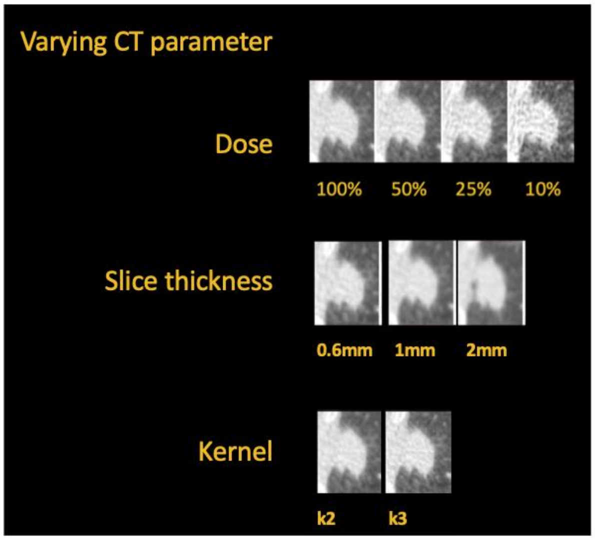


Figure 3-2. Example of images of a nodule reconstructed at the eight CT conditions included in cohorts 2-6

Table 3-2. Description of CT conditions reconstructed for patient cohorts 2-5. The first row shows the reference condition.

% Screening Dose	wFBP Reconstruction Kernel	Slice Thickness	Condition Identifier
100%	**Medium**	**1mm**	**100%, k2, st1**
100%	Smooth	1mm	100%, k1, st1
100%	Sharp	1mm	100%, k3, st1
50%	Medium	1mm	50%, k2, st1
25%	Medium	1mm	25%, k2, st1
10%	Medium	1mm	10%, k2, st1
100%	Medium	0.6mm	100%, k2, st0.6
100%	Medium	2mm	100%, k2, st2

3.1.3 Patient Cohorts 4 and 5

To train the GAN model in Chapter 6, the image data in cohort 3 was divided into a training (cohort 4) and an independent held out set (cohort 5) that included 70% (269 subjects) and 30% (116 subjects), respectively. Of the 116 cases, 55 cases had available nodule segmentation at all three slice thicknesses. Therefore, the set of 55 patients is referred to as cohort 6, and the training set is referred to as cohort 4. In testing the harmonization performance of GAN, since the availability of nodule segmentation is required for the calculation of radiomic features, cohort 6 was considered the cohort for evaluation of the GAN approach. Figure 3-3 shows the division of subjects in cohort 3.

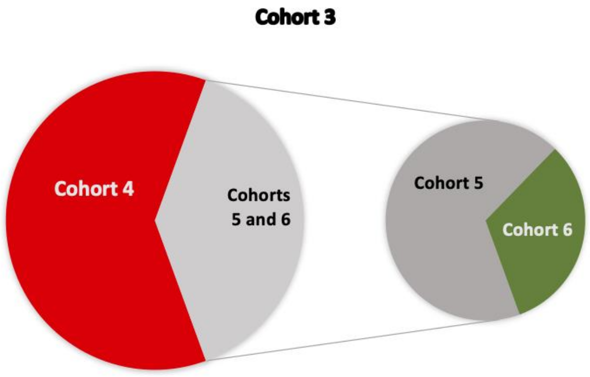


Figure 3-3. Subject-based division of cohort 3

In Chapter 8, we will compare the performance of the two mitigation algorithms of GAN and ComBat. To enable a one-to-one comparison, it was required to assess the harmonization of radiomic features calculated from the same set of subjects. As described in section 3.1.2, cohort 2 with 134 cases was used to evaluate the ComBat technique. On the other hand, the GAN model was assessed on cohort 6 with 55 cases. Therefore, after testing the potential of ComBat on a substantially large dataset (i.e., cohort 2), we ran ComBat on the unharmonized radiomic features calculated from subjects in cohort 6. Hence, we compared the two techniques of GAN and ComBat against each other through a one-to-one comparison on the same dataset.

Table 3-3 describes the sizes and composition of nodules enrolled in each patient cohort according to the clinical report obtained from the radiologist. Table 3-4 provides a summary of description of cohorts and their inclusion criteria.

Table 3-3. Description of nodules enrolled in each patient cohort according to radiologist interpretation

	Cohort 1	Cohort 2	Cohort 3	
			Cohort 4	Cohort 6
Solid ≥ 6mm	58	88	158	33
Solid < 6mm	7	14	48	7
Part-Solid ≥ 6mm w/ Solid < 6mm	9	14	14	7
Part-Solid w/ Solid ≥ 8mm	3	4	6	1
Part-Solid < 6mm	3	4	6	2
Part-Solid ≥ 6mm w/ Solid 6-8mm	1	2	4	2
Ground Glass	8	8	33	3
Total	89	134	269	55

Table 3-4. Patient cohort descriptions

Cohort	Purpose	Number of Subjects	Annotated	Raw data Required	Number of Conditions	CAD Segmentations at 3 slice thicknesses
1	Feature Robustness	89	Yes	yes	36	Yes
2	ComBat Mitigation	134	Yes	yes	8	Yes
3	GAN Mitigation (total)	385	yes	yes	8	No
4	GAN Mitigation - training set	269	yes	yes	8	No
5	GAN Mitigation - held out set	116	yes	yes	8	No
6	GAN Mitigation - test set/comparison cohort	55	yes	yes	8	Yes

Chapter 4 Investigating Reproducibility of Lung Nodule Radiomic Features across CT Acquisition and Reconstruction Parameters[2]

4.1 Introduction:

4.1.1 Background

As was discussed in Chapter 1 of this dissertation, despite the widespread use of radiomics in research, radiomics still faces uncertainties and concerns in its reliability which can inhibit its adoption into routine clinical practice. Since different choices of CT scan parameters can affect image quality differently, there is a risk that the radiomic feature quantification can differ between different datasets with heterogenous set of CT parameters.

Generalizability or reproducibility is a crucial requirement for radiomic features to serve as reliable imaging biomarkers. If the radiomic features that are used in the decision support systems are not reproducible, they may cause inconsistency of measurements and predictions. Hence, it is necessary to have an understanding of the reproducibility of radiomic features.

[2] The study in Chapter 4 is based on the previously published study of: Emaminejad N, Wahi- Anwar MW, Kim GHJ, Hsu W, Brown M, McNitt- Gray M. Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters. *Med Phys*. Published online April 13, 2021. doi:10.1002/mp.14830

4.1.2 Motivation and Purpose

The reproducibility of radiomic features has not been well established in clinical datasets. Additionally, as was discussed in Chapter 1, while there are several studies that have assessed reliability of radiomics in phantom datasets, there is still a lack of sufficient knowledge regarding the impact of CT acquisition and reconstruction parameters on radiomic features of lung nodules in clinical patient datasets. On the other hand, available patient studies may be limited in terms of the dataset and the range of parameters under investigation. Another limitation is that these studies have often conducted univariable analyses that assess parameter impacts individually without describing potential interactions among CT parameters. These limitations can be partly due to the challenge in acquiring large and controlled dataset (with multiple reconstructions per patient) that allows for validation of robustness of radiomics across wide range of CT conditions.

Motivated by these facts and to address the existing knowledge gap, in Specific Aim 2 of this dissertation we aimed to perform a systematic investigation of the reproducibility of radiomic features in patient datasets with lung nodules. Through utilizing the high-throughput computational pipeline (described in Chapter 2), we were able to overcome the challenge in assembling highly controlled dataset suitable for robustness analysis across a wide range of CT settings.

Thus, we assessed the effect of three CT technical factors of dose and weighted filtered back projection (wFBP) reconstruction parameters of kernel and slice thickness on radiomic features. The impact of these CT parameters was studied not only by univariable assessments but also by multivariable assessments that allow for a multi-faceted and simultaneous analysis of these parameters and reveals their interactions in affecting radiomic feature values. Figure 4-1 summarizes the study presented in this chapter.

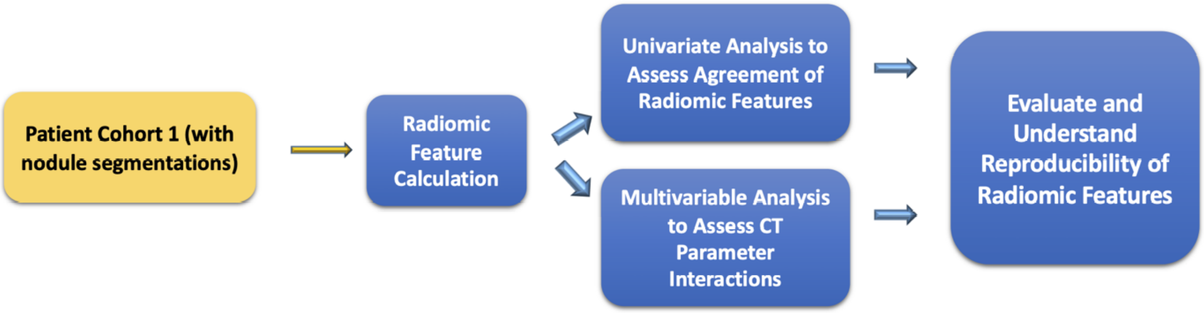

Patient Cohort 1 (with nodule segmentations)
Radiomic Feature Calculation
Univariate Analysis to Assess Agreement of Radiomic Features
Multivariable Analysis to Assess CT Parameter Interactions
Evaluate and Understand Reproducibility of Radiomic Features

Figure 4-1. Summary of the approach: a series of univariable and multivariable assessments were applied to evaluate reproducibility of radiomic features.

4.2 Materials and Methods:

4.2.1 Patient Cohort

The patient cohort 1 described in Chapter 3 was used for assessing robustness of radiomic features in this chapter. Cohort 1 consists of 89 lung cancer screening subjects who had a nodule that was ≥ 4mm in diameter in the clinical interpretation of the exam. For each patient, only the largest representative nodule was included in the study.

The raw CT projection data was processed by the computational pipeline described in Chapter 2, and a unique image dataset with 36 different CT conditions was generated. Figure 4-2 shows an example of a nodule region under these 36 image conditions, and it demonstrates how the appearance of the nodule tissue and the noise changes as CT parameters change. These conditions represent wide range of dose levels, reconstruction kernels, and slice thicknesses as shown in Table 4-1. The range of CT parameters was systematically chosen such that the resulting images cover a wide range of conditions. Additionally, this selection enabled us to push further on parameters, e.g. dose, to rigorous conditions (e.g. 10%) to understand the limits of tolerance for radiomic features.

Table 4-1. Description of CT parameters of image dataset generated by the pipeline

	Dose Level	Slice Thickness	Reconstruction Kernel[b]
CT parameter ranges	100%[a], 50%, 25%, 10%	2mm, 1mm,0.6mm	Smooth (k1), Medium (k2), Sharp (k3)

[a] 100% dose level represents the standard lung cancer screening dose with CTDIvol ≅2mGy

[b] Smooth, medium, and sharp kernels correlate to Siemens B20, Siemens B45, and Siemens B70 respectively

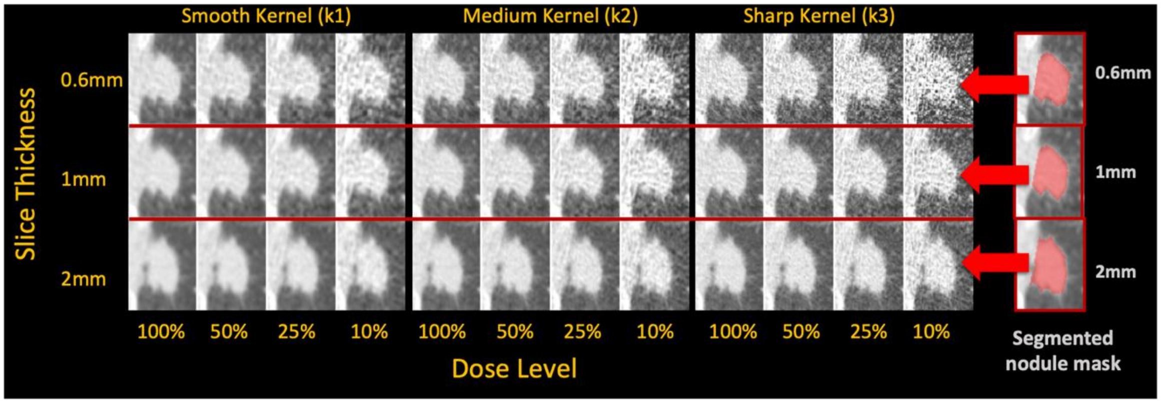


Figure 4-2. Sample nodule region at 36 different CT image conditions (four dose levels, three kernels, and three slice thicknesses) and the three different segmented nodule masks at three slice thicknesses. Each mask gets overlaid to all the images at the same slice thickness to identify the region for radiomic feature calculation.

4.2.2 Nodule Segmentation

An in-house Computer-Aided Detection (CAD) tool[43] was used to perform automatic nodule detection and segmentation. For each nodule, three volumes of interest (VOI), each segmented at a different slice thickness, were selected. As shown in Figure 4-2, each VOI was mapped to the nodule images with the same slice thickness of the VOI to perform radiomic feature calculation. The rationale for this VOI selection and mapping was as follows: since it is possible that nodule segmentations on images at different conditions vary in terms of shape and size, these variations can also impact feature values. For the purpose of this study, we aimed to control the segmentation to avoid its contribution to variation of radiomic features. Therefore, it is required to use the same VOI for feature calculations to keep nodule size and shape constant. However, because mapping the VOIs to different slice thicknesses results in inconsistencies due to different amounts of volume averaging, VOIs were only mapped to conditions with the same slice thickness. Therefore, three

VOIs (corresponding to the three investigated slice thicknesses) were selected for each case to minimize the impact on radiomic feature values caused by variation of nodule segmentation.

4.2.3 Radiomic Feature Calculation

Although the IBSI has described[17] a large number of radiomic features, we have selected a representative set of 82 well-known and frequently used features for this study. These features included features that describe intensity-based and texture-based characteristics of the nodule region. Selected features, as described by Zwanenburg *et. al.*[17], included 19 first-order descriptors of voxel intensities and heterogeneity, 12 second-order features to describe heterogeneity of nodule tissue and spatial relationships in gray level intensities from the co-occurrence matrix (GLCM), 16 gray level run length matrix (GLRLM), 16 gray level size zone matrix (GLSZM), 5 neighboring gray tone difference matrix (NGTDM), 14 Gray level dependence matrix (GLDM). Since in our study, the nodule region was kept constant within each slice thickness, radiomic features that describe nodule size or shape were not analyzed. All the descriptors used in this study were calculated using Pyradomics software package[57] using the default settings, except for GLCM features. The settings used for these descriptors are shown in Appendix A (section A.1). Each feature was calculated for each of the 89 nodules using the VOI defined for all 36 image conditions.

4.2.4 Analysis Metric for Assessing Radiomic Feature Reproducibility

A radiomic feature is considered reproducible when it shows strong agreement between its calculations under different image conditions (i.e., acquisition and reconstruction conditions). In order to evaluate the reproducibility of radiomic features among various CT image conditions, we measured the inter-condition agreements of radiomic feature values through Overall Concordance Correlation Coefficient (OCCC)[58]. OCCC is the weighted average of all pairwise Concordance Correlation Coefficients (CCC)[59] between any two image conditions (refer to section A.2 in

Appendix A). According to the proposal by McBride[60] and similar works[61,62], CCC values of equal or higher than 0.9 are considered as moderate to strong agreement, hence in this study OCCC $\geq$ 0.9 was considered as strong agreement. Therefore, a radiomic feature with OCCC ≥ 0.9 among a set of CT image conditions was considered as reproducible within that condition set.

4.2.4.1 Inter-condition Reproducibility Among All 36 Conditions

Initially, to obtain an overall understanding to determine whether radiomic features vary in response to CT parameter variations in our dataset, we assessed inter-condition reproducibility. This involved measuring the radiomic feature value agreement between all the 36 available combinations of CT parameters. In this analysis, OCCC ≥ 0.9 for each radiomic feature indicates high agreements and inter-condition reproducibility among all the 36 conditions. OCCC values for all radiomic features were then demonstrated in a bar plot.

4.2.4.2 Intra-parameter Reproducibility with Respect to Individual Parameters

The inter-condition analysis among 36 conditions provides information as to whether radiomic features show variation in general. However, to understand the details of individual CT parameter impact on radiomic features, we assessed intra-parameter agreement of radiomic feature values. For each radiomic feature, a series of univariable analysis was performed by selecting subset of conditions in which only one CT parameter varied while the two other CT parameters were kept constant. Intra-parameter agreement of radiomic feature values was measured among different levels of the varying CT parameter via OCCC. Figure 4-3 (a), (b) and (c) each show the set of univariable analyses for each of the three CT parameters and their corresponding subset of conditions. For example, to understand the impact of dose variation, intra-parameter agreement of radiomic features (Figure 4-3 (a)) was assessed as follows: subsets of conditions were selected wherein each subset, the kernel and slice thickness were fixed, but the dose varied from 100% to

10%. Each subset had a unique combination of fixed kernel and slice thickness; given three different kernels and three slice thicknesses, there were nine subsets for analysis of the effects of dose level. In each subset, agreement assessment with respect to dose variation is shown as $d.k_i\_st_j$ at kernel k_i and slice thickness st_j. The agreement ($OCCC_{d.k_i\_st_j}$) was then measured within each subset to identify whether the variation of dose impacts the feature values at kernel $k_i$and slice thickness st_j (refer to section A.2 in Appendix A).

For each CT parameter, a heatmap was generated using the OCCC values of the corresponding subsets to visualize the agreements of each radiomic feature with respect to that CT parameter. The radiomic features that had OCCC $\geq$ 0.9 across all the corresponding subsets for a CT parameter, were considered reproducible against variation of that CT parameter within the ranges that were explored in this study. For example, in Figure 4-3 (a), for a feature to be considered reproducible against dose values of 10% - 100%, it has to have $OCCC_{d.k_i\_st_j}$ $\geq$ 0.9 for all k_i and st_j levels of kernel and slice thickness.

4.2.4.3 Assessing Interaction of CT Parameters in Affecting Radiomic Feature Values

Multivariable analysis was performed to study interaction of CT parameters on radiomic feature values. For each radiomic feature (y), three-way ANOVA was fitted using kernel (α) and dose (β), and slice thickness (γ) (as categorical independent variables) as shown in equation (1). In this equation, kernel (α) is at three levels of (k_1, k_2, k_3), dose (β) is at four values of [100, 50, 25, 10], slice thickness (γ) has three values of ($0.6mm, 1mm, 2mm$). So, three main factors of kernel, dose, slice thickness, and two-way interactions of kernel and dose ($\alpha\beta$), kernel and slice thickness ($\alpha\gamma$), and dose and slice thickness ($\beta\gamma$), and a three-way interaction term ($\alpha\beta\gamma$) were included in the model. The interaction terms were tested in fitting the radiomic feature values. p-value $\leq$ 0.05 is used for the level of significance indicating the rejection of the null hypothesis (equations 2-5) and determined the significance of interaction between the corresponding CT parameters.

$$y_{ijkl} = \mu_{\ldots} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \qquad (1)$$

$where\ \ i\ = 1,2,3\ for\ kernel, j = 1,2,3,4\ for\ dose, k = 1,2,3\ for\ slice\ thickness, and\ l =$ $1\ \ldots, 89\ for\ number\ of\ patients.$ y_{ijkl} is the radiomic feature value for lth subject from a population with grand mean of $\mu_{\ldots}$ and variance of σ^2, and $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ is the error term.

$$H_0: \alpha\beta_{ij} = 0\ , \qquad H_a: \alpha\beta_{ij}\ \neq 0 \quad for\ kernel\ i\ and\ dose\ j \qquad (2)$$

$$H_0: \alpha\gamma_{ik} = 0\ , \qquad H_a: \alpha\gamma_{ik}\ \neq 0 \quad for\ kernel\ i\ and\ slice\ thickness\ k \qquad (3)$$

$$H_0: \beta\gamma_{jk} = 0\ , \qquad H_a: \beta\gamma_{jk}\ \neq 0 \quad for\ \ dose\ j\ \ and\ \ slice\ thickness\ k \qquad (4)$$

$$H_0: \alpha\beta\gamma_{ijk} = 0\ , \qquad H_a: \alpha\beta\gamma_{ijk}\ \neq 0 \quad for\ kernel\ i, dose\ j, slice\ thickness\ k \qquad (5)$$

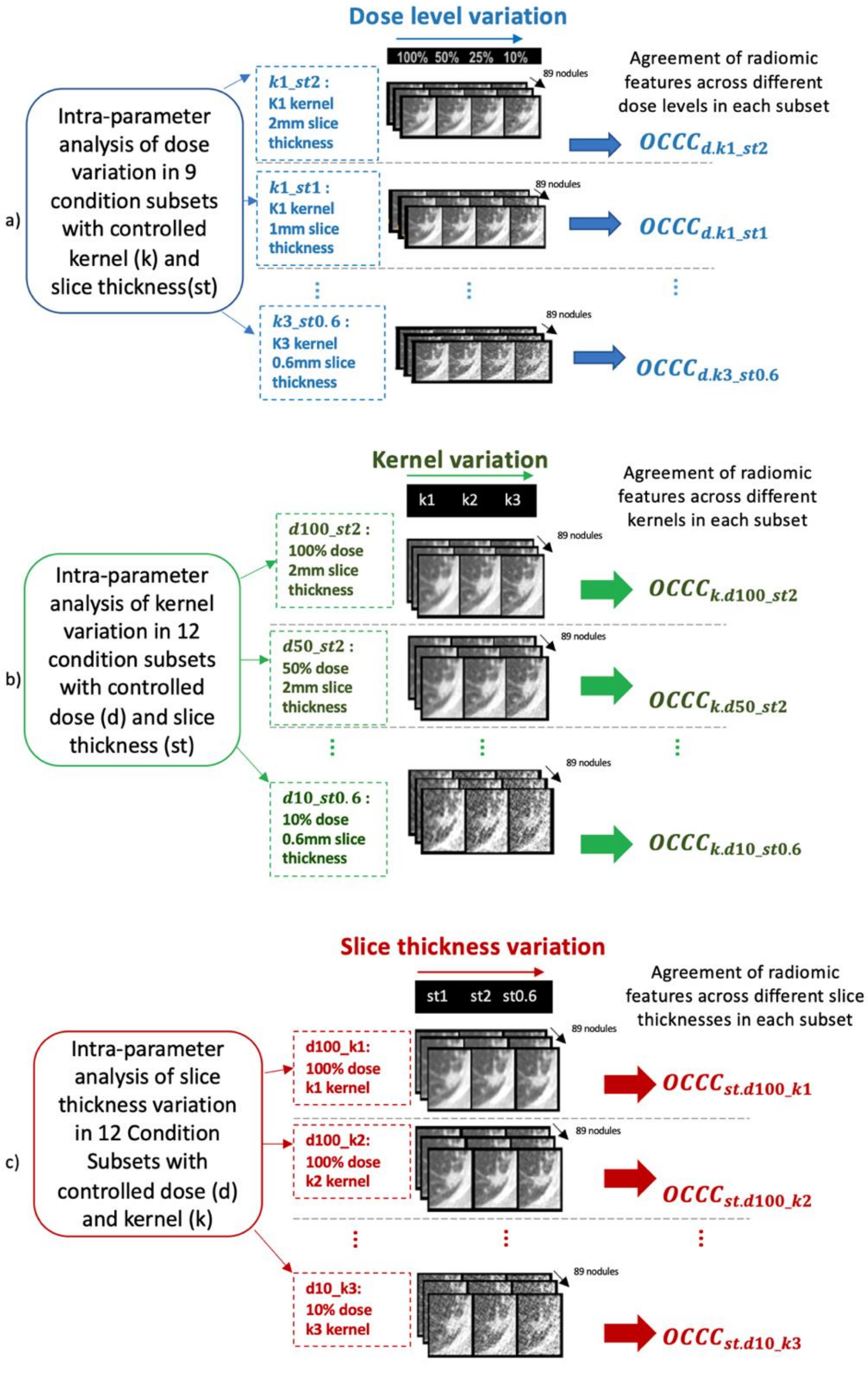

Dose level variation
100% 50% 25% 10%
89 nodules
Agreement of radiomic features across different dose levels in each subset
a)
Intra-parameter analysis of dose variation in 9 condition subsets with controlled kernel (k) and slice thickness(st)
$k1\_st2$: K1 kernel 2mm slice thickness
$OCCC_{d.k1\_st2}$
$k1\_st1$: K1 kernel 1mm slice thickness
$OCCC_{d.k1\_st1}$
$k3\_st0.6$: K3 kernel 0.6mm slice thickness
$OCCC_{d.k3\_st0.6}$
Kernel variation
k1 k2 k3
Agreement of radiomic features across different kernels in each subset
b)
Intra-parameter analysis of kernel variation in 12 condition subsets with controlled dose (d) and slice thickness (st)
$d100\_st2$: 100% dose 2mm slice thickness
$OCCC_{k.d100\_st2}$
$d50\_st2$: 50% dose 2mm slice thickness
$OCCC_{k.d50\_st2}$
$d10\_st0.6$: 10% dose 0.6mm slice thickness
$OCCC_{k.d10\_st0.6}$
Slice thickness variation
st1 st2 st0.6
Agreement of radiomic features across different slice thicknesses in each subset
c)
Intra-parameter analysis of slice thickness variation in 12 Condition Subsets with controlled dose (d) and kernel (k)
d100_k1: 100% dose k1 kernel
$OCCC_{st.d100\_k1}$
d100_k2: 100% dose k2 kernel
$OCCC_{st.d100\_k2}$
d10_k3: 10% dose k3 kernel
$OCCC_{st.d10\_k3}$

Figure 4-3. Measuring intra-parameter agreements of radiomic feature values to understand individual CT parameter impacts. Univariable agreement analysis due to variation of a) dose, b) kernel, and c) thickness (e.g. $OCCC_{d.k_i\_st_j}$ assesses impact of dose d by measuring agreement of radiomic features at fixed kernel k_i and slice thickness st_j).

4.3 Results:

4.3.1 Results of Inter-condition Reproducibility Analysis

When inter-condition reproducibility of radiomic feature was calculated among all 36 different CT conditions, all features had OCCC<0.9, as shown in Figure 4-4. This indicates that no feature is sufficiently robust to feature variation due across all 36 conditions. Among these features, first order features of mean and median intensity had OCCC≅ 0.85 had OCCC>0.8.

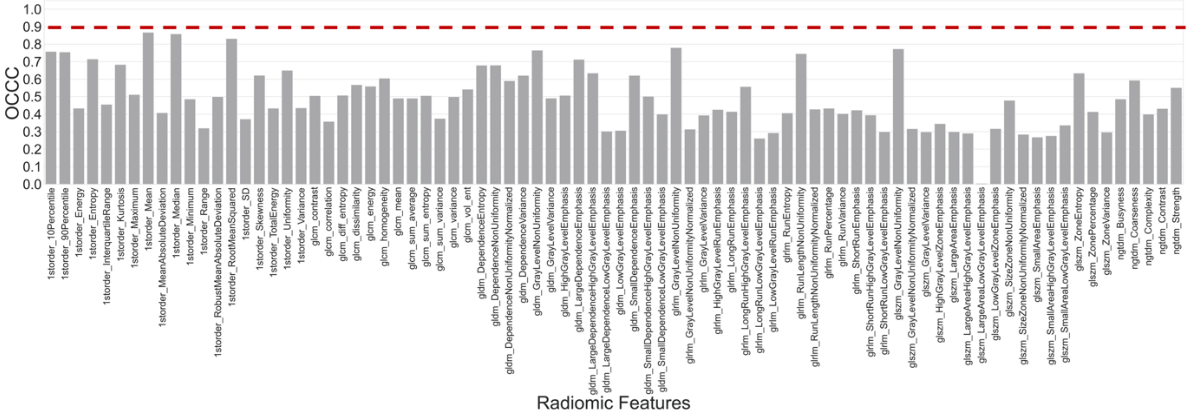


Figure 4-4. Inter-condition agreement of radiomic features among 36 conditions. Vertical axis shows agreements of each feature value. Red dashed line shows the threshold of OCCC= 0.9 to indicate reproducible features across all 36 conditions.

4.3.2 Results of Intra-parameter Reproducibility Analysis

4.3.2.1 Univariable Dose Analysis

The intra-parameter agreement of radiomic features after dose variations, measured within each of the nine subsets of CT conditions with constant kernel and constant slice thickness, indicated that several radiomic features are not reproducible against variation of dose. Figure 4-5(a) show heatmaps of $OCCC_{d.k_i\_st_j}$ for radiomic feature values in response to variation of dose in each subset. Light green and dark green colors show reproducible features (OCCC≥0.9). Table 4-2 shows that 9 radiomic features were reproducible with respect to dose variations within all nine subsets of $k_i\_st_j$. However, one first order, five GLCM, seven GLDM, nine GLRLM, ten GLSZM, and two NGTDM features were always impacted by dose variations in any given condition subsets of $k_i\_st_j$ (e.g. GLCM variance had $OCCC_{d.k_i\_st_j} < 0.9$ in all nine subsets in Figure 4-5(a)). The rest of the radiomic features responded differently to variation of dose level. These features were only reproducible at certain subsets.

Figure 4-6(a) shows the total number of radiomic features that were reproducible against variation of dose in each subset. In $k_1\_st_2$ subset, with the smoothest kernel and thickest slice, 100 features are reproducible against variation of dose while in $k_3\_st_{0.6}$ subset, that has the sharpest kernel and thinnest slice, this number reduces to 19.

Figure 4-6(a) shows the declining trend of number of features from $k_1\_st_2$ to $k_3\_st_{0.6}$ subsets. Hence, variation of dose has resulted in the least impact on radiomic feature values in $k_1\_st_2$ subset, and the most impact in $k_3\_st_{0.6}$ subset. Altogether, these results indicate that the impact of dose on radiomic feature values varied at different combinations of constant slice thickness and kernel. Overall, 94% of first order features and 42% of second order texture features were reproducible against dose variations in at least one condition subset.

4.3.2.2 Univariable Kernel Analysis:

Figure 4-5(b) shows heatmaps for $OCCC_{k.d_i\_st_j}$ of radiomic features within 12 CT condition subsets of $d_i\_st_j$ with constant dose and slice thickness. As shown in Table 4-2, the majority of features that were reproducible against dose variations in all subsets are also reproducible against variation of kernel within all 12 subsets of $d_i\_st_j$. Three first order features, ten GLCM, seven GLDM, eleven GLRLM, twelve GLSZM, and two NGTDM features were never reproducible in response to variation of kernel and had $OCCC_{k.d_i\_st_j} < 0.9$ at any given subset of $d_i\_st_j$. The rest of the radiomic features behaved differently in response to variation of kernel. According to Figure 4-6 (b), more features were reproducible at $d_{100}\_st_2$, and the number of reproducible features declined at subsets with a lower controlled dose or a thinner slice thickness. This indicates that for some radiomic features, impact of kernel on feature values varied at different combinations of dose and slice thickness. Overall, 84% of first order features and 31% of second order texture features were reproducible against kernel variation in at least one condition subset.

4.3.2.3 Univariable Slice Thickness Analysis:

Figure 4-5 (c) shows heatmaps of OCCC between radiomic features within 12 condition subsets of $d_i\_k_j$~~i_{kj}~~ with constant dose and constant kernel. Poor agreements ($OCCC_{k.d_i\_k_j} < 0.9$) among majority of radiomic features among the corresponding 12 subsets is indicative of large impact of variation of slice thickness on radiomic feature values that has resulted in only a few reproducible features in each subset as shown in Figure 4-6 (c). Among first order features, *90th percentile* feature was reproducible within four subsets in response to variation of slice thickness. First order mean intensity (referred as *1storder_Mean*) had OCCC in the range of (0.81, 0.87). One GLDM feature was also reproducible within few controlled condition subsets.

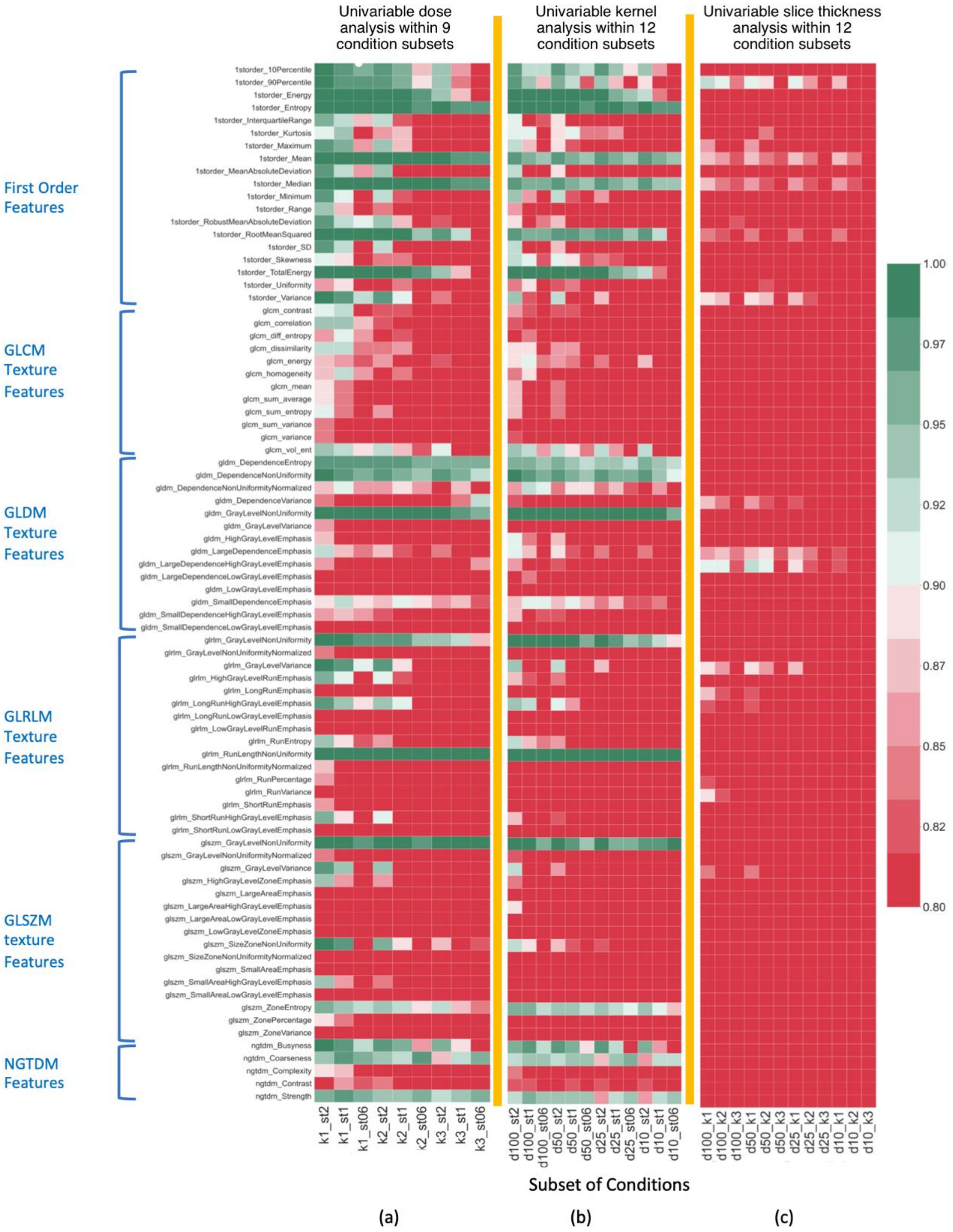


Figure 4-5. Agreement (OCCC) of radiomic features within condition subsets for analysis of a) impact of dose variation, b) impact of kernel variation, c) impact of slice thickness variation as shown by colors defined by the colormap. Colors in each column show agreements of radiomic features within the subset that is identified on the horizontal axis (e.g. $k_1\_st_2$ shows impact of dose variation at k_1 kernel and 2mm thickness). OCCC≤0.8 values were cut off at dark red color as it indicates very poor agreements.

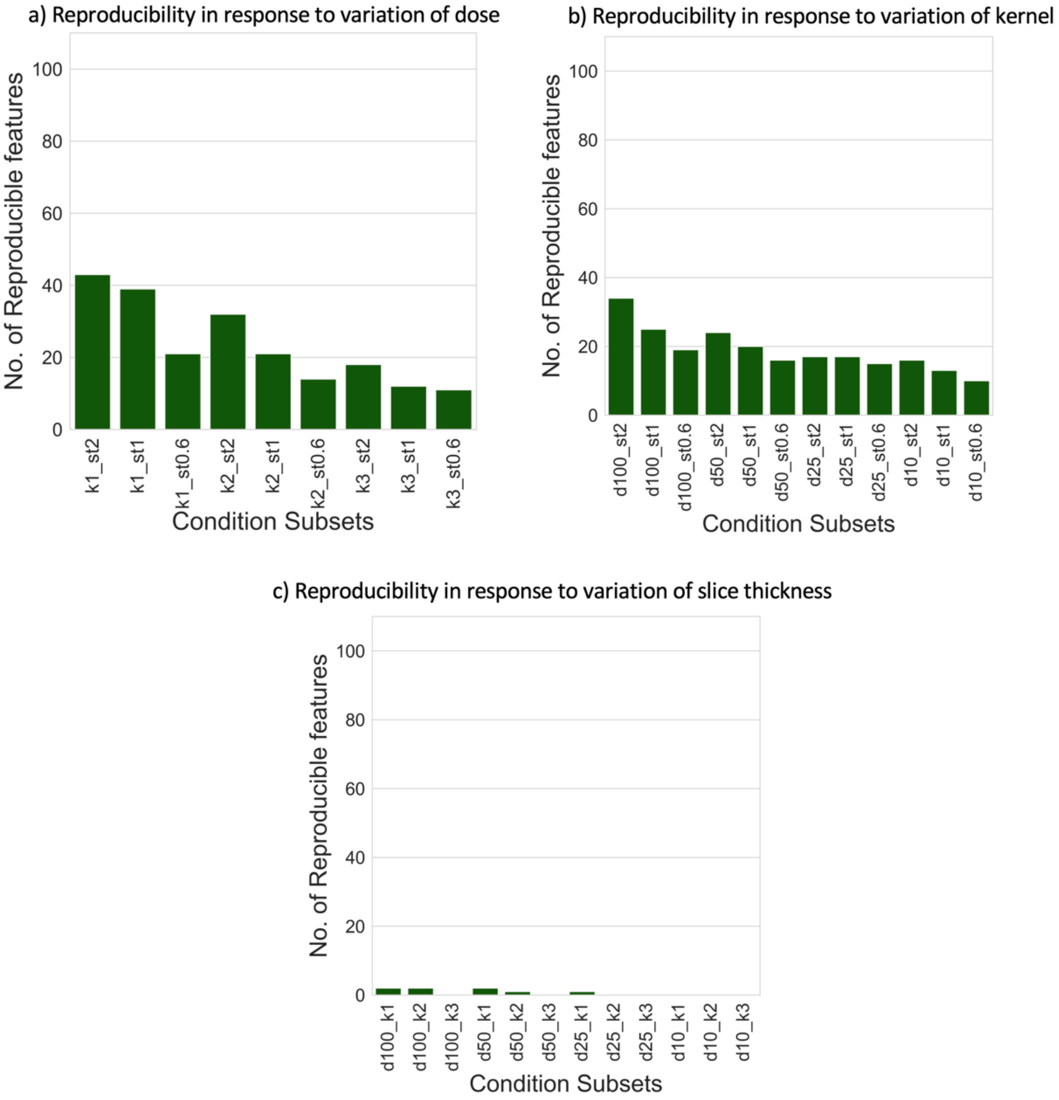


Figure 4-6. Number of reproducible features within each condition subset due to variation of an individual CT parameter when two other parameters are kept constant. (a) variation of dose in subsets with constant kernel and slice thickness, (b) variation of kernel in subsets with constant dose and slice thickness, (c): variation of slice thickness in subsets with constant dose and kernel.

4.3.3 Interaction of CT Parameters in Affecting Radiomic Feature Values

Table 4-3 summarizes the percentage of radiomic features that were impacted by interaction of CT parameters. Interaction of CT parameters affected up to 50% of radiomic features. This table demonstrates that the effect of variation of the three CT parameters (i.e., slice thickness, dose, kernel) on radiomic feature values is dependent upon each other. Interestingly, these results were in agreement with the observations in Figure 4-5. For example, a feature like mean intensity (referred as *1storder_Mean*) that has $OCCC \geq 0.9$ in response to dose and kernel variations in all corresponding condition subsets (Figure 4-5 (a) and (b)) and has $OCCC < 0.9$ in all subsets in response to slice thickness variation (Figure 4-5 (c)), is not impacted by interaction of any CT parameters ($p > 0.05$). However, for some instances of features, such as standard deviation (*1storder_SD*), radiomic features are not only dependent on variation of each individual CT parameter but are also dependent on the interaction of all three CT parameters. Other instances of features (e.g. *glcm_correlation*, *glcm_dissimilarity*, *1storder RootMeanSquared*, etc.) that have $OCCC \geq 0.9$ in few subsets and then show poor agreements in other subsets ($OCCC < 0.9$) were also among the features that were affected by interaction of CT parameters.

Table 4-2. Radiomic features that were reproducible after dose and kernel variations in all the corresponding subsets

Feature type	Reproducible against dose variations in all $k_i\_st_j$ subsets	Reproducible against kernel variations in all $d_i\_st_j$ subsets
First order	Entropy	Entropy
	Mean	Mean
	Median	Median
GLDM	Dependence entropy	Dependence entropy
	Dependence non-uniformity	Dependence non-uniformity
	Gray level non-uniformity	Gray level non-uniformity
GLRLM	Run length non-uniformity	Run length non-uniformity
GLSZM	Gray level non-uniformity	Gray level non-uniformity
NGTDM	Strength	-

Table 4-3. Percentage of radiomic features that were significantly impacted by interaction of CT parameters (with $p \leq 0.05$)

	Kernel-Dose[a] Interaction	Dose-Slice Thickness[b] Interaction	Kernel-Slice Thickness[c] Interaction	Three-way Interaction[d]
% of radiomic features	51%	59%	52%	35%

[a, b, c] To-way interactions

[d] Kernel-Dose-Slice Thickness interaction

4.4 Discussion:

The successful use of radiomics features in building reliable predictive models in a clinical setting is highly dependent on understanding and overcoming its challenges. Given that few studies have explored the robustness issue of radiomics in the context of CT image protocols in clinical datasets with chest CT scans, we aimed to expand the scope of prior patient studies[32] in understanding the reproducibility of radiomic features. Our purpose was to overcome the limitations of the current literature, such as the lack of systematic representation of CT conditions and the lack of analysis of a wide range of CT scan settings in a multi-faceted and simultaneous fashion that accounts for interactions among CT parameters.

4.4.1 Summary of Results

Our study demonstrated the lack of inter-condition reproducibility of several first order, second order texture features among 36 image conditions that consisted of a wide range of CT parameters of kernel, dose, and slice thickness (Figure 4-4). To further expand our knowledge of the impact of CT parameters on radiomic features, we assessed the individual effect of each CT parameter (Figure 4-3) along with their interactions through both univariable intra-parameter and multivariable analyses. Intra-parameter agreement analysis within several subsets of conditions with controlled parameters identified three groups of radiomic features: 1) features that were reproducible against variation of an individual CT parameter within all corresponding condition subsets ($OCCC \geq 0.9$ in all condition subsets) as shown in Table 4-2, 2) features that were never reproducible ($OCCC < 0.9$) with response to variation of an individual CT parameter in any condition subset (Figure 4-5), and 3) features that were reproducible in some but not all condition subsets (Figure 4-5).

Results of ANOVA (Table 4-3) suggest that the effect of CT parameter variation on a large number of radiomic features is bi-directional, varies in influence, and is conditional upon other CT parameters of the image. This therefore correlates with the observation that the impact of CT parameters on a group of radiomic features (group 3) varied at different condition subsets. Furthermore, from Figure 4-6, we realize that when features were calculated at images with higher noise (e.g. lower dose or thinner slice thickness) or sharper reconstruction kernels, the feature values were more susceptible to CT parameter variations as we see fewer numbers of reproducible features at these conditions.

4.4.1.1 Additional Results and Analysis

A prior study on the robustness of quantitative imaging biomarker of emphysema score reported substantial differences between the measured scores of patients over CT images with iterative (Siemens SAFIRE) reconstruction algorithms[63]. Therefore, we acquired iterative reconstructions of the same set of subjects at similar range of reconstruction CT parameters to perform a reproducibility analysis on radiomic features in SAFIRE images. The same analysis described in the method section of this chapter was applied. The description of the CT parameters and the results from the OCCC analysis is presented in Appendix A (section A.3). The results from reproducibility analysis in SAFIRE images was in agreement with findings of the prior study[63]. Radiomic features had poor reproducibility. In fact, the reproducibility of radiomic features in SAFIRE images was poorer than wFBP images. Fewer number of radiomic features had OCCC≥0.9.

4.4.2 Contribution and Comparisons to the Literature

We addressed the existing knowledge gap regarding the impact of variation of set of CT technical parameters (i.e. dose, slice thickness, and kernel) on lung nodule radiomic features extracted from patient scan datasets.

We used a unique image dataset and systematically assessed impact of wide range of CT acquisition and reconstruction conditions both individually and simultaneously. While it is not feasible to acquire multiple CT scans of patients at different dose levels, we were able to study the impact of dose on radiomic features calculated from the same patients through the application of our validated and published pipeline tool[42] and its calibrated dose simulation module[46].

Our study, compared to phantom studies, provides realistic insight on variability of radiomic features by investigating the issue of image protocol variation in a clinical dataset. For example, unlike the results from investigations on CCR phantoms[25,27,64], our intra-parameter analyses on patient dataset revealed the large impact of variation of slice thickness on a majority of lung nodule radiomic features. In clinical images, the partial volume effect and volume averaging between image object (nodule) and its background (lung tissue) at thicker slice thicknesses impacts various characteristics such as nodule's mean intensity. On the contrary, in CCR or water phantom images, since no nodule object is present, the regions depicted for feature calculations are not different compared to their background; hence, when slice thickness changes, volume averaging does not impact the radiomic feature value. Meanwhile, the impact of slice thickness has been previously reported in anthropomorphic phantom studies as well, where a nodule object different than background is present. For instance, results of studies by Kim *et. al.*[31] and Zhao *et. al.*[30], on anthropomorphic phantom images of lung with phantom nodules, also showed large variation of radiomic features due to slice thickness variation. While the nodule phantoms deviate from patient nodules - as they consist of uniform regions as opposed to possible non-uniform and heterogenous tissue of patient nodules - our results confirm that slice thickness variation impacts patient nodule radiomic features (on CT images within the explored range of image conditions) as well.

Within the intra-parameter comparisons, more radiomic features were reproducible when varying the dose level (i.e. $OCCC_{d.k_i\_st_j} > 0.9$), as compared to variation of kernel and slice thickness (as shown in Figure 4-6 (a) compared to Figure 4-6 (b) and Figure 4-6 (c)). This result is important as it indicates that dose reduction in CT imaging may be possible without affecting reproducibility of a set of radiomic features. The majority of texture features, unlike the first order features, were not reproducible in response to dose variations ($OCCC_{d.k_i\_st_j} < 0.9$). Similarly, Zhao *et. al.* [40] and Kim *et. al.*[65] reported a large variation of most texture features between two different reconstruction settings. The reproducibility of radiomic features in response to dose and kernel variations had trends that were in agreement with findings from phantom studies as well; Shafiq-ul-Hassan *et. al.*[27] reported a large dependency of texture features to kernel variations compared to dose dependency of these features. MacKin *et. al.*[66] found that most phantom radiomic features were robust against dose variations at FC18 reconstruction kernel and 5mm slice thickness within heterogenous CCR cartridges compared to homogenous cartridges.

While results of the current work support prior observations in showing reproducibility of a set of radiomic features to CT technical parameters, the findings also expand our knowledge regarding the details of reproducibility of radiomic features on a wider variation of CT parameter combinations in a clinical dataset. The current work is a systematic study that has been performed in a multivariable fashion by exploring the multi-dimensional space of possible combinations of CT settings (i.e., at 36 conditions with varying dose, kernel, and slice thickness), including the interactions of these three parameters (e.g., low dose, thin slice and sharp kernel). This has enabled us to observe that reproducibility of some radiomic features varied between different subsets of controlled CT parameters (Figure 4-5).

4.4.3 Clinical Implications

Results of our study can have clinical implications. This study can be helpful for radiomic studies focused on low-dose lung cancer screening CT cases to enable early cancer diagnosis. Since the National Lung Screening Trial (NLST) provided evidence that low-dose CT can reduce lung cancer mortality rate[67], various studies have explored the predictive power of radiomic features in lung cancer diagnosis and have found encouraging results in early cancer risk assessments [7,8]. This is an important contribution as, to the best of our knowledge, this is the first time that the reproducibility of radiomic features is assessed in depth in the context of low-dose screening CT.

The findings of this study can contribute to the design of future studies involving radiomic feature values and predictive models based on radiomic features: we have provided details of the reproducibility of a set of well-known radiomic features to variation of image acquisition protocols that possibly occur in retrospective or prospective image datasets of radiomics studies.

Interestingly, a set of radiomic features that were found as powerful prognostic biomarkers for NSCLC patients such as, GLRLM gray level non-uniformity and first order energy features, reported by Aerts *et. al.* [6], and first order features of entropy and mean intensity that were reported by Anh *et. al.*[68], were reproducible in response to dose and kernel variations in a majority of condition subsets in our study. This encourages researchers to consider a careful assessment of radiomic features before making a selection for the features to incorporate in radiomic research and predictive modeling. On the contrary, features like first order kurtosis and skewness, previously identified as prognostic and associated with genetic mutations of NSCLC patients[69], were impacted by dose, kernel, or slice thickness variations in several condition subsets. This observation this warns against use of different CT reconstruction parameters, especially slice thickness, interchangeably. Furthermore, this implies that if a high-performing prediction model (e.g. machine learning models) is achieved by training on non-reproducible radiomic features from an image dataset with homogenous set of acquisition protocols, the model's performance may not generalize well to radiomic features of CT images acquired at other protocols. On the other hand, if by using radiomic features from a heterogenous image dataset (e.g. multicenter data with heterogeneous acquisition protocols), a poor-performing model is achieved, it is possible that model's performance may improve with proper selection of reproducible radiomic features or with harmonization and preprocessing approaches[70,71].

4.4.4 Limitations

Our study has its own limitations. We have not explored other potential factors in CT medical imaging that can impact robustness of radiomic features, such as inter-scanner variabilities, other CT parameters (field of view, kV, pitch, etc.), nodule segmentation algorithm, or the impact of variation of feature definition itself or software packages. For example, recently McNitt-Gray *et. al.*[19] and Foy *et. al.*[72] reported the possible impact of use of different feature calculation software on radiomic feature (first-order and second-order GLCM feature) quantification, especially when features are computed with default software package parameters. These studies found that after applying a harmonization on parameter choices or feature computation implementations, agreement of radiomic features increased. Hence, it is expected to see similar trends in radiomic feature reproducibility against variation of CT parameters if other radiomic software packages are used at consistent settings compared to the settings used in this study.

The current study did not address diagnostic or prediction power of radiomic features, and mainly focused on robustness of these feature values. It is critical to also understand how the variations in

CT image acquisition protocol can impact the radiomic feature power and its downstream predictions. Li *et. al.* [73] found CT slice thickness as a significant factor impacting EGFR mutation prediction ability of a set of reproducible radiomic features and Kim *et. al.*[74] reported variation of nodule classification performance of radiomic-based models due to variation of CT reconstruction algorithm. Further investigation into variation of predictive performance of radiomic features can be achieved by collecting prospective image dataset with raw CT projection data as well as patient diagnosis information. Prediction power of radiomic features and their agreement at different acquisition conditions can then be assessed using OCCC or Kappa agreement index as well.

While the choice of OCCC threshold was obtained by recommendations in literature[60], it is also of interest to perform a sensitivity analysis with respect to the OCCC threshold. Additionally, while we have used a unique dataset of clinical patients with a wide range of CT reconstruction parameters and dose levels for the same patient since the dataset in hand is only from low-dose screening scans, it would be helpful to also explore these effects in images at higher dose level or in a different patient population, which remains as a future step. However, given that a wide variation and combination of different levels of CT parameters were examined in this study and the fact that findings of this study were in line with other patient studies at diagnostic dose level, further explorations may reveal similar trends in variation of radiomic features on images acquired with diagnostic scan acquisition parameters.

4.4.5 Final Notes

There are a set of important considerations for this study. First, in designing the range of CT parameters in question, we chose the range of slice thickness and kernel that reflect the current clinical practice of lung cancer screening CT scans. However, for dose, we have intentionally pushed the range to low dose levels so that we obtain an understanding from tolerance of radiomic features. While this has resulted in a range of low dose levels (e.g. 10% of screening dose) that are not currently in clinical use, lower dose levels are being explored for lung cancer screening CT. For example, recently, Fletcher *et. al.*[75] assessed nodule detectability at low radiation dose levels down to $\mathrm{CTDI_{vol}}$ of 0.4mGy (i.e. corresponding to 20% of screening dose in this study).

Additionally, while variation of CT parameters may result in variation of lung nodule segmentation itself which can then turn into further impact on radiomic feature values, in this study, we decided to isolate radiomic feature variations to differences in acquisition and reconstruction parameters by controlling the segmentation. Hence, investigation into contribution of segmentation variation to radiomic feature variability remains as a future work. In this context, it should be noted that, while we aimed to keep the nodule VOI as constant as possible between different conditions, it was not possible to use and map only one nodule VOI across all slice thicknesses due to inconsistencies and lack of precision in mapping one region from one slice thickness to a different slice thickness. Hence, for each subject, we used a different VOI for each slice thickness but kept the VOI constant among all image conditions within each slice thickness. The volume of these different VOIs used for feature calculation were in high agreement, having an OCCC of more than 0.97. It should be noted that even in the scenario of mapping of one VOI to all 36 conditions, there are still inevitable segmentation variations due to volume averaging or oversampling. Hence, the technique used in this study was identified as the best possible scenario to achieve consistent mapping of VOIs while maintaining shape as much as possible. Though, it should be noted that in our investigations, in the scenario with only one VOI mapping, slice thickness was still the CT parameter that had the greatest impact on radiomic feature values.

4.5 Conclusion:

In this chapter, we have explored the reproducibility of a set of well-known radiomic features in response to variation of CT image acquisition and reconstruction parameters of dose, kernel, and slice thickness. Since, in routine clinical imaging and between different clinical institutions, there is a possibility of differences in image acquisition protocols, it is important to understand how these differences impact the reliability of radiomic analysis. The work presented here constitutes a widely applicable experimental technique and methodology for assessing the robustness of radiomic features.

Results of this study determine that several radiomic features are impacted by the variation of CT parameters. Among the CT parameters investigated, slice thickness had the largest, and dose had the least impact on lung nodule radiomic feature values. This indicates that dose reduction may be possible without affecting the reliability of a set of radiomic features, but different slice thicknesses may not be used interchangeably. The multi-dimensional exploration of radiomic feature variability has revealed existing interactions between CT parameters in impacting radiomic feature quantification. These results can be leveraged to identify strategies for ensuring the reliability of radiomic analysis.

Chapter 5 Understanding Sources of Variability of CT Radiomic Features: Comparisons Among Phantom Datasets

5.1 Introduction

In Chapter 4, we evaluated and reported the lack of reproducibility of CT radiomic features of patient lung nodules in response to variation of dose, kernel, and slice thickness. In the Specific Aim 2 of this dissertation, we aimed to understand how the radiomic features get impacted by the variation of non-biological factors related to the CT image acquisition and reconstruction parameters. Therefore, the purpose of the study in the current chapter was to take one step further toward the goal of understanding the underlying reasons for variability of radiomic features due to the variation of slice thickness, kernel, and dose.

Since the actual texture or the true quantitative medical imaging feature values of patient tumor tissues are unknown, it was more helpful to elaborate and assess the questions asked in this study by analyzing phantom images instead of patient images. In phantom images, it is possible to have a gold standard for specific feature values (e.g., mean intensity) or to have information regarding the inherent texture of the material used in producing the phantom (i.e., whether it is homogenous or heterogeneous).

While the depicted volume of interest (VOI) on the CT scan of a phantom may not perfectly represent all the different characteristics of patient nodules on CT images, a series of phantom images are publicly available. Each phantom represents particular characteristics of patient nodules on CT scans. As an example, Credence Cartridge Radiomic (CCR) texture phantom developed by Mackin *et al.* [24] consists of man-made material that has been shown to have similar CT Hounsfield Unit (HU) values to different tumors[76]. However, the CCR phantom deviates from patient lung nodules in other aspects as it consists of blocks that are made up of one material (Figure 5-1 (b)), and there is no object with a clear boundary that differentiates itself from background material (in terms of characteristics such as intensity values, contrast, etc.). This is opposed to solid lung nodules (Figure 5-1 (d)) that are differentiated from the lung parenchyma (in terms of intensity and contrast). On the other hand, a different phantom, with synthetic nodules inserted into an anthropomorphic thoracic phantom by Gavrielides *et al.* [29], can represent various characteristics such as the existence of an object with a similar size or shape (Figure 5-1 (c)). Though, probably unlike lung nodules, the synthetic nodule[29] has a uniform density.

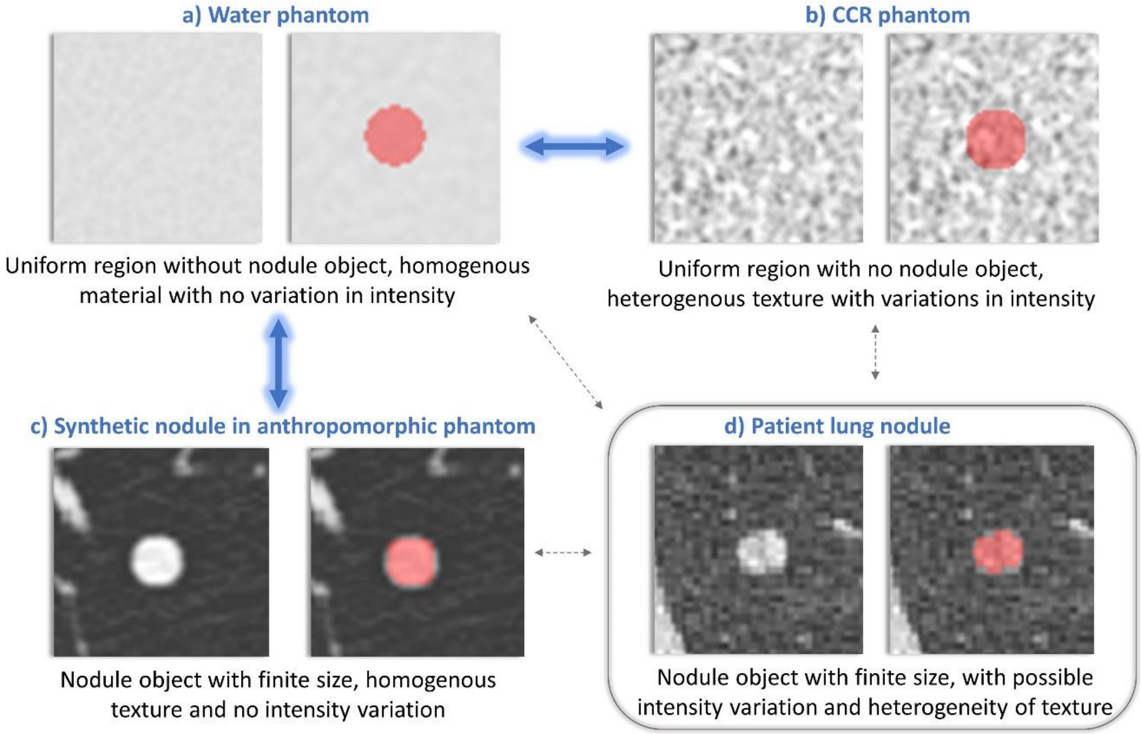

a) Water phantom
b) CCR phantom
Uniform region without nodule object, homogenous material with no variation in intensity
Uniform region with no nodule object, heterogenous texture with variations in intensity
c) Synthetic nodule in anthropomorphic phantom
d) Patient lung nodule
Nodule object with finite size, homogenous texture and no intensity variation
Nodule object with finite size, with possible intensity variation and heterogeneity of texture

Figure 5-1. Sample images from CT scans of phantoms and the depicted VOIs (in red) in comparison to sample scans of patient lung nodule and its VOI (in red). (a) water phantom, (b) CCR phantom c) Synthetic nodule in an anthropomorphic phantom. Blue arrows show the pairwise comparisons performed between the phantoms in this study.

In this study, we anticipated that by analyzing data from these different phantoms, with known properties, and by focusing on the similar aspects between patient tumor images and phantom images, we would be able to gain more insight into the underlying mechanisms in the variability of CT radiomic feature in lung nodules. For example, the findings enable us to understand how slice thickness variation and volume averaging impact the radiomic features. Additionally, we can obtain insight into how these variabilities differ between radiomic features calculated over phantoms with and without nodules. Moreover, we can understand whether inherent disparities between the homogeneity/heterogeneity of different materials result in a distinction in the variability of radiomic features in response to variation of CT image parameters of dose and reconstruction kernel.

Our prior study and other works in the field reported the lack of reproducibility of radiomic feature in CT scans in patient studies[40,77,78] and phantom studies. For example, among phantom studies, some have used the same image data or scanned the same phantom used in this study. Shafiq-Ul-Hassan *et al.* [27] assessed the impact of kernel and dose variation across a wide variety of protocols and scanners in the CCR phantom and reported the effect of the kernel on radiomic feature reproducibility. Mackin *et al.* [66] found radiomic features calculated from rubber particles in the CCR phantom were robust to dose variation due to tube current differences. Another study using the CCR phantom data[79] calculated 85 radiomic features (from all groups) and found 65 features to be robust to voxel size variation (Coefficient of variation < 50%). Several other studies[30,31,80] assessed the reproducibility of radiomic features in the anthropomorphic phantom CT scans and reported the impact of slice thickness variations on radiomic features.

While the studies mentioned above offer valuable insight into radiomic feature reproducibility using similar phantom image data to this study, each work is focused on only one type of phantom data. Thus, there are no comparisons between variability of radiomic features in different types of phantoms under consistent CT scan protocols. We extended the investigations mentioned above into comparisons between the different phantom types by analyzing the variability of a large group of radiomic features. We anticipated that comparisons between radiomic feature variability in various kinds of phantoms can reveal important information regarding underlying mechanisms that play a role in the variability of CT lung nodule radiomic features.

The rationale is that, while each phantom has certain similarities to patient lung nodules, the phantoms themselves have particular distinctions as described above (Figure 5-1). Hence, the differences in variability of radiomic features in different phantoms can help us understand the characteristics responsible for radiomic feature variability induced by varying CT parameters. For example, while slice thickness showed a large impact on radiomic feature in anthropomorphic studies[30,31], it had less impact in a CCR study[79]. Thus, analyzing such deviation can be helpful to understand the origin of radiomic feature variability.

Furthermore, due to possible difficulties in acquiring a proper and representative patient dataset that enables a comprehensive uncertainty analysis, reproducibility studies and studies that derive harmonization techniques for mitigation of radiomic feature variability[79,81] may be achieved through utilization of phantom image datasets instead of patient datasets. Therefore, it is necessary to compare and contrast these phantoms with each other and with patient data regarding radiomic feature reproducibility to identify any potential deviation of results.

A critical challenge in enabling phantom data comparisons is the availability of scans of both phantoms at the same CT protocols and the same scanner; the publicly available phantom datasets used in this study do not have scans with matching CT protocols. For example, in assessing the impact of kernel variations, the available CCR images were acquired at different dose levels than anthropomorphic phantom images. Furthermore, our prior study[78] showed that CT parameters interact with each other in affecting radiomic feature values. This suggests that the impact of the kernel on radiomic features can be dependent on the CT scan dose level. Hence, in assessing the effect of one varying CT parameter between different phantom datasets, we need to control the non-varying CT parameters to be fixed at the same level. In the following sections, we will discuss how we suppressed the need for having phantom datasets with matching CT protocols by including a third image dataset from water phantoms as a benchmark to be compared against the two phantom image datasets.

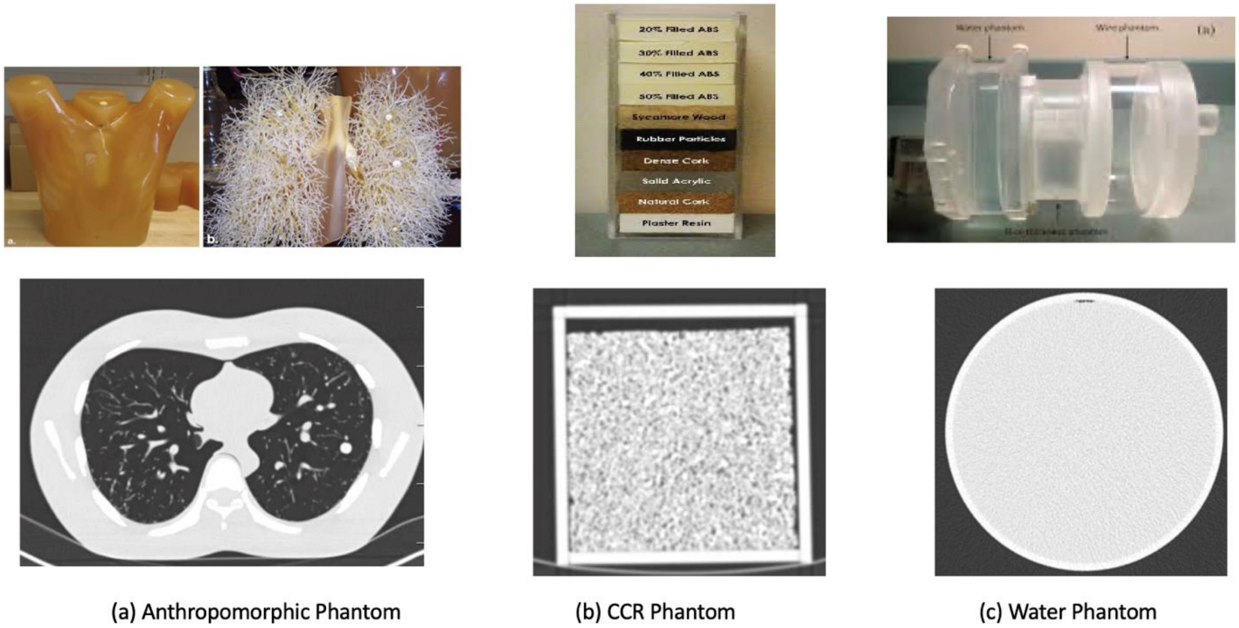
In summary, the three key contributions of this study are that, firstly, due to the lack of its availability, we performed an analysis that carried out a comparison of radiomic feature variability for three different types of phantoms (CCR, water, and anthropomorphic) in an approach that radiomic feature variations for these phantom images can be compared consistently. Secondly, we overcame the limitation in the variability of phantom datasets under matching CT acquisition protocols. Finally, we derived insight on underlying lung nodule characteristics that play a role in radiomic feature variability induced by CT parameter variations.

5.2 Materials and Methods:

5.2.1 Data

In this study, we have utilized CT scans of three different phantoms 1) a textured phantom, 2) an anthropomorphic thoracic phantom, and 3) a water phantom. Figure 5-2 shows these phantoms and an example of their CT scan.

To be consistent and to avoid inter-manufacturer variabilities, we only included image data acquired from Siemens CT scanners.



(a) Anthropomorphic Phantom (b) CCR Phantom (c) Water Phantom

Figure 5-2. Images of the three phantoms: a) anthropomorphic phantom with synthetic lesions, b) CCR phantom with heterogeneous material, c) Water phantom with a homogenous region

5.2.1.1 CCR phantom

Shafiq ul Hassan et al. [82] provided open-access CT scans of all the different CCR phantom cartridges made with various types of materials using multiple scanners and multiple CT protocols with different CT parameters. We only used the rubber particle cartridge as it was shown to be the most similar material to NSCLC tumors in terms of CT number and standard deviation[76] and has been shown to have effective atomic numbers similar to human tissues[83,84]. Table 5-1 shows the CT acquisition and reconstruction parameters that were included in our study to assess CCR images. Figure 5-3 (b) shows the CCR phantom and its CT scan. Scans were acquired with a Siemens Definition AS and Siemens Sensation 64 scanner with different dose levels, kernel, and thickness at a collimation of 64mm x 0.6mm, FOV of 250, and 120 kV.

Table 5-1. CT scan protocols for CCR phantom used in this study. CT parameters considered as reference protocol are in bold.

CT parameter being varied	Scan Protocols				Reference Protocol [a]
	wFBP Kernel	Dose (mAs)	Slice thickness (mm)	FOV	
Kernel[b]	**B20f**, B50f, B70f	**65**	**1.5**	**250**	**B20f, 65, 1.5mm, 250 FOV**
Dose[c]	**B31f**	50, 100, 200, **300**	**1.5**	**180**	**B31f, 300, 1.5mm, 180 FOV**
Thickness[d]	**B31f**	**250**	**1.5**, 2, 3	**250**	**B31f, 250, 1.5mm, 250FOV**

[a] The protocol that was considered as the reference to compare against radiomic feature values at other protocols to assess the impact of each CT parameter

[b,d] scanned via Siemens Definition AS scanner

[c] Scanned via Siemens sensation 64 scanner

5.2.1.2 Anthropomorphic Thoracic phantom

Gavrielides et al. [29] has provided an open-access CT image dataset of synthetic lung nodules inserted into an anthropomorphic thoracic phantom (Kyotokagaku Incorporated, Kyoto, Japan). These images have been acquired using multiple scanners and image protocols and were available on the Cancer Imaging Archive (TCIA) website. We have analyzed the radiomic features of the spherical 8mm nodule placed in the right lung in the scan layout 3. Table 5-2 shows the range of included image protocols acquired at collimation of 64mm x 0.6mm, kVp of 120, and pitch of 1.2 with the 64-row Siemens Definition scanner. Figure 5-3 (a) shows the anthropomorphic phantom and its CT scan.

In assessing variation of slice thickness in anthropomorphic phantom, we considered the condition with the thinnest thickness (0.75mm) as the reference protocol. This helped to understand the impact of slice thickness variation in one direction (i.e., from thin to thick).

Table 5-2. Scan protocols for the anthropomorphic phantom used in this study. Parameters for the condition considered as reference are in bold.

CT parameter being varied				Reference Protocol[a]
	wFBP Kernel	Dose (mAs)	Slice thickness (mm)	
Kernel	**B40f**, B60f	**100**	**1.5**	**B40f, 100, 1.5mm**
Dose	**B40f**	25,100,**200**	**1.5**	**B40f, 200, 1.5**
Thickness	**B40f**	**100**	**0.75**, 1.5, 3	**B40f, 100, 0.75**

[a] The protocol that was considered as the reference to compare against radiomic feature values at other protocols to assess the impact of each CT parameter

5.2.1.3 Water Phantom

A practical approach in understanding differences in variations of radiomic feature between two different phantom image datasets in response to CT parameter variations is to compare images acquired at a matching set of CT parameters. Since CT scans of the CCR phantom and anthropomorphic phantom were not available at the same image protocols, we included CT scans of a third phantom (water phantom) to serve as a benchmark.

The Water phantom images were from the water section of the QA phantom of our CT scanner (Definition AS, Siemens Healthcare, Forchheim Germany). Figure 5-3 (c) shows the water phantom and its CT scan. We scanned the water phantom with a Definition AS. The water phantom scans were acquired at the same CT acquisition and reconstruction parameters available for both of the two other phantom datasets (Table 5-1Table 5-1. CT scan protocols for CCR phantom used in this study. CT parameters considered as reference protocol are in bold. and Table 5-2) to enable consistent paired comparisons between water-CCR phantoms water-anthropomorphic phantoms at similar CT protocols. Therefore, in comparing radiomic features of the water phantom with the CCR phantom, water phantom images with the scan protocols determined in Table 5-1 were used. In comparing radiomic features of the water phantom with the anthropomorphic phantom, water phantom images with the scan protocols specified in Table 5-2 were used.

5.2.2 Comparisons between phantom images

Paired comparisons were made between water-CCR phantoms and water-anthropomorphic phantoms at similar CT protocols, shown in Table 5-1 and Table 5-2, respectively.

There is a substantial difference in the homogeneity of the water phantom and CCR phantom. So, the pairwise comparison of radiomic feature values of water-CCR phantoms, induced by each CT parameter fluctuations, reveals whether dissimilarities in the heterogeneity of VOI material can result in deviation of radiomic feature variability.

On the other hand, while both water phantom and synthetic nodule insertions in anthropomorphic thoracic phantom have homogenous texture, they are different in terms of existence or non-existence of an object differentiated from its background. Therefore, the synthetic nodule can better represent lung nodules in terms of shape and size. Hence, the pairwise comparisons of water-anthropomorphic phantoms can reveal insight into how the slice thickness variations impact the reproducibility of radiomic features in different phantom images and patient lung nodules.

5.2.3 VOI and radiomic feature calculation

For the water phantom and the CCR phantom, spherical VOIs with 8mm diameter were overlaid on the central part of the scanned image for radiomic feature calculation. The VOIs are the red regions overlaid on the CT scans shown in Figure 5-1(a) and Figure 5-1(b). For the synthetic nodules, the VOI was acquired by applying a computer-aided detection and segmentation software (CAD)[43] on each CT scan. The VOI segmentations were then visually reviewed for quality assurance. The red region overlaid on the CT scan shown in Figure 5-1(c) is the segmented VOI. 82 radiomic features were calculated using both an in-house feature calculation tool and the Pyradiomics[57] software. Of this total, 63 texture features were calculated: 11 gray level co-occurrence matrix (GLCM) features were calculated using in-house software with 16 gray level discretization, distance 1, and via 8-connected neighborhood in 2D with the corresponding direction vectors of: (1, 0, 0), (1, 1, 0), (0, 1, 0) and $(-1, 1, 0)$. 16 gray level run length matrix (GLRLM), 16 gray level size zone matrix (GLSZM), 5 neighboring gray tone difference matrix (NGTDM), 14 Gray level dependence matrix (GLDM) via the Pyradiomics software[57] with gray level discretization done at 16 bin count and default Pyradiomics settings. In addition to the texture features, 19 first order (density histogram-based) radiomic features were also calculated via the Pyradiomics software[57] using the default settings.

5.2.4 Data analysis

According to Figure 5-3 and Figure 5-1 and Table 5-2, for each phantom, radiomic feature variability was assessed in response to variation of one CT parameter (kernel, dose, or slice thickness) at a time where the CT parameter in question varied between different levels. In contrast, all other CT parameters were kept constant. Each time, one protocol was considered as a reference and was compared to other protocols (i.e., non-reference) to assess the variability of radiomic features induced by the fluctuations of the CT parameter in question. The variability of radiomic features was then quantitatively described by the ratio of radiomic feature values at non-reference to reference protocols. For example, in Figure 5-3, when assessing the impact of dose level variations for the CCR phantom, three non-reference protocols ($m = 3$) with different dose levels (50 mAs, 100 mAs, 200 mAs) at the same slice thickness of 1.5mm and reconstruction kernel of B31f will be compared against the reference protocol with 300 mAs and the same slice thickness of 1.5mm and reconstruction kernel of B31f.

Ratios that are very close to one indicate that the non-reference radiomic feature is very close to the reference radiomic feature. Ratios <1 suggest a variation between the reference and non-reference radiomic feature values and that the non-reference radiomic feature is less than the reference. Ratios >1 indicate that the non-reference radiomic feature is higher than the reference.

Subsequently, the variability of radiomic features was then compared between the phantoms via paired comparisons of water-CCR phantoms and water-anthropomorphic phantoms. When the mean intensity for the water phantom had a value equal to zero, it was excluded from calculating the ratios[3]. Hence, the histogram of voxel intensities (in HU) for the VOIs on different phantoms was also examined and was compared with other phantoms.

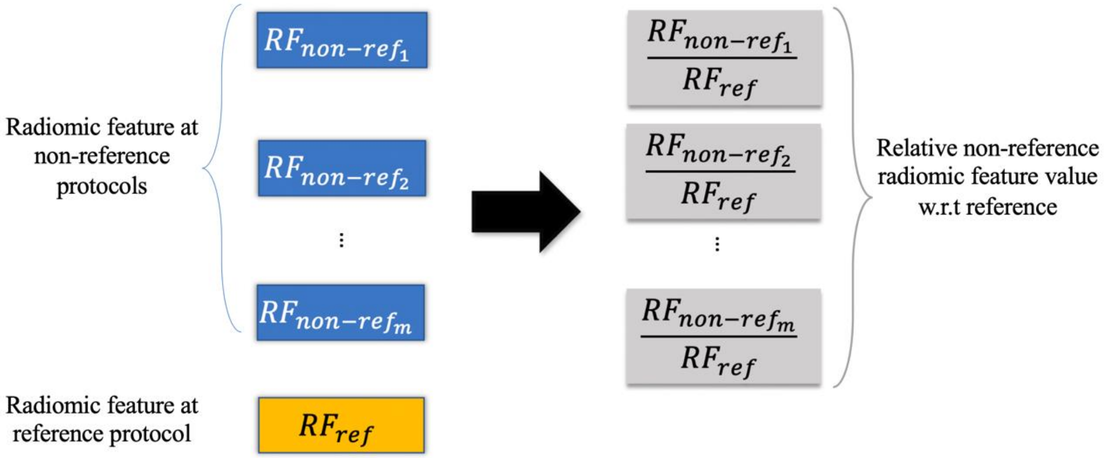

Radiomic feature at non-reference protocols
$RF_{non-ref_1}$
$RF_{non-ref_2}$
⋮
$RF_{non-ref_m}$
Radiomic feature at reference protocol
RF_{ref}
$\frac{RF_{non-ref_1}}{RF_{ref}}$
$\frac{RF_{non-ref_2}}{RF_{ref}}$
⋮
$\frac{RF_{non-ref_m}}{RF_{ref}}$
Relative non-reference radiomic feature value w.r.t reference

Figure 5-3. Comparison of radiomic feature (RF) values between reference and non-reference protocols shown in Figure 5-1 and Table 5-2. Within each phantom and for each radiomic feature, non-reference feature values ($RF_{non-ref_m}$) at non-reference condition m were compared with reference feature (RF_{ref}) values through calculating the ratios.

[3] To avoid infinity values as a result of having a denominator equal to zero.

5.3 Results

5.3.1 Radiomic Feature Variations Due to Slice Thickness Variations:

5.3.1.1 Slice Thickness Variations: Comparison of Anthropomorphic and Water phantom

Figure 5-4 demonstrates ratios of a set of first-order radiomic feature values between non-reference protocol (at 100mAs, B40f, 1.5mm thickness) and reference protocol (at 100mAs B40f, 0.75mm thickness) for water and anthropomorphic phantoms. The density histogram-based features of the VOI, such as Inter Quartile Range (referred to as IQR) and Standard Deviation (referred to as stddev), and Mean Absolute Deviation (referred to as MAD) that describe non-spatial deviations and randomness of intensity values with respect to the mean intensity had higher ratios for non-reference protocol relative to the reference protocol ($\frac{RF_{non-ref_m}{}^{Anthropomotphic}}{RF_{ref}{}^{Anthropomotphic}} > 1$). Therefore, these non-reference radiomic features had a shift toward higher values than the reference.

However, the ratios for the radiomic features of the water phantom were less than one ($\frac{RF_{non-ref_m}{}^{Water}}{RF_{ref}{}^{Water}} < 1$). For some radiomic features shown, such as entropy for both anthropomorphic and water phantoms and Root Mean Square (RMS), the ratios are very close to one, indicating low variability of the radiomic feature.

Figure 5-5 shows images of the synthetic nodule in the anthropomorphic phantom, and Figure 5-6 shows the images of the VOI over the water phantom. These figures visualize the VOI in lung and soft window settings.

To assess variation of voxel intensity values due to slice thickness variations, Figure 5-7 plots the histogram of intensity values (in HU) and fitted distributions for the synthetic nodule in the anthropomorphic phantom and the VOI depicted over the water phantom.

Thicker slice thickness reconstruction involves averaging of larger scanned volume. Since in the anthropomorphic phantom, the synthetic nodule is surrounded by a background (air) with highly negative intensity values (Figure 5-5), the volume averaging had shifted the intensity values of synthetic nodule toward negative intensity values for the VOI (see Figure 5-7 (a) and (b)). In Figure 5-5 this phenomenon is apparent as the synthetic nodule and its segmentation mask are shown both for lung window (level: -600 HU, width: 1600 HU) and soft tissue window (level: 50 HU, width: 350 HU) at three different thicknesses. It is visible that at all three slice thicknesses, the upper or lower slices of the synthetic nodule have a very low density that they disappear in the soft tissue image. As the image is reconstructed at thicker slice thickness, more slices of the nodule had vanished from the soft tissue window. This is indicative of the impact of volume averaging with the background. Therefore, the substantial variability of the other radiomic feature values calculated over the synthetic nodule can be attributable to the effect of volume averaging as well. The volume averaging has increased the randomness of pixel values in the VOI (an increase of noise). This can explain why in Figure 5-4 for standard deviation (stddev), we observed a ratio of more than one for the non-reference condition (i.e., $\frac{{RF_{1.5mm}}^{Anthropomotphic}}{{RF_{0.75}}^{Anthropomotphic}} > 1$).

However, no substantial difference is visualized between images of the VOI over the water phantom at the soft tissue window when their slice thickness changes. The VOI is surrounded by a uniform and homogenous background in the water phantom with an intensity equal to the VOI (see Figure 5-6). So, volume averaging does not result in noticeable variation in intensity values for the VOI at thicker slice thickness. As shown in Figure 5-7 (c), the variation of slice thickness in water phantom images, at the same protocols compared to Figure 5-7(a) and Figure 5-7(b), has not resulted in variation of intensity values or mean intensity of the VOI.

According to Figure 5-7(a) and Figure 5-7(b), the mean intensity values of the VOIs for the synthetic nodule have changed from 127.8 HU at 0.75 mm thickness to a negative value (-95.9 HU) at 3mm thickness. However, the gold standard for this nodule was reported at 100 HU by Gavrielides *et al.* [29]. Our reported mean intensity value is closest to this gold standard at 100mAs, B40f, 1.5mm.

In total, 37 out of 63 texture features and 13 out of 19 first-order features had larger ratios in anthropomorphic phantom compared to water phantom.

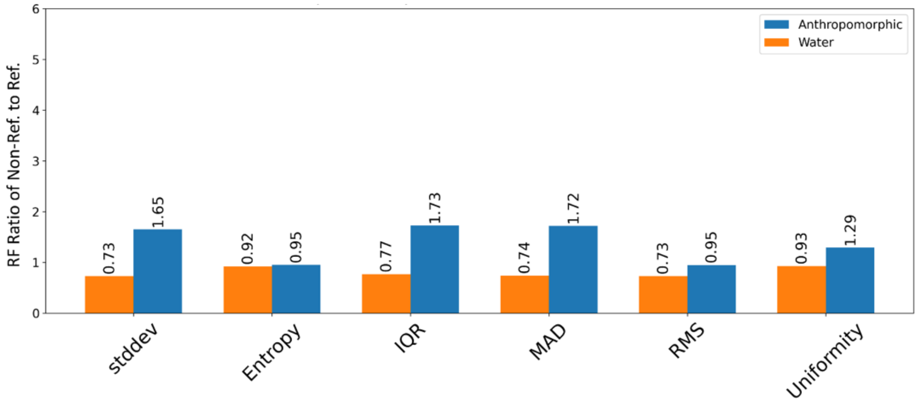


Figure 5-4. Variation of a set of first-order radiomic feature values in terms of the ratio of radiomic feature value at non-reference protocol (1.5mm, B40f, 100mAs) to reference protocol (0.75mm, B40f, 100mAs). The ratios are written for bars belonging to each feature and each phantom. Stddev: Standard Deviation feature, IQR: Inter Quartile Range feature, MAD: Mean Absolute Deviation feature, RMS: Root Mean Square feature.

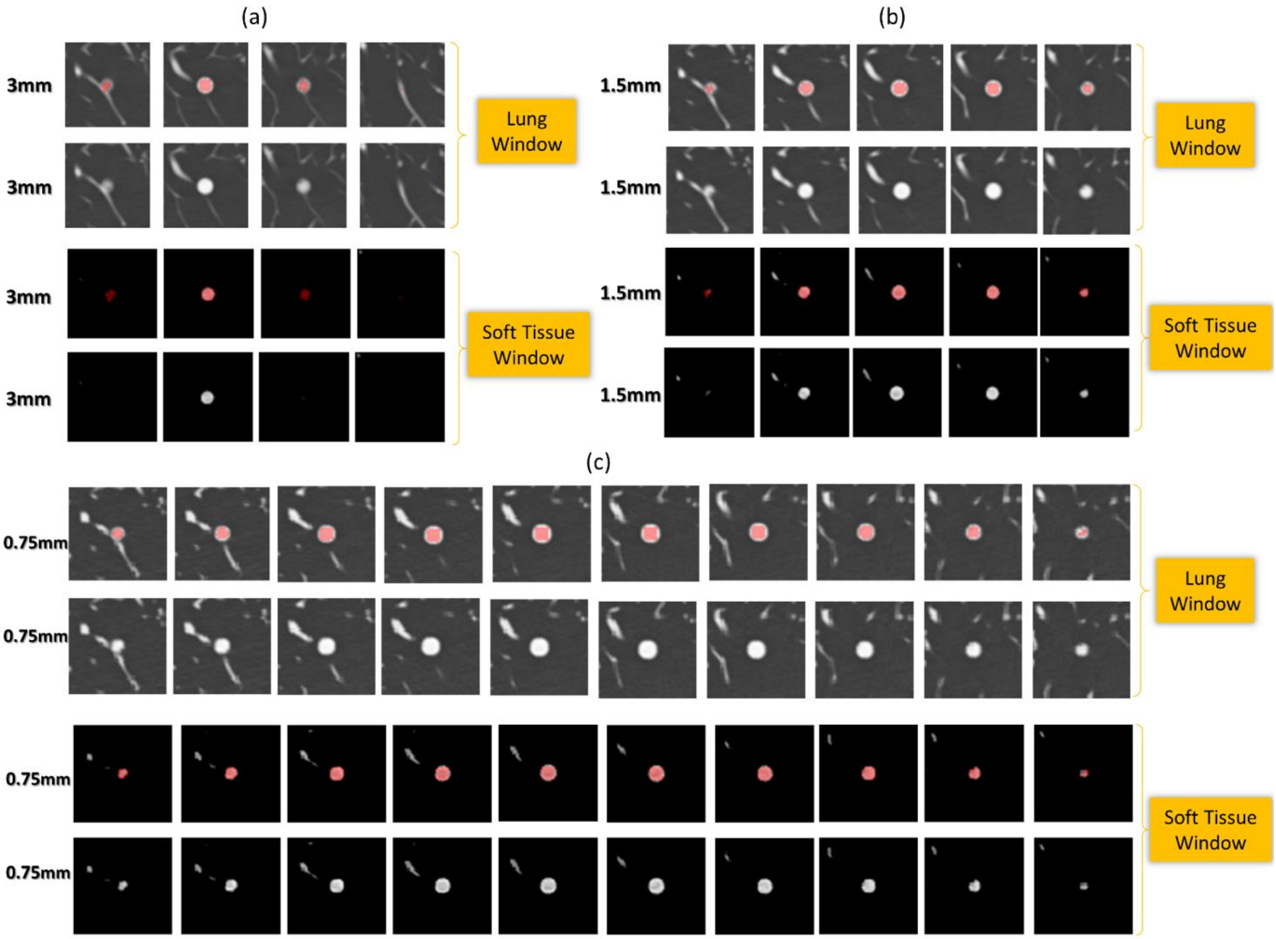


Figure 5-5. Images of the synthetic nodule in anthropomorphic phantom visualized in lung window setting (level: -600 HU, width: 1600 HU) and soft tissue window setting (level: 50 HU, width: 350 HU) with and without the overlay of the segmented VOI (the region in red) for scans with different slice thicknesses. (a) B40f, 100mAs, 3mm, (b) B40f, 100mAs, 1.5mm, (c) B40f, 100mAs, 0.75mm.

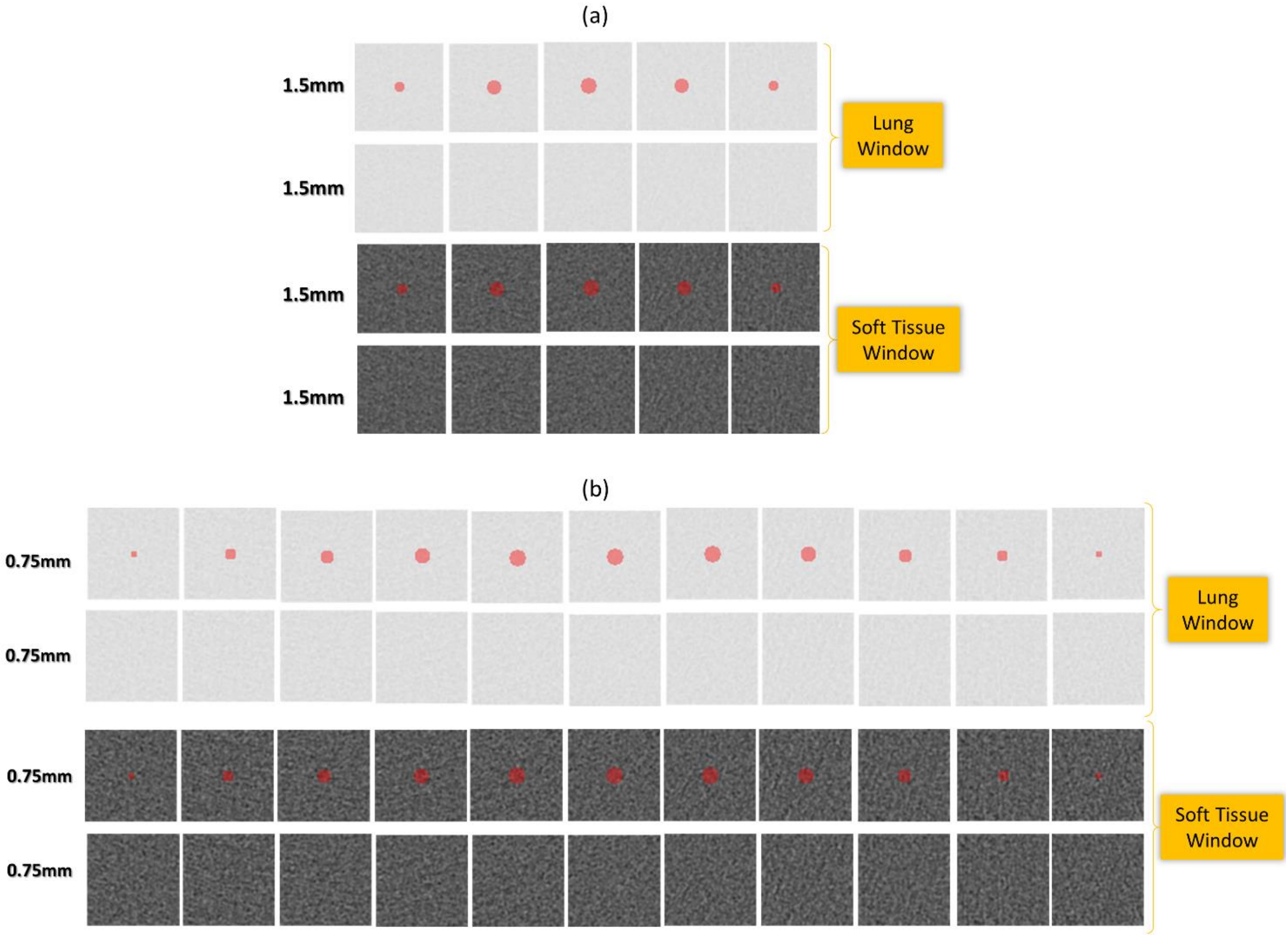


Figure 5-6. Images of the VOI over water phantom visualized in lung window setting (level: -600 HU, width: 1600 HU) and soft tissue window setting (level: 50 HU, width: 350 HU) with and without the overlay of the VOI (the region in red) for scans with different slice thicknesses. (a) B40f, 100mAs, 1.5mm, (b) B40f, 100mAs, 0.75mm.

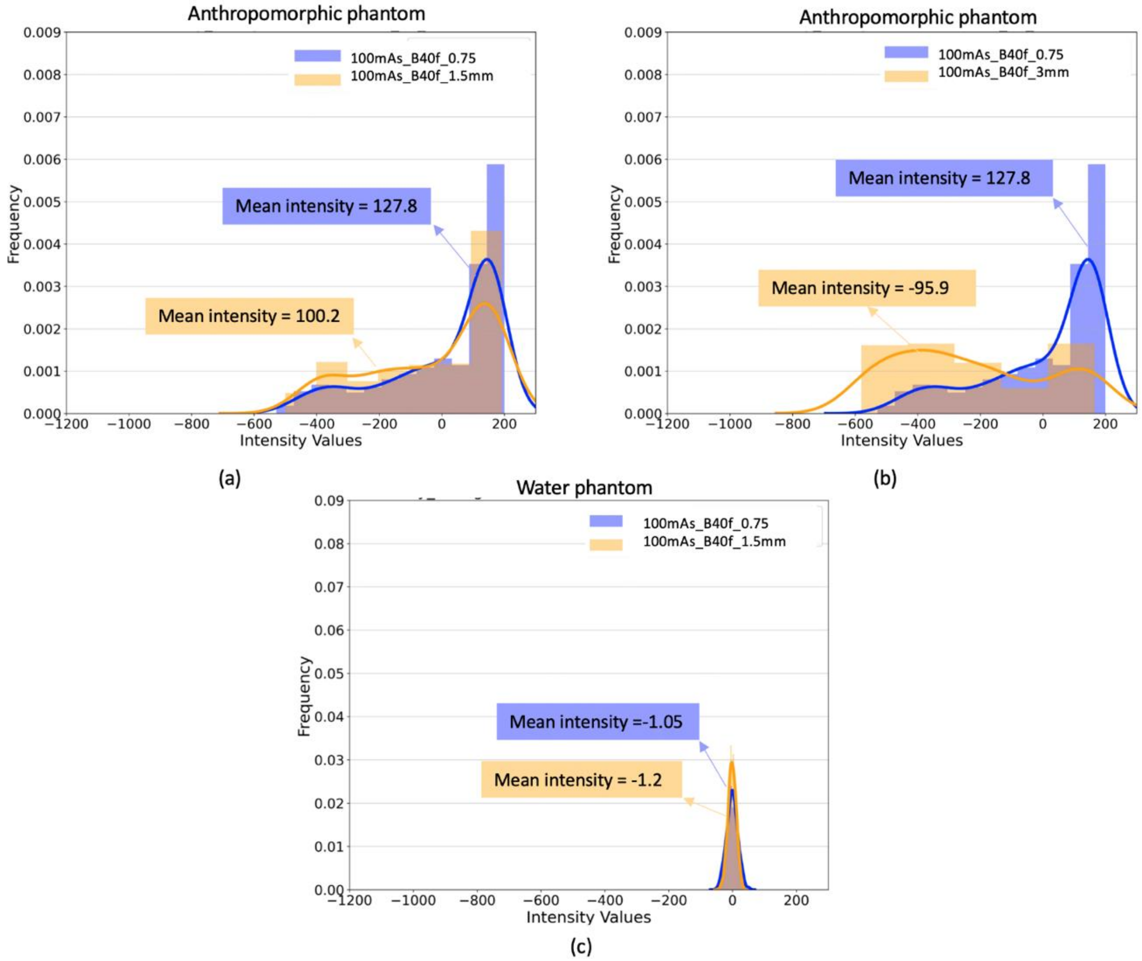


Figure 5-7.Variation of intensity histogram and mean intensity of VOIs due to variation of slice thickness (at constant kernel and dose) in anthropomorphic phantom (a and b) and water phantom (c)

Figure 5-8 shows the histogram of the ratios between reference and non-reference protocol for all the 82 radiomic feature values in both anthropomorphic phantom and water phantom. A histogram centered at 1 with a low standard deviation has radiomic feature values with ratios very close to 1. This will be indicative of low variability of radiomic feature values between reference and non-reference protocols. According to Figure 5-8, in anthropomorphic phantom and at the thicker slice thickness (3mm), the ratios between the reference and non-reference radiomic feature values are more noticeable compared to thinner slice thickness (both the mean (μ) and standard deviation (σ) of the histogram are larger). When comparing the 1.5mm protocol with the 0.75mm protocol, the radiomic feature ratios are higher for anthropomorphic phantom than the ratios for water phantom.

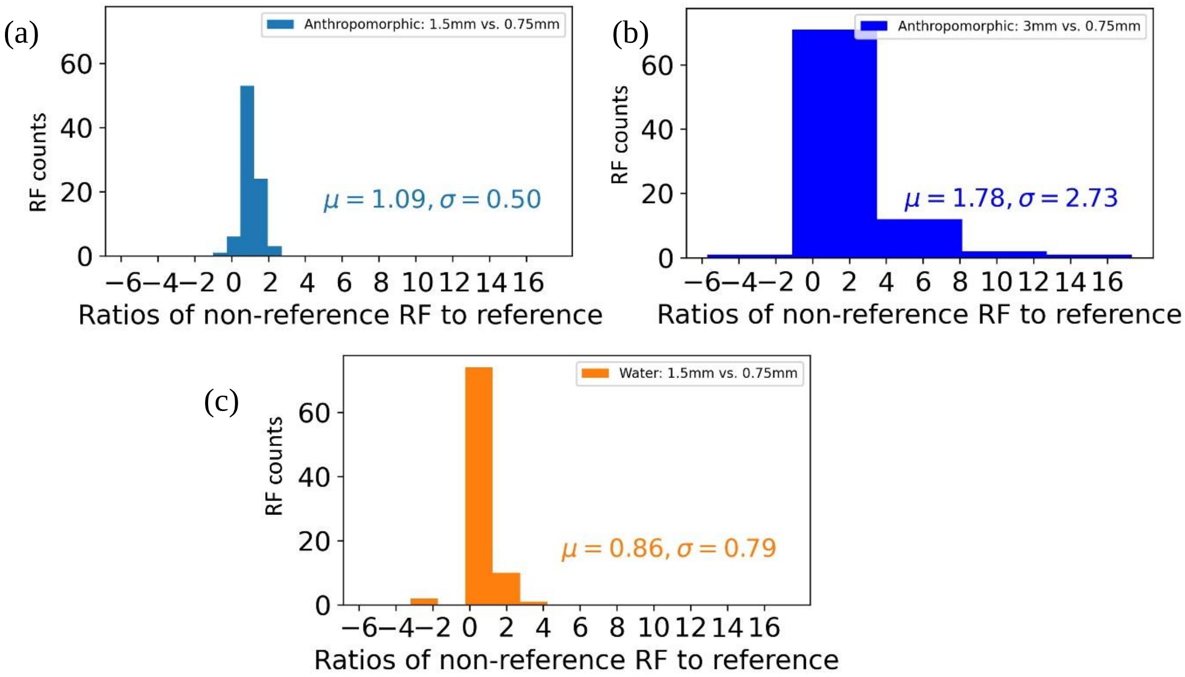


Figure 5-8. Distribution of the ratios between reference and non-reference protocols for 82 radiomic features in response to thickness variation at constant kernel and dose. The mean (μ) and standard deviation (σ) of each histogram is shown. (a and b) from synthetic nodule in the anthropomorphic phantom. (c) from VOI over water phantom

In Figure 5-8(a) Figure 5-8(b), a few features have negative ratios. The ratios for *kurtosis*[4] and the *10th percentile* features are negative in Figure 5-8(a). In Figure 5-8(b), the ratio for the *10th percentile* feature is negative. This is because the kurtosis feature is negative at the 3mm image of the synthetic lesion. Additionally, the *10th percentile* feature of the synthetic lesion is negative at the 1.5mm and the 0.75mm images. The negativity of the *10th percentile* for thicker slice thickness is due to the shift of pixel intensity values to negative values (as was shown in Figure 5-7). The negativity of the kurtosis of the lesion is also visual in Figure 5-7(a) and Figure 5-7(b); the distribution of pixel values has a smoother peak than the normal distribution and is flatter than the reference distribution.

[4] The *kurtosis* feature of VOI is calculated by $\frac{\mu_4}{\sigma_4} - 3$

The *10th percentile* feature for water phantom VOI is negative at both slice thicknesses, so its ratio is positive. But the ratio for *skewness* feature is negative in the water phantom VOI as the skewness for the VOI at the 1.5mm was slightly negative ($skewness_{1.5mm=-0.12}$), indicating wider (or less symmetric) intensity distribution for thicker slice thickness (1.5mm) compared to thinner slice (0.75mm).

5.3.1.2 Slice Thickness Variations: Comparison of CCR and Water Phantom

Figure 5-9 shows the images of the VOI over the water phantom visualized in lung and soft window settings. In the CCR phantom, the VOI is surrounded by a background with similar average intensity and texture to the VOI (Figure 5-9). No substantial difference is visualized between images of CCR phantom at soft tissue window when their slice thickness changes.

Figure 5-10 shows the radiomic feature ratios between reference and non-reference protocol when slice thickness has changed under constant kernel and dose level in CCR and water phantom images (protocols shown in Table 5-2). Unlike the anthropomorphic phantom, the shown radiomic features have ratios <1. This indicates that the radiomic feature has smaller values at non-reference protocol (with thicker slice thickness) than reference protocol (with thinner slice thickness). First-order features, such as *stddev*, *IQR*, *skewness*, and texture features such as *GLRLM GrayLevelNonUniformity*, *GLSZM ZoneVariance*, *etc.*, are among the features with ratios <1. These features describe the spatial and non-spatial deviations and heterogeneity of intensity values within the VOI. This result indicates the reduction of variations and heterogeneity and the increase in the image's smoothness with thicker slice thickness.

The ratios for both phantoms were close to each other. So, the differences between the radiomic feature variation in water phantom versus CCR phantom were not substantial. Figure 5-11 shows the histogram of ratios for non-reference to reference radiomic feature values. The average radiomic feature ratios for CCR and Water phantoms are similar and are close to 1.

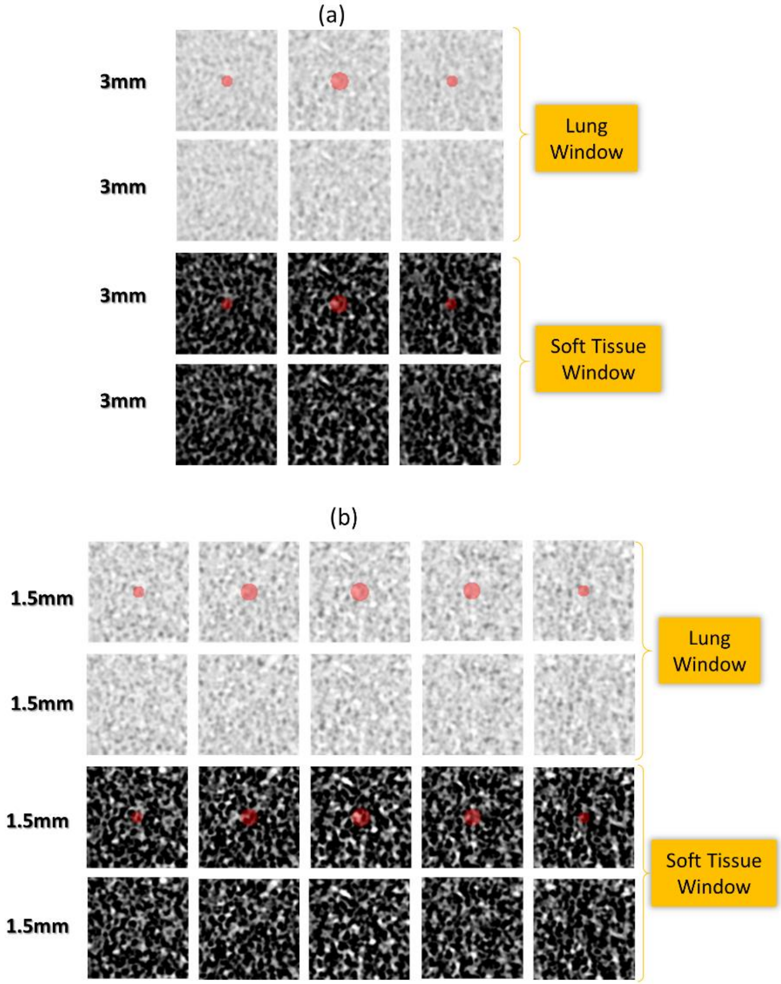


Figure 5-9. Images of the VOI over CCR phantom visualized in lung window setting (level: -600 HU, width: 1600 HU) and soft tissue window setting (level: 50 HU, width: 350 HU) with and without the overlay of the VOI (the region in red) for scans with different slice thicknesses. (a) B31f, 250mAs, 3mm, (b) B31f, 250mAs,1.5mm.

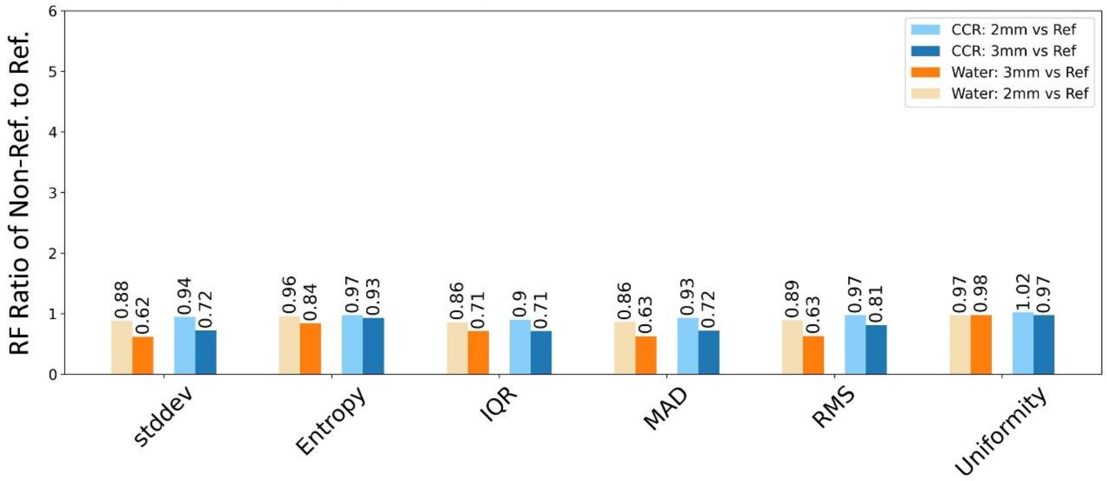


Figure 5-10. Impact of slice thickness variations on a group of first order radiomic features. radiomic feature variability is shown in terms of the ratio of radiomic feature value at non-reference protocol (with B31f, 250mAs) to radiomic feature value at reference protocol (1.5mm, B31f, 250mAs).

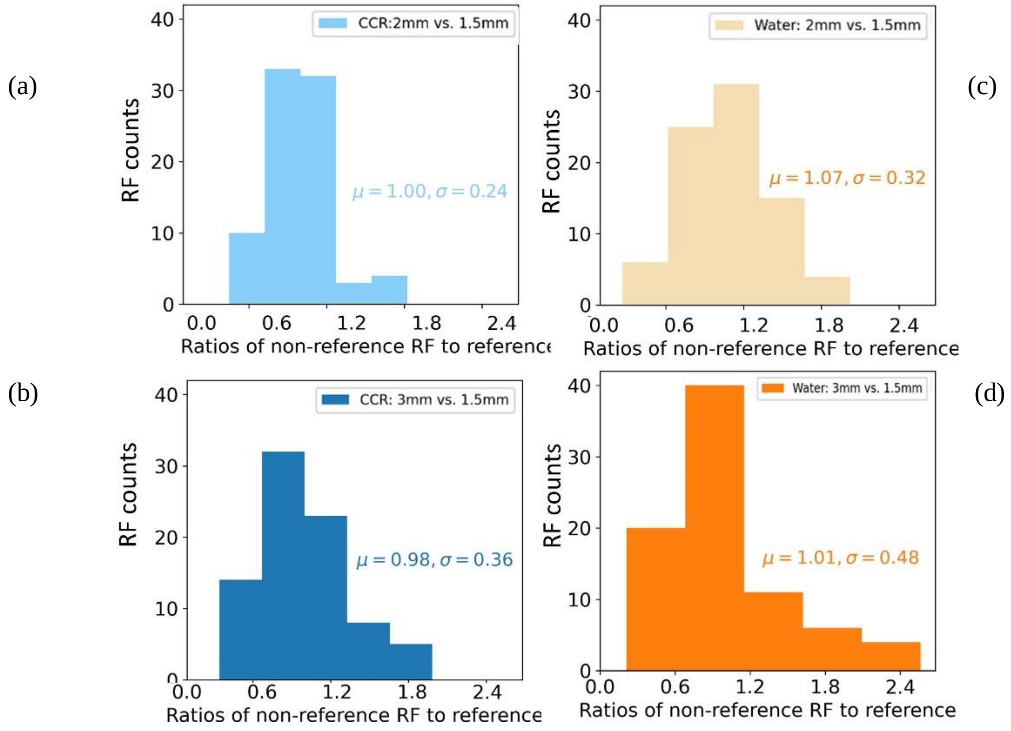
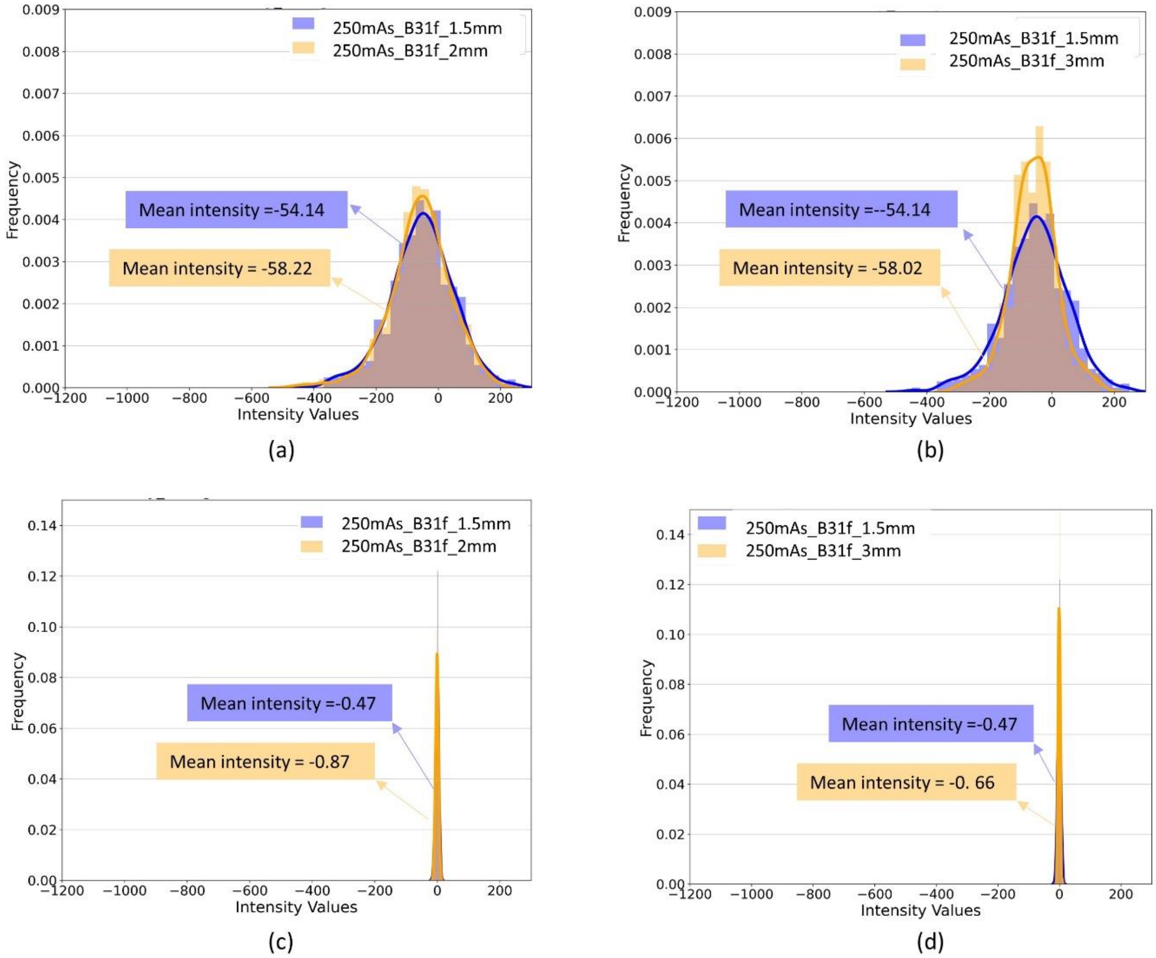


Figure 5-11. Distribution of the ratios between reference and non-reference protocols for 82 radiomic features in response to thickness variation at constant kernel and dose. (a and b) from VOI in CCR phantom. (c and d) VOI in the water phantom mean (μ) and standard deviation (σ) of each histogram is shown.

Unlike what was seen for the anthropomorphic phantom, no variations are observed in intensity values for the VOIs in CCR phantom due to variation of slice thickness (Figure 5-12 (a and b)).

Additionally, Figure 5-12(b) for CCR and Figure 5-7(c) for water clearly show that, unlike anthropomorphic phantom, the intensity histogram is narrower at thicker slice thickness. Since the thicker slice thickness involves a signal from a higher number of photon counts, there will be a reduction in standard deviation. The reduction of standard deviation is more for water phantom as it consists of homogenous material, but for CCR, the standard deviation is higher as it consists of heterogeneous material.



(a)

(b)

(c)

(d)

Figure 5-12. Variation of intensity histogram and mean intensity of VOIs due to variation of slice thickness (at constant kernel and dose) in CCR phantom (a-b) and water phantom (c-d).

5.3.2 Radiomic Feature Variations due to Variation of Dose Level

5.3.2.1 Dose Variations: Water-CCR Phantom Comparisons

Figure 5-13 demonstrates the variability of radiomic feature induced by variation of dose level both for water phantom and CCR phantom images at consistent protocols (protocols shown in Table 5-1). In water phantom, when kernel and slice thickness were constant, but the dose reduced by a factor of 6, features such as stddev, IQR, MAD, and RMS increased to twice the feature value at the reference protocol (300 mAs). However, in the CCR phantom, the values for these features at non-reference protocol stayed similar to the feature value at the reference protocol even after dose reduction up to 50 mAs (ratio ~ 1 even with the dose reduction of a factor of 6). In response to dose variations (and at constant slice thickness and kernel), 14 first-order features (including mean intensity feature) and 39 texture features had higher water phantom ratios than CCR phantom. Among the highly variable radiomic features were some features that 1) describe intensity distribution, such as *Skewness*, S*tandard Deviation*, etc. 2) describe randomness and spatial nonuniformity of intensity and texture, such as *NGTDM Contrast*, *GLSZM Size Zone Non-Uniformity*, *NGTDM Busyness, etc. 3)* that measure the magnitude of voxel values, such as energy. Figure 5-14 shows the histogram of the ratios of non-reference to reference protocols for all radiomic features due to variation of dose. In Figure 5-14 average ratios for both CCR and water phantom are close to 1, except that for water phantom, at lower dose level (50mAs), the average and standard deviation of the histogram is slightly higher.

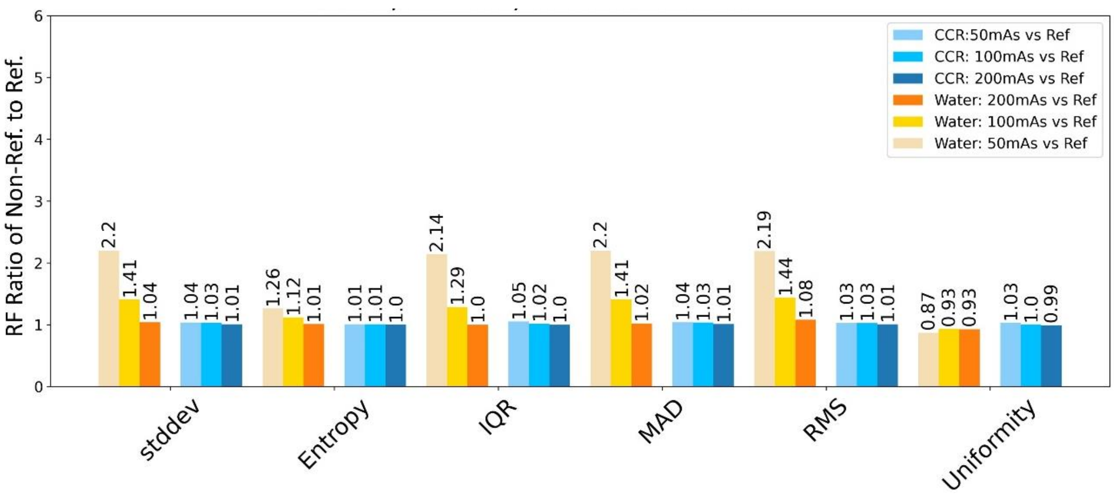


Figure 5-13.Impact of dose variations on a set of first-order radiomic features. radiomic feature variability is shown in terms of the ratio of radiomic feature value at non-reference protocol (at B31f, 1.5mm) to radiomic feature value at reference protocol (300mAS, B31f, 1.5mm)

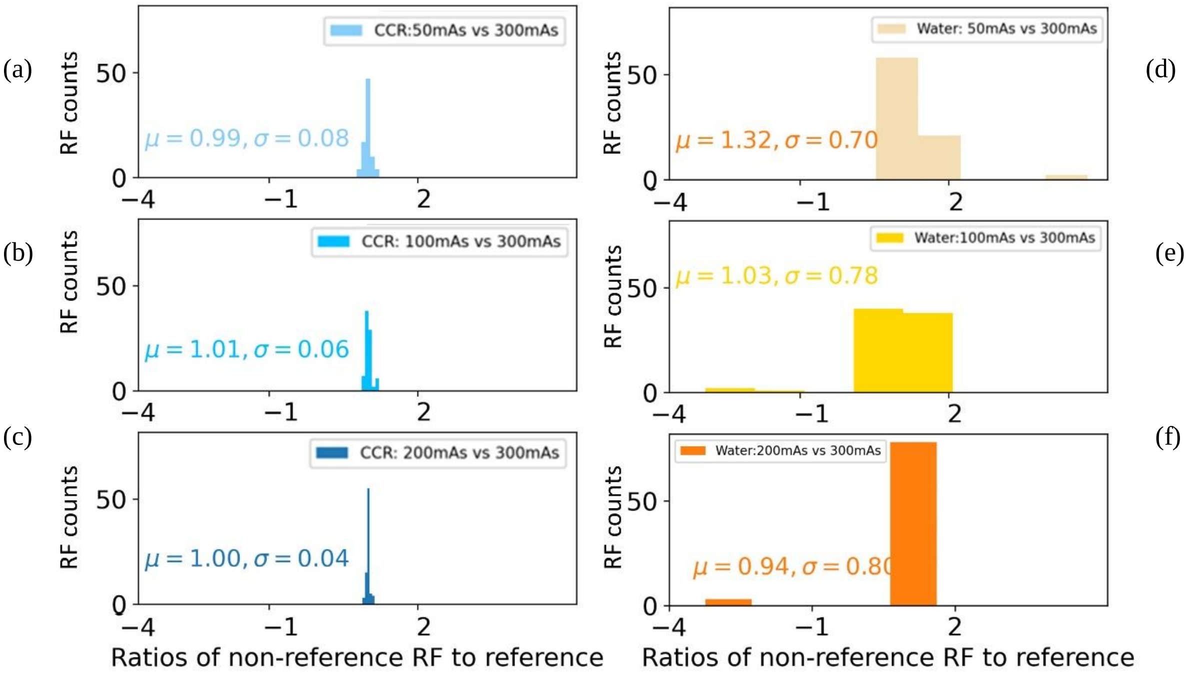


Figure 5-14. Distribution of the ratios between reference and non-reference protocols for 82 radiomic features in response to dose variation at constant kernel and thickness. (a-c) from VOI in CCR phantom. (d-f) from VOI in the water phantom. The mean (μ) and standard deviation (σ) of each histogram are shown.

5.3.2.2 Dose Variations: Water-Anthropomorphic Phantom Comparisons

Figure 5-15 shows ratios of non-reference to reference radiomic feature in response to variation of dose for anthropomorphic and water phantom (for protocols shown in Table 5-2). Interestingly, the displayed radiomic feature values had higher variability (higher ratios) in the water phantom for dose variations. In total, in response to dose variation, 48 radiomic features (out of a total of 82 radiomic features) had larger ratios in water phantom than the anthropomorphic phantom.

Figure 5-16 shows the histogram of the ratios of non-reference to reference radiomic feature for variation of dose in the anthropomorphic and the water phantom. The average ratios for the anthropomorphic phantom are very close to 1. However, in the water phantom, the ratios have a higher deviation from 1, indicating more considerable radiomic feature variability in the water phantom due to variation of dose.

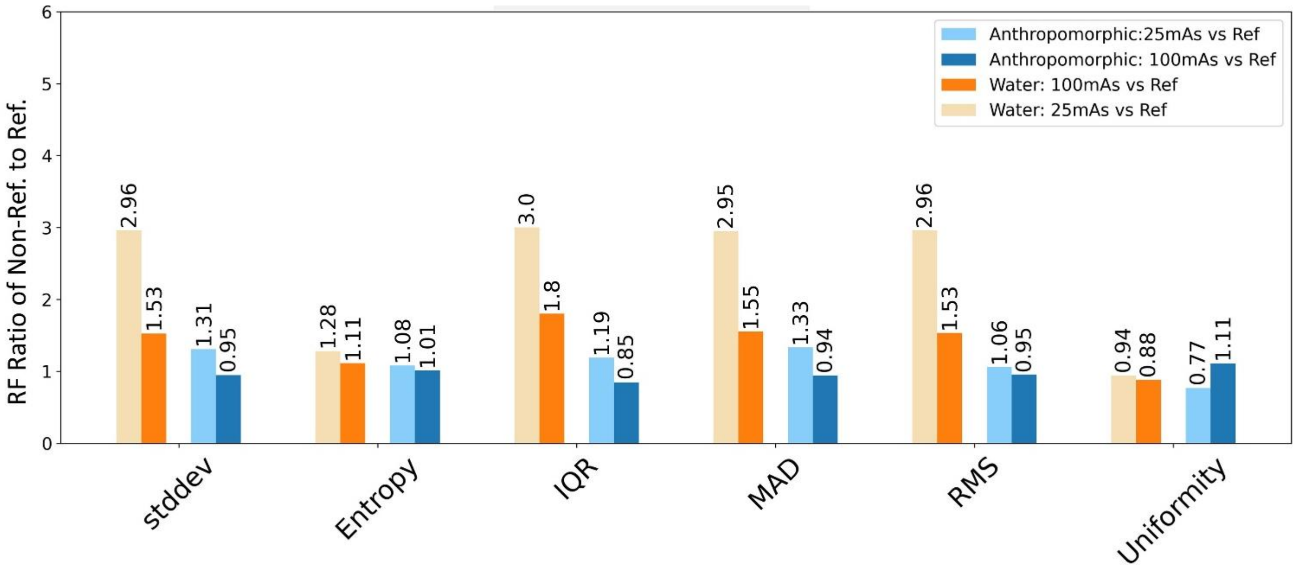


Figure 5-15. Impact of dose variations on a set of radiomic features calculated from water phantom and anthropomorphic phantom. The ratio of radiomic feature value at the non-reference protocol to radiomic feature value at reference protocol (200 mAs, B40f, 1.5mm) is shown.

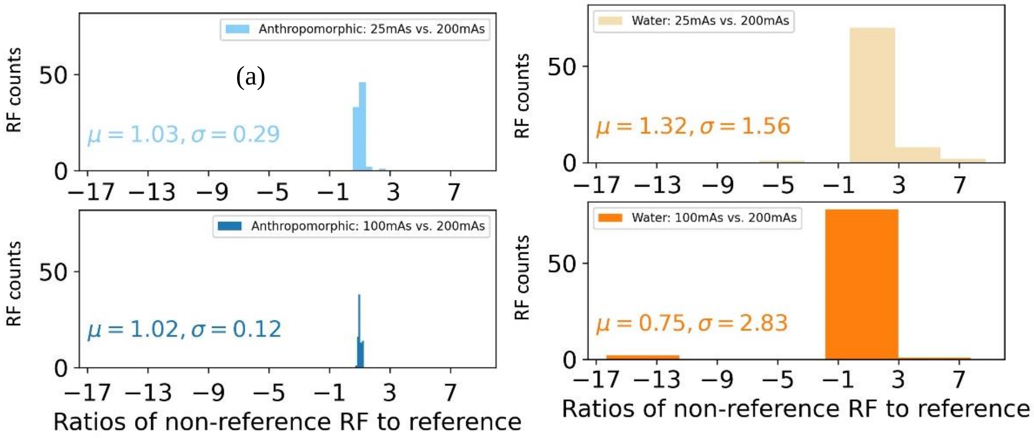


Figure 5-16. Distribution of the ratios between reference and non-reference protocols for 82 radiomic features in response to dose variation at constant kernel and thickness. (a-b) from synthetic nodule in the anthropomorphic phantom. (c-d) from VOI in the water phantom. The mean (μ) and standard deviation (σ) of each histogram is shown.

5.3.3 Radiomic Feature Variation Due to Variation of Kernel

5.3.3.1 Kernel Variations: Water-CCR Phantom Comparisons

For the radiomic features shown in Figure 5-17, there is more variability in the water phantom VOI in response to kernel variations. For example, 17 texture features and 18 first-order features had larger water phantom ratios than CCR phantom. The same first-order features impacted by dose variation in the water phantom were also affected by kernel variations. The mean intensity feature was not affected by kernel variations in either of the two phantoms.

In comparing the impact of kernel variations with the effects of dose variations in these paired comparisons, we observe that in Figure 5-17, kernel variation resulted in more considerable variability for the shown radiomic feature compared to dose variations shown in Figure 5-13 (*stddev, IQR, MAD, RMS* have ratios of more than 10 when the non-reference protocol is compared to the reference protocol).

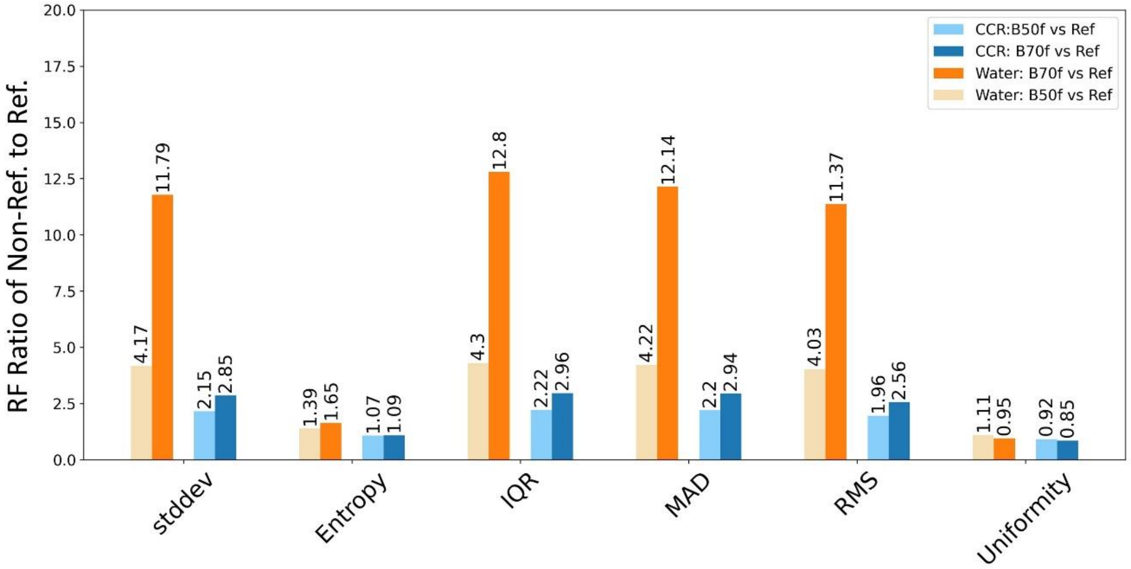


Figure 5-17. Impact of reconstruction kernel variations on a set of first-order radiomic features. radiomic feature variability is shown in terms of the ratio of radiomic feature value at non-reference protocol (65mAs, 1.5mm) to radiomic feature value at reference protocol (B20f, 65mAs, 1.5mm)

Figure 5-18 shows the histogram of the ratios of non-reference to reference protocols for all radiomic features due to variation of the kernel. In Figure 5-18, for the CCR phantom, the average radiomic feature ratios are very close to 1. However, the average ratios are higher for water phantom, especially for the ratios between B70f and B20f protocols. In addition, the intensity variance and energy features had substantial ratios (>100) between B70f and B20f protocol. These results indicate substantially more considerable variability of radiomic features in water phantom with homogenous material compared to the heterogeneous material of CCR phantom.

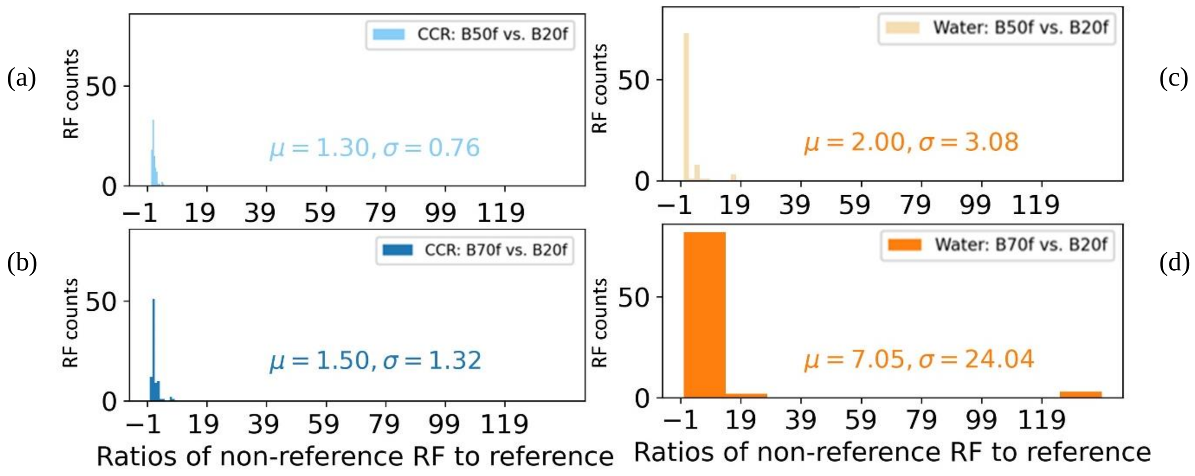


Figure 5-18. Distribution of the ratios between reference and non-reference protocols for 82 radiomic features in response to kernel variation at constant thickness and dose. (a-b) from VOI in CCR phantom. (c-d) from VOI in the water phantom. The mean (μ) and standard deviation (σ) of each histogram are shown.

5.3.3.2 Kernel Variations: Water-Anthropomorphic Phantom Comparisons

Figure 5-19 shows ratios of non-reference to reference radiomic feature in response to variation of the kernel for anthropomorphic and water phantom (for protocols shown in Table 5-2). The illustrated radiomic features had higher variability in the water phantom for kernel variations. In total, in response to kernel variation, 53 radiomic features (out of a total of 82 radiomic features) had larger ratios in the water phantom than the anthropomorphic phantom.

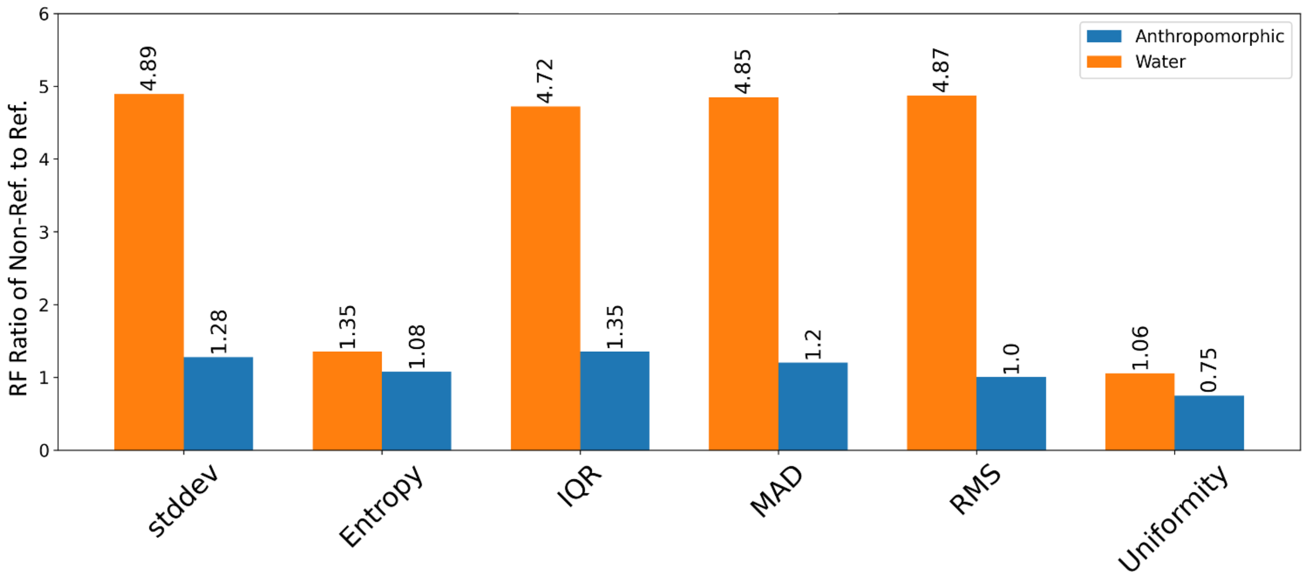
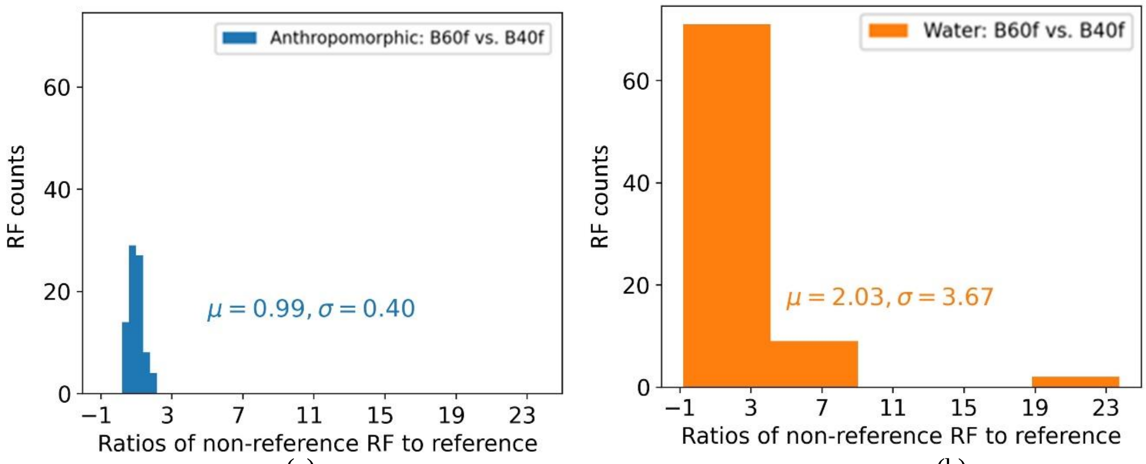


Figure 5-19. Impact of kernel variations on a set of radiomic features calculated from water phantom and anthropomorphic phantom. The ratio of radiomic feature value at non-reference protocol (100 mAs, B60f, 1.5mm) to radiomic feature value at reference protocol (100 mAs, B40f, 1.5mm) is shown.

Figure 5-20 shows the histogram of the ratios of non-reference to reference radiomic feature for variation of the kernel in the anthropomorphic and the water phantom. The average ratios for the anthropomorphic phantom are very close to 1. However, in the water phantom, the ratios have a higher deviation from 1, which indicates larger radiomic feature variability in the water phantom due to the kernel variation.



(a)

(b)

Figure 5-20. Distribution of the ratios between reference and non-reference protocols for 82 radiomic features in response to kernel variation at constant dose and thickness. (a) from synthetic nodule in the anthropomorphic phantom. (b) from VOI in the water phantom. The mean (μ) and standard deviation (σ) of each histogram is shown.

5.4 Discussion

5.4.1 summary:

Implementation of quantitative medical imaging techniques in the clinic and enabling the application of radiomics in building decision support tools that assist physicians in diagnosis or prognosis of a disease requires an understanding of limitations and sources of variability of this technique[20]. Recent research has indicated the sensitivity of CT radiomic features to variation of CT image protocols. Further investigation into the underlying sources of variability unravels the involved mechanisms, and this can inform us about potential solutions or proper data inclusion strategies in future radiomic studies.

In this study, we examined the impact of variation of CT scan protocol on radiomic features that are often used to describe pathologies such as lung nodule characteristics. The purpose was to understand how the radiomic features are influenced by slice thickness, dose, and kernel. Table 5-3 summarizes the findings of this study regarding the impact of each of the three investigated CT parameters on the radiomic features of the three phantom datasets.

Table 5-3. Summary of the impact of variation of CT parameters on radiomic features of the three phantoms

Parameter	**Homogeneous Water Phantom**	**Heterogeneous Texture Phantom (CCR)**	**Anthropomorphic Phantom with homogeneous synthetic lesion**
Slice thickness	Little to no effect	Little to no effect	Large effect
Dose	Medium	Little to no effect	Little to no effect
Kernel	Large effect	Medium effect	Medium effect

This was done by assessing the variation of three main CT parameters of radiation dose level, reconstruction kernel, and slice thickness in CT scans of phantoms with different complexity and distinct characteristics. Three phantom image datasets of a textured phantom (CCR phantom[24]), an anthropomorphic phantom[29], and a water phantom were used in this study. It is clear that phantom data may not be a perfect surrogate for patient data. While in our prior study, we have already assessed radiomic feature reproducibility directly in patient in this study, we decided to use phantom data as each phantom data had similarities to specific characteristics of lung nodule CT scans, and knowing a gold standard for intensity values, size or having a reference textured phantom helped in interpretation of radiomic feature variability in response to variation of CT parameters. Another key contribution of this study is that, by comparing and contrasting these phantom datasets, we understand the extent to which the results of phantom studies in reproducibility of radiomic feature or the harmonization of radiomic feature variability can be generalized to patient datasets. Currently, no studies compare the three different types of phantom data used in this study in terms of variability of radiomic feature that describe density characteristics and texture of tissue in response to variation of CT image acquisition. Notably, the public image datasets for two of the phantoms used (i.e., CCR and anthropomorphic phantom) are acquired on disparate CT protocols in a way that consistent comparisons between the two phantoms were not possible. Hence, we enabled a parallel comparison between these phantoms by scanning a third phantom (i.e., water phantom). By making comparisons at matching set of protocols for each of the two phantoms and controlling for CT parameters while only one parameter in question varied, we could better discern the relationship between the radiomic feature and the varied parameter.

In this study, slice thickness variation in images of a synthetic nodule in anthropomorphic phantom showed the most considerable radiomic feature variability compared to water phantom or the CCR phantom. In the anthropomorphic phantom, at the thicker slice thickness, the first-order features that describe deviations and randomness of gray level intensity values increased (Figure 5-4), voxel intensity values were shifted toward negative values (Figure 5-7), and the average ratios of non-reference to reference protocol deviated largely from 1 (Figure 5-8) indicating of dissimilarity of radiomic feature between reference and non-reference protocol when all CT parameters were fixed, and only the image slice thickness changed. According to Figure 5-8 and Figure 5-11, within each phantom, the ratios of features have a wider range when the reference and the non-reference protocol have a wider deviation in terms of slice thickness (e.g., in comparison of the distribution of CCR ratios of 2mm vs. 1.5mm to 3mm vs. 1.5mm).

These results were consistent with findings of similar phantom studies. Shafiq-Ul-Hassan *et al.* [79] studied CCR phantom data, which overlaps with the data used in this study, found 65 features to be robust to voxel size variation. Larue *et al.* [85], found the first-order *Energy and GLRLM Run Length Non-Uniformity* feature in CCR phantom being highly impacted by slice thickness variation; likewise, our results showed that at thicker slice thickness, these radiomic features decreased by 70% and 50%, respectively, when comparing thicker slice to thinner slice. The study by Zhao *et al.* [30] in the analysis of 14 radiomic features, which overlaps with our set of radiomic features, in the same anthropomorphic phantom indicated the lack of reproducibility against variation of slice thickness. It showed that with an increase in the difference between image slice thicknesses, the larger the deviations of radiomic feature values.

Additionally, radiomic feature variations observed due to slice thickness variation in the synthetic nodule are consistent with our prior studies[78] and other patient studies in literature[86,87]. Kim *et al.* [31] also assessed the variation of slice thickness in scans of the anthropomorphic phantom but within a different range of slice thicknesses than this study. Though their findings agree with our results, and their conclusions are complementary and consistent with this study.

Furthermore, previous studies[88–91] have assessed the impact of slice thickness on the variation of volume, surface area, or diameter of nodules in phantom images or patient scans and have indicated a significant or substantial impact of thickness on these measurements. Though, this study has extended these investigations by assessing a larger group of quantitative metrics (radiomic feature) that can describe various tumor characteristics such as density and texture.

As a summary, through comparison of findings from slice thickness variation in anthropomorphic phantom, water phantom, and CCR phantom along with visualizations of VOIs in Figure 5-5, Figure 5-6 and Figure 5-9, and the agreement of results with patient studies, it appears that the mechanism of volume averaging that occurs in the reconstruction of the image at thicker slice thickness creates a more considerable impact on radiomic feature of a nodule object. Therefore, the anthropomorphic phantom had the most similarity to patient data in terms of slice thickness variation.

It is worthwhile to note that while a substantial impact can be seen for lung nodules due to variation of slice thickness, this may not be the same for the case of tumors in other organs. For example, CT radiomic features of liver tumors can behave differently against variation of slice thickness as the distinction of a liver tumor with respect to its background is not as substantial as lung nodules. Understanding and considering these disparities between phantoms and patients and even for different tumor types are essential when assessing the reproducibility of radiomic features in response to slice thickness variation. One important consideration is the role of the segmentation algorithm; the appearance or disappearance of the segmentation mask in the soft tissue window of Figure *5-5* depends upon the thresholds and the specifics defined within the segmentation algorithm. Hence there can be an interaction between the segmentation technique and variability of radiomic feature. For example, a multicenter study[19], in their investigation of the reproducibility of radiomics, determined that differences in segmentation algorithms are an important factor in the reproducibility of radiomic features. Also, Hoye *et al*. [33] highlighted the interaction of the performance of segmentation algorithms in the quantification of radiomic features in their study. Additionally, this study only assessed spherical nodules. Nodules with irregular shapes may be impacted differently by the variation of slice thickness. Therefore, the radiomic features of irregular-shaped nodules may behave differently in response to variation of slice thickness.

In exploring differences between radiomic feature variability in homogenous material versus heterogeneous material, we found that kernel and dose variations can have different impacts on radiomic feature measured from these different materials. It seems that when varying kernel and dose in scans of a heterogeneous material (e.g., in CCR), the inherent heterogeneity contributed more to the measured standard deviation ($stddev$) or the quantified textures via radiomic feature than the variations induced by the CT parameter. For example, in Figure 5-13, many of the first-order radiomic features shown are constant (non-reference to reference ratio ≅ 1) in CCR but have larger ratios in the water phantom. Similarly, as shown in Figure 5-17, the feature *stddev* had larger ratios (and therefore more variability) when kernel changed in water than in CCR. Considering that the *stddev* feature in VOI accumulates the standard deviations due to the imaging protocol (σ_p) and the deviations due to heterogeneity of the material (σ_m) ($stddev = \sqrt{\sigma_p{}^2 + \sigma_m{}^2}$), when the scanned material is textured (such as in CCR), $\sigma_m{}^2$ can be much higher than deviations caused by imaging protocol variation ($\sigma_m{}^2{"}\,\sigma_p{}^2$). Therefore, variation of image acquisition or reconstruction does not impact *stddev* value substantially in CCR. This can explain the different behavior of other radiomic feature values in CCR versus water phantom. According to Figure 5-18, the average ratios of 82 radiomic feature between non-reference and reference protocols for CCR is very close to 1 as opposed to water phantom's ratios that deviate from 1, indicating more variability in the water phantom for many of the calculated radiomic features.

When dose and kernel variations were examined in anthropomorphic phantom and were compared to water phantom (Figure 5-15 and Figure 5-19), we saw a similar trend. The density histogram of the synthetic nodule (Figure 5-7) showed that the nodule's pixel values have more heterogeneity and deviations than the water phantom. Thus, this can be why the kernel and dose variations had resulted in less impact in radiomic features that were calculated from synthetic nodule compared to the water VOI.

The observed disparities between radiomic feature variability in homogenous versus heterogeneous material get important when considering the possible differences between patient tumor tissues. These findings suggest that tumor tissues with minimal texture may be more impacted by noise variations in the image than a textured tumor.

Within each phantom, reconstruction at sharper kernel resulted in more variability of non-reference radiomic feature than a smoother kernel. Furthermore, dose variations did not result in substantial variability in the CCR and the anthropomorphic phantoms. The average of the ratios between non-reference and reference protocols was very close to 1 when dose changed, and other CT parameters were kept constant. However, the radiomic features of the water phantom were slightly impacted at the lowest dose levels.

Similar findings have been previously reported in the literature, where dose variation resulting from changing the tube current has not impacted radiomic features[66]. Furthermore, other studies have identified reconstruction kernel as an important factor affecting radiomic feature values in phantom studies[92] and patient studies[27,32,40] .

The findings of this study suggest that, while usage of phantoms can be beneficial in evaluating the robustness of a quantitative imaging technique (such as radiomics or tumor segmentation, etc.), we should take into account the intrinsic differences of different phantom datasets with each other and with patient data. Relying on one type of phantom image data can result in deviations from the actual observations in patient data.

The scope of this study was limited as we aimed only to address the impact of variation of three CT parameters on radiomic feature to understand differences in the radiomic feature reproducibility in the presence of volume averaging and in the presence of heterogeneous/homogenous material. Clearly, the radiomics technique faces a couple of other sources of variability[93] that was not explored in this study. For example, it has been previously reported that usage of different methods or definitions for radiomic feature calculation can result in discrepancies[19,35]. Therefore, it will be most optimal if features are compared according to a reference manual such as Image Biomarker Standardization Initiative guidelines [17]. This initiative offers a reference of feature definitions that can be used to benchmark radiomic features and provides recommendations for reporting pre-processing methods. Another potential source of variation is inter-scanner variability (between different manufacturers or between different scanner models of a manufacturer). This has been explored in a study[94] that assessed variability of radiomics features of the CCR phantom across multiple clinics and scanners using chest and head scan protocols. This study found existing variability of radiomic features between different scanners and sites compared to inter-patient variabilities. Our study was limited to CT scans from one manufacturer (Siemens). However, concerning the existing variations among scanners as reported in literature[24], it will also be worthwhile to compare phantom datasets using different scanners.

For future studies, it may be helpful also to obtain images of a synthetic nodule with heterogeneous and known texture inserted into the anthropomorphic phantom to conduct similar investigations and compare with CCR or patient data. Robins *et al.* [92] has performed a study in line with this idea.

The authors assessed radiomic feature variability within a computational phantom with anatomically informed texture. Another potential future direction for this work is understanding the interaction of reproducibility of radiomic features and their diagnostic performance.

5.5 Conclusion

The current study compares the reproducibility of radiomic features in CT scans of phantom datasets with different complexity to obtain insight into the role of specific characteristics of lung nodules in the variability of radiomic feature due to variation of CT parameters of slice thickness, kernel, and dose level.

Slice thickness variations can impact a wide range of radiomic features (first-order and second-order texture features) due to the variations in photon statistics and the volume averaging phenomenon, especially in the synthetic lesion of the anthropomorphic phantom. We found that the variability of radiomic features can rely on the inherent heterogeneity/homogeneity of the material. Tube current (dose) variations did not substantially impact radiomic features in the CCR and the anthropomorphic phantom. However, the reconstruction kernel had a substantial impact on radiomic features. More considerable variability of radiomic features was observed for heterogeneous materials.

Findings regarding the variability of radiomic feature due to slice thickness variation in synthetic nodule of the anthropomorphic phantom were similar to results of patient studies. In addition, the trends seen for impacts of kernel and dose on radiomic feature from all three phantoms were also consistent with patient studies.

While it may be easier to acquire phantom data than acquiring patient data to perform uncertainty analysis or evaluate the performance of quantitative imaging techniques, differences of phantom characteristics and how these characteristics interact with the variability of quantitative metrics should be taken into account. This study can be helpful for interpretation of such discrepancies in response to variation of CT parameters of slice thickness, kernel, and dose.

Section II Mitigating Radiomic Feature Variations

A.1 Background and Problem

Radiomics is receiving increasing attention in medical imaging research, and several studies have shown the promise of radiomics in the development of diagnostic imaging biomarkers. On the other hand, in Chapter 4, we found that the variation of CT image acquisition and reconstruction parameters of dose (as a function tube current), reconstruction kernel, and slice thickness impact most radiomic features of patient lung nodules. In addition, other studies have reported variability of radiomic features due to other image acquisition parameters or image analysis such as tube voltage[41], segmentation algorithm[33], feature definitions[19,35], streak or beam-hardening artifacts[95], etc.

When radiomic features get influenced by non-biological processes such as image artifacts or variation of image quality due to variation of image acquisition protocols, their ability to provide accurate information about tumor phenotype may be negatively impacted. For example, in PET/CT and MR imaging studies, radiomic features affected by differences in image acquisitions between different centers had poor performance when radiomic feature variations were not taken into account and were not corrected[96,97].

This issue raises reliability concerns in using radiomic features as a diagnostic or prognostic biomarker in practice. Therefore, for incorporation of radiomic features into radiomic studies or clinical trials, standardization procedures are required to control and deal with potential sources of variability (i.e., image acquisition, image analysis, radiomic feature calculation, etc.).

A.2 Motivation and Purpose:

To ensure maximum information gain from radiomic features, we have to ensure that radiomic signatures or radiomic-based prediction models perform robustly against variation of non-biological factors. Therefore, it is crucial to consider the variabilities as mentioned earlier and take standardization steps before model building or before investigating associations between biological properties and radiomic features.

By including an effective standardization or harmonization step, radiomic features will become comparable across different image acquisitions. Therefore, a pre-trained model may generalize well to harmonized radiomic features acquired with acquisition protocols different from the training set. Similarly, in a multiple-time point study, even if the baseline and follow-up scans of a patient are acquired with various protocols, after a harmonization step, the radiomic features can reflect the deviations that are genuinely due to the variation of the biological properties.

Currently, only a limited number of studies have explored the mitigation of radiomic features. On the other hand, only a few studies[97–99] have included a harmonization procedure in building prediction models. Therefore, the exploration of mitigation techniques for radiomic features has remained an open question that deserves further evaluation and, perhaps, the establishment of novel methods.

In specific aim 3 (SA3) of this dissertation, we aimed to take steps toward testing strategies that can improve the reliability of radiomic features in characterizing biological properties of lung nodules. Specifically, we aimed to test whether it is possible to harmonize the variability of radiomic feature values in response to variation of CT acquisition protocols by utilizing mitigation techniques. To this aim, we explored the potential of two different mitigation techniques in the harmonization of radiomic features against variation of the three CT parameters that were shown to have a consequence on radiomic feature values in the SA2 of this dissertation.

One mitigation technique explored is the deep learning approach of Generative Adversarial Networks (GAN)[100] that works in the image domain. The GAN model learns the mapping function between images acquired with different acquisition parameters, applies an image transformation, and outputs images that result in harmonized radiomic features. The other investigated technique in this dissertation is a data-driven technique called ComBat[101] that works in the feature domain rather than the image domain. It treats the impact of variation of image acquisitions (acquisition shift) as an additive and multiplicative effects. It estimates these effects using Bayes estimation and adjusts the radiomic feature values by removing the estimated effects from data to acquire a harmonized radiomic feature data with a reduced impact of non-biological factors of CT parameters. Figure 21 summarizes how the two methods operate.

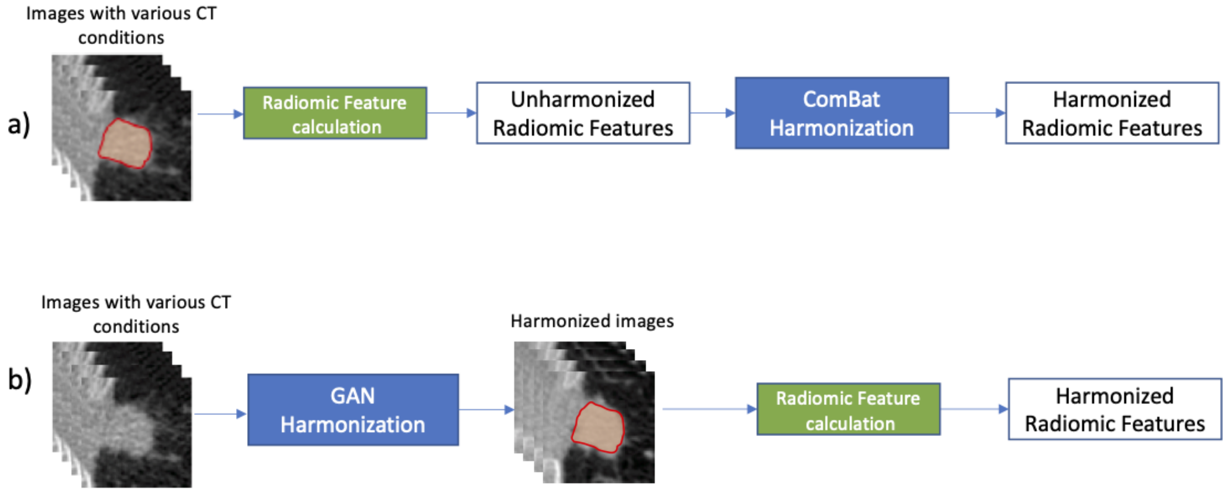

Images with various CT conditions
a)
Radiomic Feature calculation
Unharmonized Radiomic Features
ComBat Harmonization
Harmonized Radiomic Features
Images with various CT conditions
b)
GAN Harmonization
Harmonized images
Radiomic Feature calculation
Harmonized Radiomic Features

Figure 21. The operation of the two mitigation techniques explored in this dissertation. a) Harmonization in feature domain: ComBat adjusts the radiomic feature values. b) Harmonization in image domain: GAN is applied to the images before radiomic feature calculation

A.3 Outline of the Following Section

Section 2 of this dissertation is organized as follows: we will first provide an overview of existing radiomic feature mitigation studies that explore approaches other than ComBat and GAN. Then, in Chapter 6, we will first provide a literature background for the GAN technique compared to few other deep learning-based methods. Finally, we will also describe the GAN theory, our experiments, and the results from GAN harmonization in Chapter 6.

Chapter 7 will provide an overview of studies that have utilized ComBat in the radiomic feature field, and we will describe ComBat theory, our experiments, and the corresponding results.

In Chapter 8, we will compare and contrast the two techniques of GAN and ComBat by making an analogy between their harmonization performances on the same set of data.

A.4 Literature Review: Standardization of Radiomic Features

As mentioned earlier, while there is a growing number of radiomic studies investigating the association of radiomic features to patient outcomes or endpoints, few studies have explored or identified standardization strategies in tackling the lack of robustness of radiomic features.

One standardization strategy could be controlling for scan protocols across research and clinical institutes. For example, Ger *et al.* [94] identified a controlled CT protocol (120 kVp, 200 mAs, pitch of 1, 50cm field of view) that minimized differences in radiomic feature values of a phantom between scans across different CT scanner vendors in multiple imaging institutes. This study[94] can encourage prospective radiomic studies to restrict acquiring CT images to those dose at a reference scan protocol and therefore enable obtaining comparable radiomic features between different imaging centers. However, it may not always be possible to adapt a controlled protocol across many imaging centers. So, there is still a need for other standardization strategies, especially for retrospective studies. Therefore, in our study, we have focused on strategies that can handle variability issues in retrospective studies or scenarios where it is impossible to obtain multicenter images with a standard protocol.

Shafiq-ul-Hassan *et al.* [27] demonstrated a correction for variation of reconstruction kernels through measuring the noise power spectrum (NPS) of a standard phantom for a scanner. The authors used the measured NPS and maximum intensity to adjust and correct noise texture in the CT image of CCR phantom data and improved feature robustness by 30–78%. However, further investigation is needed to understand the generalizability of this technique in lung nodule radiomic features as well.

Our findings (reported in Chapter 4) and conclusions of other studies[40,87] have determined that slice thickness deviations have a strong impact on radiomic feature values. Previous efforts[79,85] have proposed approaches for harmonization of the effect of slice thickness variation by resampling or interpolating voxels. These two studies found that after resampling images of the CCR phantom[24], the coefficient of variation of radiomic features and concordance correlation coefficient (CCC) improved for many radiomic features. However, in the study by Ger *et al.* [94], image resampling did not help compensate variability of radiomic features in phantom data. Similarly, in our initial experiments on patient images, we did not significantly improve the radiomic feature reproducibility after resampling the images with 2mm and 0.6 mm slice thickness to 1mm thickness (see Appendix B, Table B-4). Shafiq-ul-Hassan *et al.*[102] findings in a separate study on patient lung tumors verified these findings. The deviation between results in patient images versus findings from phantom images can be due to the existing differences in volume averaging between CCR phantom images and lung nodule images, which we explored in Chapter 5 of this dissertation. Moreover, the resampling (down-sampling or up-sampling) of the image causes uncertainty in mapping the region of interest for radiomic feature calculation.

Various denoising approaches widely used to enhance image quality can also be explored to understand whether they have a role in remedying the poor reproducibility of radiomic features. For example, Hoffman *et al.*[103] utilized adaptive bilateral denoising in increasing robustness of emphysema scoring metric (RA-950). However, the RA-950 metric is density-based, so it is different than the second-order radiomic features. As a result, this denoising approach was not influential in correcting the variability of several radiomic features. Among other image denoising techniques, the Iterative reconstruction algorithms (IR) also provide a denoised image reconstruction of CT raw data as opposed to the Weighted Filter Back Projection (wFBP) algorithm. Firstly, one downside of this approach is that it requires raw projection data. Secondly, it is time-consuming to obtain IR reconstructed images. Thirdly, according to our results in Chapter 4, the radiomic features calculated from images with IR were still not reproducible.

Basic methods for image normalization are also among other simple techniques that can be explored to harmonize radiomic features. These methods include histogram normalization concerning a global or a local mean and standard deviation. So, these methods may modify the image content and the radiomic feature values. However, such normalization techniques do not consider the impact of individual CT parameter variations in adjusting the image data or radiomic feature data as they only adjust an overall variation; therefore, these methods have not been shown to be very effective. For example, Foy *et al.* [81] investigated the role of a set of different techniques such as histogram normalization (initially proposed by Loizou *et al.* [104]), Butterworth low-pass filtering (offered by Mackin *et al.* [76]), and ComBat technique in the harmonization of radiomic features from a cadaveric liver CT scan in response to variation of CT acquisition parameters (dose, pitch, thickness, field of view, and kernel). This study found that while the other harmonization techniques reduced the number of unstable radiomic features, the ComBat technique significantly decreased the number of unstable features to zero. An essential distinction of the ComBat method compared to other normalization methods is that it estimates the total variance due to the impact of variation of imaging protocols. As a result, the ComBat method appears to be a compelling tool to be investigated in our research study.

In another approach, Gang *et al.*[105] proposed a model, tested on phantom data, that predicts changes of histogram-based and GLCM based radiomic features as the result of noise or spatial resolution corruption. The predicted variations will be then adjusted through a recovery technique that uses two deconvolution steps to adapt either the additive noise or the blur in an image. While this approach provides a general recovery framework, it does not examine mitigation of a particular CT parameter (e.g., slice thickness). Nonetheless, according to the authors[105], since this approach is directly applied to images, it has the potential to be combined with (or compliment) techniques such as ComBat that work in the feature domain.

Zhovanik *et al.*[106] performed a proof of principle study in which they characterized the dependence of radiomic features to signal-to-noise ratio (SNR) due to variation of mAs (from 30 to 460 mAs). The correction method used an additive correction factor obtained from a regression model fitted to the scan exposure. The technique was developed using a phantom[107] with 17 inserts of different tissue-like properties (Gammex 467 Tissue Characterization phantom, Sun Nuclear, Melbourne, FL). The study[106] assessed the overlap of four radiomic features of NSCLC patients with the distribution of the four radiomic features across 17 different tissue inserts. Unfortunately, it did not specify which tissue insert has the most considerable overlap with the NSCLC radiomic features to understand if the technique is sufficiently efficient for lung tumors. Hence, the method was not explicitly evaluated on radiomic features of lung tumor tissue and can serve as a general-purpose harmonization approach.

A.5 Significance and Contribution

While several studies in the literature have already illustrated the predictive power of radiomics, the topic of mitigating radiomic features has only recently gained attention. Therefore, mitigating radiomic features has not been adequately explored. Furthermore, although the previously mentioned efforts have either proposed or assessed some potential solutions, fewer have been thoroughly validated in clinical patient datasets. Additionally, most current approaches tend to only address the variability of radiomic features in response to one CT parameter. Nevertheless, it may be more practical to have a versatile framework that addresses the lack of reproducibility of radiomic features in response to various sources of variability all in one. Hence, in this dissertation, we conducted experiments in which the two mitigation techniques of ComBat and GAN will be tested to harmonize radiomic feature variability in response to changes of three different CT parameters.

The two techniques of ComBat and GAN have recently received growing attention in the medical imaging research field. Some studies have shown promising performance by the ComBat method to harmonize radiomic features in MRI[108], PET[109], and CT[110]. However, only a limited number of studies have investigated the application of image generation by GAN for preprocessing or harmonization of images to correct the variability of radiomic features. Nonetheless, further investigations and verifications are needed to understand whether each of these two techniques generalizes well to different imaging datasets across a wide range of CT protocols. An additional overview of the characteristics of our study, on ComBat and GAN techniques, compared to available studies in the literature, will be provided in the following chapters.

Another contribution of this study is in providing a one-to-one comparison of the two techniques of GAN and ComBat. It is valuable to understand the distinction of various solutions in stabilizing radiomic features. Currently, there exists no direct analogy between the harmonization performance of these two techniques.

Chapter 6 Application of Generative Adversarial Neural Networks (GAN) to improve the reproducibility of radiomic features

6.1 Introduction

6.1.1 Problem Statement and Purpose

In this chapter, we discuss our investigations in testing a mitigation strategy to remedy the impact of the variation of image acquisition and reconstruction parameters (acquisition shift) on radiomic features. The idea was to apply a technique that performs image standardization, before radiomic feature calculation, through an image transformation. Image standardization can reduce the impact of the acquisition shift caused by the variation of CT parameters. Therefore, the radiomic features calculated from the transformed images may be more similar to the radiomic features of images without any acquisition shift (referred to as the reference protocol).

In this chapter, we will discuss and test an image transformation technique for the improvement of radiomic feature reproducibility. We utilized a Generative Adversarial Neural Network (GAN) that performs image-to-image translation. We investigated whether the application of GAN on non-reference CT images can synthesize standardized images that yield radiomic features with improved reproducibility.

6.1.2 Background on GAN

The application of Generative Adversarial Networks (GAN) in image generation and image translation has received increasing attention in the computer vision field[111]. The framework has shown remarkable power in fitting to complex, multi-modal data distribution and generating realistic and high-quality images in a variety of tasks that has been a challenge for other methods such as synthesizing image from text[112], generation of music[113], unsupervised feature learning[114], impressive super-resolution images[115], etc.

GAN was initially introduced by Goodfellow *et al.* [100]. GAN's framework consists of two models - a discriminative model (discriminator) and a generative model (generator). The discriminative model, in general, is a classifier model that can be trained in a supervised manner and predict the conditional probability of classes for input or distinguish between different objects. Contrary to discriminative models, generative models can learn the probability distribution of the input; Hence they can be used to generate samples from the learned distribution of a target domain. The generator and the discriminator models can each be built Convolutional Neural Network (CNN).

The GAN model training involves an adversarial game. The generative model is responsible for the generation of sample data (fake data) that looks like real data. The discriminator is used against the generator to distinguish the samples as real or fake, and therefore it serves as a means to guide the generator's training. Unlike other generative models such as Variational Auto Encoders (VAE)[116] that make inference of the latent variables through approximation of posterior distributions of the latent variable, GAN uses the feedback from the discriminator in each iteration to estimate the model parameters and learn the data distribution.

Several variations of GAN frameworks have been proposed to adapt to a variety of tasks in the computer vision field. Recently GAN applications have further expanded into the medical imaging field. Different frameworks of GANs have been used or proposed for a variety of tasks. For example, GAN frameworks have shown satisfactory performance for data augmentation[117], segmentation[118,119], synthesizing MR scan from CT scan[120] or CT scan from bi-planar chest X-ray[121], etc. A contribution of this study is in providing insight into the potential of the GAN technique in harmonizing radiomic feature variability.

There are a variety of deep learning denoising techniques[122,123] that are applied in the image domain and may improve the impact of dose reduction but not necessarily harmonize the impact due to variation of parameters such as slice thickness variation and/or its downstream effect on radiomic features. Among these techniques is the application of CNNs that have been used to estimate a high-dose image from a low-dose image. However, some studies have shown excellent performance by GAN models in comparison to CNN-only models in the task of image denoising[124], normalizing, or image translation. Wolternik et al. [125] showed that in noise reduction of low dose CT for phantom images and patient cardiac CT data, implementation of a GAN model is superior to using a CNN model alone. Their results showed that by the addition of an adversarial loss function to the voxel-wise loss function of a CNN, the model captures the image statistics of full-dose images better and avoids unnecessary smoothing of the image. Furthermore, they compared their method with an IR method and saw a superior performance by their GAN model.

You *et al.* [126] implemented a residual CNN-based network in a CycleGAN[127] framework to recover high-resolution anatomical data from low-resolution CT images in an unpaired fashion in which the model training does not require a one-to-one relationship between training images at low or high resolution. They found that peak signal-to-noise ratio (PSNR) oriented algorithms for super-resolution imaging apply higher noise reduction and smoothing, while their GAN approach generates images with better texture details as identified by expert radiologists.

Additionally, GAN models showed success in studies that focus on standardization of Positron Emission Tomography (PET) imaging that synthesized routine-dose PET from low-dose scans.

Ouyang et al. [128] compared the performance of a CNN-only model[94] that used an encoder-decoder to a GAN model that used an encoder-decoder plus a discriminator for standard-dose image generation. The GAN model outperformed the CNN model in the qualitative assessment of expert radiologists as well as quantitative metrics of structural similarity (SSIM), PSNR, and root means square error (RMSE). An in-house preliminary study[129] also showed that the implementation of GAN model standardization of CT images between three different screening dose levels and slice thicknesses resulted in higher quality image generation than a CNN model.

6.1.3 Significance

Since the problem of image generation is a difficult task, considering existing challenges in other techniques such as VAE models, and by following the findings of the aforementioned studies in their comparisons between CNN models and GAN models, we were inspired to tackle the problem of poor reproducibility of radiomic features by testing the potential of GAN framework in mitigating variability of radiomic features.

The significance of our study in comparison to other studies using a GAN [129,130] is that we explored the potential of GANs in generating images that can be used directly for extracting quantitative imaging biomarkers to be used in diagnostic prediction models. Thus, the focus here was specific to the preservation and similarity of the high-level features (radiomic features) of the generated images to the images in the target domain. For example, Armanious *et al.* [131] proposed a new GAN framework for medical image translation with the purpose of using the generated images for improving post-processing tasks rather than performing diagnosis. Furthermore, many existing GAN studies in the CT medical imaging field have been judged and evaluated based on general-purpose metrics such as SSIM, PSNR which can correlate with human perception but not with a diagnostic task. Additionally, while for tasks such as data augmentation, the qualitative feedback of expert radiologists matters, for quantitative tasks such as building prediction models, this feedback is unnecessary. Therefore, the question of whether a GAN framework that achieves high general-purpose quality metrics can also achieve high performance in matching quantitative imaging features (radiomic features) is a question that has not been explored thoroughly. Hence, a contribution of this study is in offering such insight.

Among the studies that have used GAN image synthetization for standardization of CT images, studies by Wei *et al.* [129] and Liang *et al.* [130] have studied the variability of radiomics and GAN's impact on the reproducibility of radiomic features. Our study is an expansion of these studies in some aspects and is a complement to these investigations in other aspects. For example, we encompassed the assessment of the reproducibility of a larger number of radiomic features from different categories. Liang *et al.* [130] measured radiomic features over randomly selected 2.5D patches from the whole CT scan. However, we based our measurements and conclusion on the segmented nodule region as in the context of cancer diagnosis, reliability of radiomic features of the nodule region is much more important than the radiomic feature of patches of non-nodule soft tissue. Furthermore, we have evaluated model performance using Concordance Correlation Coefficient (CCC), which has been shown to be a very accurate reproducibility metric[59] and is a very sensitive metric that focuses on the deviation of radiomic features of individual nodules. Our study is a complement to the aforementioned studies because, firstly, we have explored the potential of a different GAN framework than these studies. Secondly, we took a different approach in the training of the GAN model such that rather than training the model on images (patches) of various parts of the lung CT scan, we only trained the model on a sub-volume around each nodule region; This approach allowed the model to only focus on the volume of interest (i.e., the nodule) and its surrounding and reduced model complexity (input size, model depth, etc.).

Hence, the insight provided in this study can be compared to findings of other frameworks used in other studies and can show whether it is sufficient to train the GAN model only based on the nodule and its surrounding.

Since we assessed mitigation of a large number of well-known radiomic features for lung nodules, our study takes a step toward building reliable approaches in facilitating future multicenter radiomic studies or clinical trials to enable clinical adaptation of radiomic techniques.

6.1.4 Overview of the Approach in this Chapter

In the methods section, we will describe how we tested a popular GAN model framework in mitigating the variability of radiomic features through the synthesis of CT scans of lung nodules across a range of CT image conditions by varying CT dose level, reconstruction kernel, and slice thickness.

In this study, the goal was to perform image standardization and image transformation, so we utilized a popular image-to-image translation framework, Pix2Pix, proposed by Isola *et al.* [132] that has been able to generate realistic images for a variety of domains and tasks. This model transforms images from a source domain to a target domain. The training process of this model involves optimization of a loss function that measures the difference between source domain images and target domain images. In addition to testing the Pix2Pix model with its original configuration and hyperparameter set, we explored fine-tuning of the model hyperparameters and configuration to investigate further improvement of the performance of the GAN model.

6.2 Materials and Methods

6.2.1 Theory and Preliminaries

6.2.1.1 Initial GAN Framework and Theory of Operation:

The initial GAN framework proposed by Goodfellow *et al.*[100] consists of two neural networks that get trained simultaneously: a generative model G and a discriminative model D. In this framework, the generator is tasked to capture data distribution (p_{data}) and map input (in form of a noise vector $z \sim p_{z(z)}$) into the data space. On the other hand, the discriminator is a classifier that estimates the probability of whether a sample (x) is from the real data distribution (p_{data}) or from the generated (fake) image by $G\sim$ (p_g). There is adversarial process between G and D in form of a minimax game. G will be trained to maximize the probability that D makes a mistake, and D is trained to minimize making errors. The competition between G and D will enable them to improve themselves through training iterations. The D's output will not only be used for backpropagation of its own network for updating weights but is also used for updating the G's weights. Hence, D's output will guide G toward generating more realistic samples. Simultaneously D attempts to learn more meaningful features to better distinguish between real and fake images. This competition continues until the two models have reached Nash Equilibrium, in which both models have been improved enough, and the fake images are no longer distinguishable from real images, and convergence happens. The minimax game between G and D is formulated as the following:

$$min_G max_D \mathcal{L}_{GAN} = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \qquad \text{Eq. 1}$$

D is trained to maximize the probability assigned to the correct label, meaning that for real images ($x\sim pdata$) it is aimed to maximize $\log D(x)$ and for fake images ($G(z)$) it is aimed to maximize $\log(1 - D(G(z))$. Though G gets trained to minimize $\log(1 - D\big(G(z)\big))$[5].

[5] The theoretical approach may not provide sufficient gradient for generator early in the training, as the discriminator may easily identify $G(z)$ and this can result in saturation of $\log(1 - D(G(z))$. Therefore, in practice instead of minimization of $\log(1 - D(G(z))$ maximization of $\log(D(G(z))$ is performed.

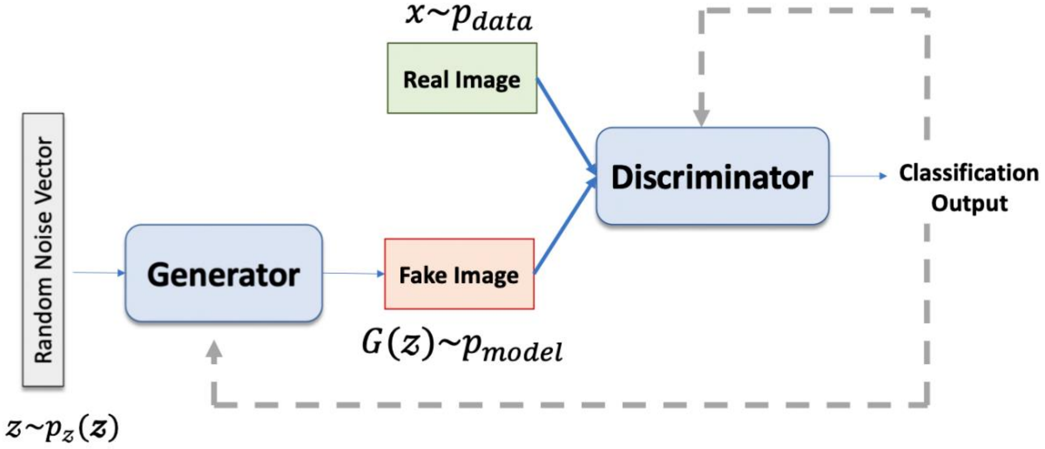


Figure 6-1. GAN model training framework

Once the training has resulted in convergence, the discriminator will no longer be needed for inference, and the trained generator is used to generate new images.

6.2.1.2 Conditional GAN (cGAN):

The initial GAN framework described in the previous section is capable of learning the data distribution and effectively generate new random plausible images for the given dataset. However, its limitation is that there are no controls on the type/class of the generated images; in other words, it is unconditioned. The conditioned version of GAN (cGAN), proposed by Mirza *et al.* [133], on the contrary, is a type of GAN that allows for the targeted generation of images. With cGAN, we can target the generation of images that belong to a specific category of interest. If class labels (y) for image data are available, we can extend the GAN into cGAN by feeding the extra information about class labels into both D and G[133] (Figure 6-2). This is achieved by conditioning both discriminator and generator models on y. So, the cost function slightly changes by replacing $D(x)$ and $G(z)$ with $D(x|y)$ and $G(z|y)$.

As an example, in the generation of dog images, if the breed of dogs is known, the information about the breed class can be used in training the models so that once the model is trained and ready, it can be used to generate images of a certain breed of interest. In cGAN, the cost function penalizes the joint configuration of the output and its minimax game shown in Eq. 2.

$$min_G max_D \mathcal{L}_{GAN} = \mathbb{E}_{x \sim p_{data}}[\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \qquad \text{Eq. 2}$$

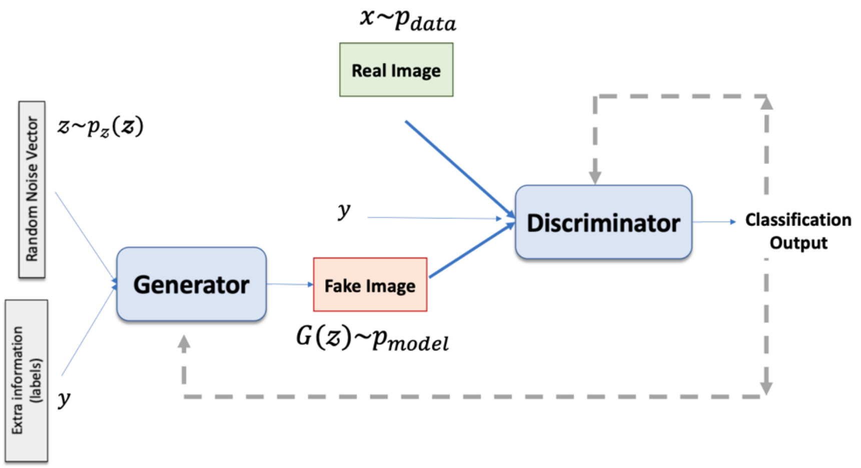


Figure 6-2. cGAN: generator and discriminator are conditioned on extra information provided by y.

6.2.1.3 Image-to-Image Translation with GAN:

Isola *et al.* [132] adapted the conditional generation concept of cGAN in a new framework (Pix2Pix) to synthesize a new image from an input image and therefore predict pixels from pixels. This framework conditions both D and G on y which is an image from a source domain (instead of a class label as was described in the previous section). The GAN training is aimed toward translation of the input image into an output image that is in a target domain. The translation is achieved through the mapping function learned by G. Therefore, the adversarial loss will be as shown in Eq. 3.

$$min_G max_D \mathcal{L}_{cGAN} = \mathbb{E}_{x \sim p_{data}}[\log D(x, y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z, y), y))] \qquad \text{Eq. 3}$$

Pix2Pix performs a paired image-to-image translation meaning that the model is fed with pair of images (one from the source domain and one from the target domain), and each pair belongs to one subject. Therefore, this translation can be considered as supervised since there is a ground truth (target image) for each source image.

The authors[132] combined the adversarial loss ($\mathcal{L}_{GAN}$) with an additional loss function in form of L1 distance between generator's output and the ground truth (target image), so that the model will be trained to both fool D and try to generate images that are structurally similar to the ground truth. Hence the objective for Pix2Pix is as follows:

$$min_G max_D \mathcal{L}_{cGAN} + \lambda_1 \mathcal{L}_{L1} \quad \text{Eq. 4}$$

Where λ_1 is a weighting hyperparameter for L1 loss and

$$\mathcal{L}_{\mathrm{L1}} = \mathbb{E}_{\mathrm{x,y,z}}[\|\mathrm{x} - \mathrm{G}(z, \mathrm{y})\|_1] \quad \text{Eq. 5}$$

In practice, the authors found that while having an addition of noise ($z \sim p_{z(z)}$) in input is necessary for the model to avoid the production of deterministic outputs, it is possible to include this noise in the form of dropout of neurons both during training and inference (testing) of the generator. Dropout in the training of a deep neural network happens by randomly dropping or ignoring neurons by temporarily removing them from the network. This adds randomness and results in making the training process noisy, and this is why the authors chose to use this technique instead of inputting a noise. Another benefit in using the dropout technique is that it helps to avoid overfitting in the training of deep neural networks; The dropping out of neurons will result in layers looking like a different layer with a different number of nodes. So, it seems like in each iteration; a different model architecture is getting trained, and therefore during the whole training procedure, an ensemble of different network architectures has been trained. Since ensemble modeling approaches are known to reduce overfitting, dropout will help to avoid overfitting too. When a node is dropped out, the remaining nodes have to take on more responsibility and get to be updated with a different view from previous layers, and this can lead to a more robust training and a more generalizable network[134].

There are two main differences between discriminator and generator in the Pix2Pix model as compared to other image-to-image translation GANs. Since the problem of image-to-image translation is aiming to find a mapping function between two images that represents the same underlying structure but with a different appearance (with different noise or quality), the structures are aligned. Furthermore, the input image and the ground truth image share the low-level features. Therefore, it will be helpful to pass the low-level information into the network during the generation of the image. The generator in Pix2Pix made the preservation of location and context possible by the usage of U-Net architecture. In the U-Net architecture, as opposed to other encoder-decoder architectures, the feature maps learned by the contractive blocks at different depth levels will be concatenated with the up-sampling blocks (through skip connections). So, the addition of the high-resolution feature vectors from different depth levels during reconstruction (in the up-sampling path) both helps for the preservation of context and for localization accuracy.

Since an L1 loss function was used in the objective function, it enforces the correctness of low-frequency image data. In order to ensure the correctness of high-frequency data, the authors[132] implemented the architecture of the discriminator in the form of a PatchGAN that restricts attention to local image patches and only penalizes image patches by classifying each patch as fake or real. Therefore, this takes care of attention to high-frequency data. The discriminator output will then be an average of outputs on all the patches across the image. The idea of application of patched base convolution and averaging for modeling and prediction is similar to the idea of modeling and transferring of image texture through quilting of patches by Efros *et al.*[135,136]. Hence, it can be expected to have this discriminator capable of understanding texture. The size of the patches or the effective receptive field size for the discriminator determines the size of the input window that was mapped to each unit in an output activation map. With an activation map with a given output size, the receptive field from the input is calculated by:

$$\text{receptive field} = ((\text{output size} - 1) \times \ stride) + kernel\ size \qquad \text{Eq. 6}$$

The procedures for training of the GAN is shown in Figure 6-3.

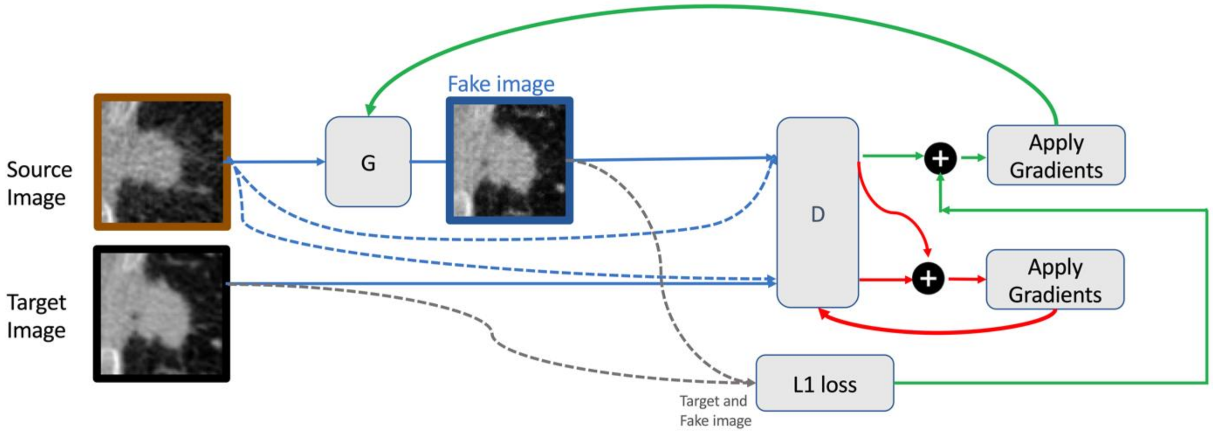


Figure 6-3. The training process for the GAN.

6.2.2 The architecture of The GAN Model Used in this Study

In this study, we investigated the application of image-to-image translation GAN for the transformation of lung nodule CT images with the purpose of improving the reproducibility of radiomic features between a reference condition (100% dose, medium kernel, and 1mm thickness) and non-reference conditions.

We studied the transformation of non-reference images acquired at seven different CT conditions. As a result, we trained and tested seven models, one for the transformation of each condition

(Figure 6-4). In experimenting with different hyperparameter sets and configuration tuning, all the seven models were trained and tested with the same architecture.

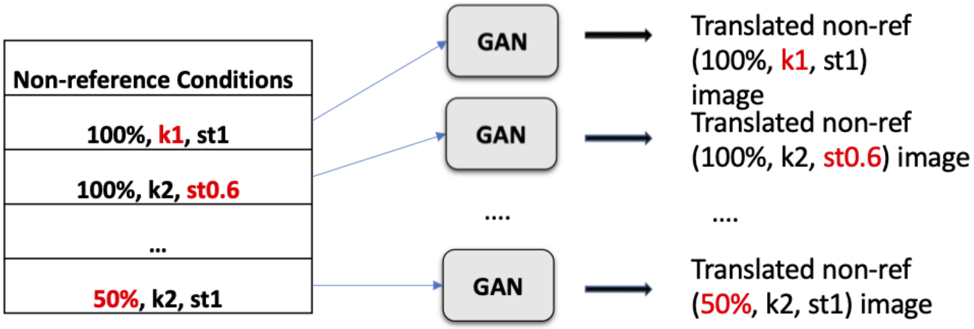

Non-reference Conditions
100%, k1, st1
100%, k2, st0.6
...
50%, k2, st1
GAN
GAN
....
GAN
Translated non-ref (100%, k1, st1) image
Translated non-ref (100%, k2, st0.6) image
....
Translated non-ref (50%, k2, st1) image

Figure 6-4. For standardization of each non-reference condition, a separate model was trained

For each of the seven models, a 3D GAN architecture was implemented, which was inspired by the 2D Pix2Pix[132]. The 3D version of Pix2Pix with few modifications was adapted from a framework called Vox2Vox proposed by Cirillo *et al.* [137]. The generator was built with a U-Net and Resnet[138], and the discriminator was built with a PatchGAN[132] (see Figure 6-5 and Figure 6-6). The input is consisted of two 3D images (source image and target image), each with 1 channel. Source domain consists of images acquired and reconstructed at a non-reference. The target domain consists of images acquired and reconstructed at the reference condition. The GAN is then trained to transform the source domain images into the target domain. Since the GAN model performs paired image-to-image translation, the model is fed with pairs of images, and each pair consists of reference and non-reference images of the same nodule.

The input images only encapsulate the nodule and its surrounding (i.e., the input does not cover the whole lung). The two images are concatenated at the channel level before feeding to the models.

6.2.2.1 The Generator's Architecture

The generator, similar to the U-Net, has a contracting path and an expanding path. For symmetry, there is equal number of contracting and expanding blocks. The contracting path consists of down-sampling blocks with 3D convolutional layers where the number of feature maps doubles at each depth to learn more complex information effectively. A dropout probability of 0.2 was considered for these blocks. As shown in Figure 6-5, each down-sampling block included the same padding, instance normalization, and activation function (Leaky RELU). The bottleneck is a layer between the contraction path and expansion path. A bottleneck consisted of 4 residual blocks, and the input of each block is the concatenation between the previous block's input and output (black arrows in Figure 6-5). Each residual block was made from 3D convolutional layers with kernel size 4, stride 1, same padding, instance normalization, and activation function of Leaky RELU. The size of inputs and outputs remains constant. The expansion path consists of up-sampling blocks of 3D transposed convolutional (deconvolutional) layers where each feature map gets in half after each block to maintain symmetry with contracting blocks. The up-sampling blocks were implemented with kernel size 4, stride 2, instance normalization, and Leaky RELU (slope of 0.2) activation function.

Each decoding block was concatenated with the encoding block of the same depth. The last layer of the decoder consisted of 3D transpose convolution with kernel size 4 and stride 2 followed by a hyperbolic tangent activation function. Although Figure 6-5 shows a U-Net with 2 layers, we tested with different numbers of layers as well. But we chose the simpler model with 2 layer as shown in Figure 6-5 (More information about configuration tuning will be provided in section 6.2.5). The output of the generator is in the shape of input image size.

The basis of the generator's loss function according to Eq. 4 is the sum of L1 loss (using mean absolute error between generated image and the target image), and adversarial loss which is in the form of sigmoid cross-entropy applied to the output of the discriminator on generated images $D(G(z))$ and a matrix of ones with the shape of discriminator's output. In the following subsections, we will describe our experiment in testing the addition of another loss function (perceptual loss) to the objective function of the generator.

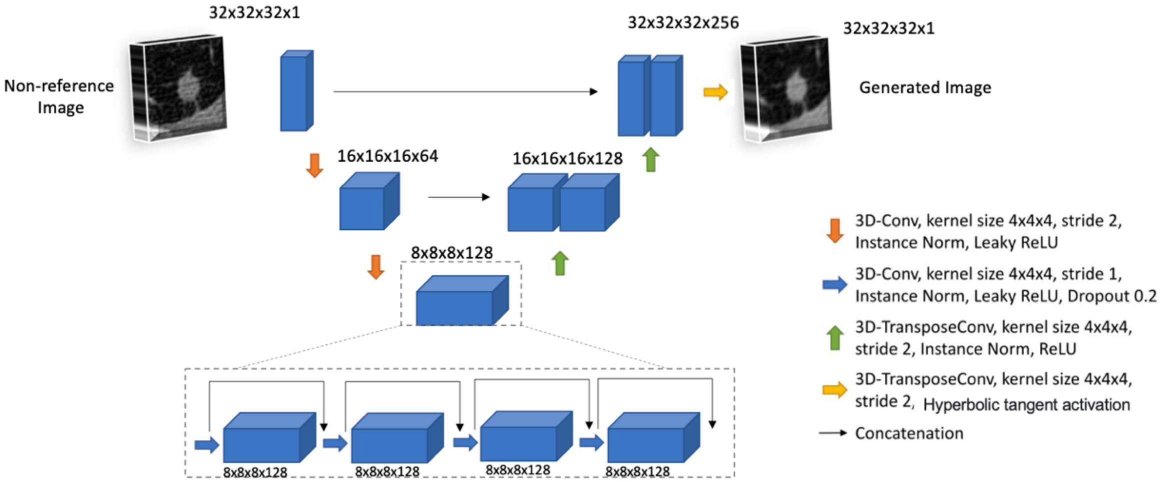

32x32x32x1
Non-reference Image
32x32x32x256
32x32x32x1
Generated Image
16x16x16x64
16x16x16x128
8x8x8x128
8x8x8x128
8x8x8x128
8x8x8x128
8x8x8x128
3D-Conv, kernel size 4x4x4, stride 2, Instance Norm, Leaky ReLU
3D-Conv, kernel size 4x4x4, stride 1, Instance Norm, Leaky ReLU, Dropout 0.2
3D-TransposeConv, kernel size 4x4x4, stride 2, Instance Norm, ReLU
3D-TransposeConv, kernel size 4x4x4, stride 2, Hyperbolic tangent activation
Concatenation

Figure 6-5. The final architecture used for the generator model (adapted from Cirillo et al.[137])

6.2.2.2 The Discriminator's Architecture

Each discriminator was configured with 3D blocks of convolution-Instance Normalization-Leaky RELU activation (with slope 0.2). Convolutional layers had kernel size of 4 and stride of 2 except for the last layer that had stride 1. The number of feature maps doubled after each block. The last layer has a sigmoid activation.

In configuration tuning, we trained and validated different number of discriminator layers (from 1 to 3) by using the validation set, but the discriminator model with 1 convolutional layer (Figure 6-6) was selected as the final choice since it resulted in better training convergence and better harmonization performance in the validation set. More information about configuration tuning will be provided in section 6.2.5.

Each unit in the output classifies a portion of the input image with the size that is calculated by Eq. 6. Therefore, each voxel in the output layer maps to a 7x7x7 receptive field in the input layer. Since the discriminator is a binary classification model, it predicts the probability of belonging to a real or fake class and is in the range of [0, 1]. The loss function for the discriminator is binary cross-entropy loss for classification. Since the discriminator receives two inputs, it classifies two pair of images, real pair (source image, target image) and fake pair (source image, fake image). Therefore discriminator's loss is the sum of real loss (loss on real images) and fake loss (loss on fake images). The loss is calculated by comparing images with a matrix of target values (with the same shape as output). The target value is 0 for fake images and is 1 for real images.

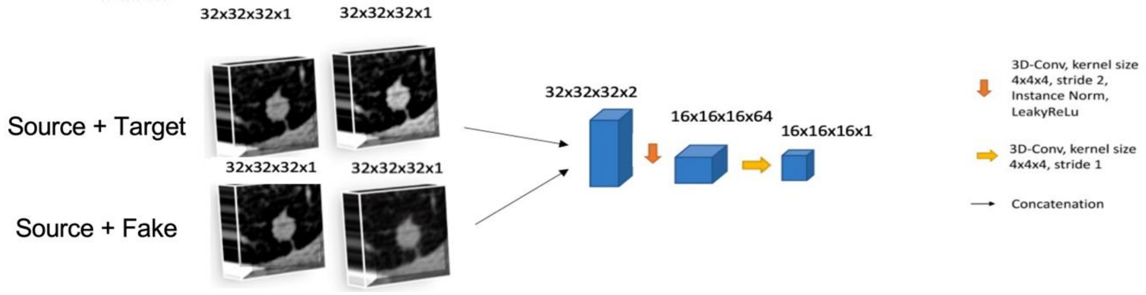


Figure 6-6. The final architecture used for the discriminator model. (adapted from Cirillo et al.[137]). The discriminator receives pair of images. Gan image and reference image (target) are concatenated with the source image.

6.2.3 Data:

Cohort 3 described in Chapter 3 was used for this study. Images of 385 cases were available, and for each case, only one nodule (the largest if there was more than one for a given patient) was enrolled in the study. The image data was divided into a training and an independent testing set that included 70% (269 cases) and 30% (116 cases) and is referred to as cohort 4 and cohort 5, respectively, of the data for training and evaluating of the GAN model. 25% of the training cases were randomly assigned to a validation set for hyperparameter tuning and monitoring the model performance. The condition (100%, k2, st1) was considered as the reference condition as it is the most similar to the lung cancer screening protocol used in our institution.

Table 6-1. Description of CT conditions included in the study. (Values in bold show the reference condition). This table shows the variation in reconstruction kernel (2 conditions), dose level (3 conditions) and slice thickness (2 conditions)

% of Screening Dose	wFBP Reconstruction Kernel	Slice Thickness	Condition Identifier
100%	**Medium**	**1mm**	**100%, k2, st1**
100%	Smooth	1mm	100%, k1, st1
100%	Sharp	1mm	100%, k3, st1
50%	Medium	1mm	50%, k2, st1
25%	Medium	1mm	25%, k2, st1
10%	Medium	1mm	10%, k2, st1
100%	Medium	0.6mm	100%, k2, st0.6
100%	Medium	2mm	100%, k2, st2

6.2.3.1 Preprocessing

Initially, from each CT scan, a sub-volume of size 64x64x64 voxels that were centered at the nodule region was depicted such that it captures the complete segmented nodule region. The goal was to only train the GAN model on a 3D patch that contains the nodule rather than training on the whole lung.

Afterward, Chest CT images underwent resampling to a pixel size of (0.7, 0.7) in the axial plane and pixel depth equal to the slice thickness in the z-direction. Images were normalized so that HU values lie in the range of (-1, 1). These normalizations are required to ensure consistency in data distribution. Pixel value normalization ensures that for all input images, the learned weights and biases are in a similar range that is also small and not large. If pixel values are large or largely deviate, weights will deviate largely, and gradients will not act uniformly and may get out of control. Pixel size normalization ensures that each pixel relates to a fixed size (0.7mm) of an image in each direction. If pixel sizes differ for images of a subject across different conditions, the amount of intensity averaged in each pixel differs, and this can cause another inconsistency in addition to CT condition variation.

6.2.3.2 Augmentation

In order to improve the ability of the GAN model so that it generalizes well to diverse range of images that have different orientations and nodule appearances, each image in the training set were also rotated at 90, 180, and 270 degrees. This resulted in a total of 1076 samples in the training set. Furthermore, in order to avoid the model using the same images in each epoch, images were randomly flipped along the x or y-axis online and during the training of GAN models.

6.2.3.3 Radiomic Feature calculation

Out of 116 cases (cohort 5), 55 cases from the held-out test set had nodule segmentation by an in-house Computer-Aided Detection (CAD)[43] software. These cases are referred to as cohort 6. Before the application of GAN, three CAD segmentation masks were available for each subject and for each of the three different slice thickness reconstructions. Therefore, at each slice thickness, the corresponding mask was used to calculate the radiomic features (referred to as unharmonized radiomic features). Figure 6-7(a) illustrates how the masks were used for radiomic feature calculation at non-reference conditions before GAN application. Figure 6-7(b) demonstrates the radiomic feature calculation of the reference condition.

The aim of the GAN application was to transform non-reference images such that they match with the reference condition with 1mm thickness. The GAN transformation of non-reference images (at 0.6mm and 2mm) resulted in images that appeared more like the 1mm thickness images. Therefore, after GAN application, the mask segmented at 1mm condition was used for radiomic feature calculation. Figure 6-7(c) demonstrates the overlay of 1mm segmentation masks on GAN-generated images.

Radiomic features were extracted from ROIs defined by these segmentation masks using Pyradiomics Software[57]. 93 radiomic features were extracted, including features from different radiomic feature groups of density histogram-based (first-order), Gray Level Run Length Matrix (GLRLM), Gray Level Dependence Matrix (GLDM), Gray Level Size Zone Matrix (GLSZM), Gray Level Co-Occurrence Matrix (GLCM), Neighboring Gray Tone Difference Matrix (NGTDM). For all features, images were discretized by 16 bins. For GLCM, a step distance of 1 was used, and matrices were merged slice by slice. Other settings that were used were those set as default in Pyradiomics. The name of all the 93 features are available in Table B-2 of Appendix B.

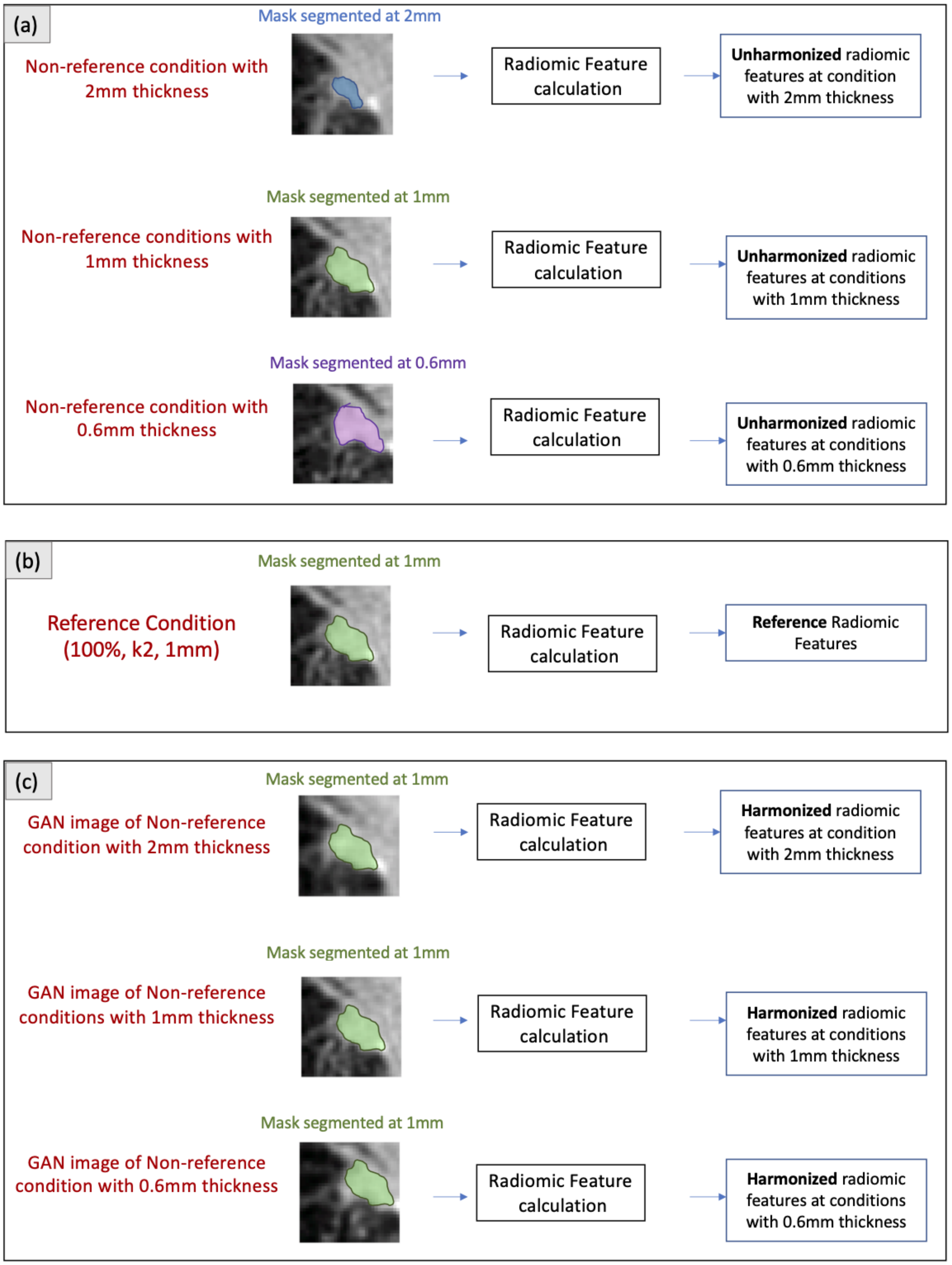

(a)
Mask segmented at 2mm
Non-reference condition with 2mm thickness
Radiomic Feature calculation
Unharmonized radiomic features at condition with 2mm thickness
Mask segmented at 1mm
Non-reference conditions with 1mm thickness
Radiomic Feature calculation
Unharmonized radiomic features at conditions with 1mm thickness
Mask segmented at 0.6mm
Non-reference condition with 0.6mm thickness
Radiomic Feature calculation
Unharmonized radiomic features at conditions with 0.6mm thickness
(b)
Mask segmented at 1mm
Reference Condition (100%, k2, 1mm)
Radiomic Feature calculation
Reference Radiomic Features
(c)
Mask segmented at 1mm
GAN image of Non-reference condition with 2mm thickness
Radiomic Feature calculation
Harmonized radiomic features at condition with 2mm thickness
Mask segmented at 1mm
GAN image of Non-reference conditions with 1mm thickness
Radiomic Feature calculation
Harmonized radiomic features at conditions with 1mm thickness
Mask segmented at 1mm
GAN image of Non-reference condition with 0.6mm thickness
Radiomic Feature calculation
Harmonized radiomic features at conditions with 0.6mm thickness

Figure 6-7. Illustration of calculation of radiomic features for (a) unharmonized non-reference conditions, (b) reference condition, and (c) harmonized non-reference condition. (a) at each non-reference condition, a unique mask is used for images at different slice thickness. (c) after GAN application the mask segmented at the reference condition (1mm) is used for feature calculation.

6.2.4 Evaluation

Currently, in the general computer vision field, there are a couple of metrics and scoring algorithms proposed by researchers used for the evaluation of GAN's performance. Yet, the question of which metric is the best is left as an open question, and there is no general-purpose metric that can be used to judge GAN's performance as opposed to other prediction tasks like classification. Hence, it will be best if the judgment relies on task-specific evaluations. Among general-purpose metrics, Structural Similarity Index (SSIM)[139] and Peak Signal to Noise Ratio (PSNR) have been used widely in the literature for reporting and comparison of different GAN models. SSIM is a metric for quality assessment between two images based on the degradation of structural information and measures similarity in terms of structure, contrast, and luminance. SSIM measures image quality in accordance with the human visual perception of structure in an image. PSNR is a measure of quality generally used for evaluating the reconstruction of a signal. In a comparison of two images, it is calculated by the distance of the two images (mean squared error) scaled by the maximum fluctuation of image data and measures the peak error.

While we measured quality metrics of average SSIM and PSNR for transformed images of the sub-volumes for each non-reference condition, we performed task-specific evaluations, and we based our judgment and conclusions on task-specific evaluation metrics.

The purpose of the study was to test the potential of GAN in the mitigation of variability caused by CT acquisition shift in images in a way that results in the harmonization of radiomic features. Once the hyperparameters were tuned, and an optimal model configuration was identified, we used the generator trained with optimal configuration to transform non-reference images. The data that was used to test the GAN approach included CT images with different conditions (a reference condition and seven non-reference conditions) for all subjects (55 cases) in the held-out test set.

We first performed reproducibility analysis between radiomic features of non-reference conditions ($RF_j\ for\ j = 1, \dots ,7$ non-reference conditions) and radiomic features of the reference condition (RF_{j_ref}) for the 55 cases. We then applied GAN on non-reference conditions and calculated radiomic features (RF_{j_GAN}) from the transformed non-reference conditions and made comparisons to the reference radiomic features.

RF_{ref} served as the figure of merit and was compared to RF_j and RF_{j_GAN}. To measure the agreements to RF_{ref} and understand the reproducibility of RF_j and RF_{j_GAN}, Concordance Correlation Coefficient (CCC)[59] was calculated to quantify the agreement. CCC measures the deviation from the 45° identity line between paired data. Initially, and before GAN transformation, for each non-reference condition and for each feature (i), CCC_{ij1} was calculated between each non-reference and reference radiomic feature. After the GAN application, CCC_{ij2} was calculated between the reference radiomic features and the radiomic features from the transformed images.

We hypothesized that when GAN transformation is applied on CT images of the subjects at a non-reference condition, agreement of radiomic features improves (i.e., $CCC_{ij2} > CCC_{ij1}$). We assessed this hypothesis through the utilization of a mixed-effects linear model that models differences between CCC values ($CCC_{ij2} - CCC_{ij1}$) as a function of the harmonization effect of each 7 non-reference condition j (j = 1, 2, …,7). The model was described as:

$$Y_{ij} = \mu + u_i + \beta_j + e_{ij} \qquad \text{Eq. 7}$$

Since we are interested in the improvement of CCC_{ij2} compared to CCC_{ij1}, Y_{ij} is determined as $CCC_{ij2} - CCC_{ij1}$ for condition j, feature i ($i = 1, 2, \dots , p$). μ is the average of differences in CCC values. β is the fixed effect from harmonization of the j^{th} non-reference condition on Y_{ij}, and u_i is a simple, scalar random effect for features and is added to the intercept to adjust for repeated measurements.

. $u_i$codes which observation belongs to which feature. The random effect for the i^{th} the feature is the deviations in the intercept of that feature from the population values and follows $\mathrm{N}(0, {\sigma_i}^2)$. So it takes into account the potential between feature variabilities. There will be a total of p random effects, one for each feature. e_{ij} is the error that is assumed to follow a normal distribution with mean zero and variance σ^2.

Since we are interested in seeing a positive effect on CCC values such that $CCC_{ij2} - CCC_{ij1}$>0, one-sided test with the null hypothesis (H_0) and alternative hypothesis (H_a) for the coefficients of the fixed effects was determined as shown in the following at the confidence level of 95%:

$$H_0: \beta_j \leq 0, \qquad H_a: \beta_j > 0 \qquad \text{Eq. 8}$$

For each non-reference condition j

A confidence interval (CI) was also calculated for the one-sided test as [$\bar{X} - t_\alpha {}^{S}/_{\sqrt{n}}$,1].

As expected, a large number of radiomic features were correlated. Therefore, features were filtered to acquire a group of uncorrelated features with $correlation\ coefficient\ (r) < 0.7$.

In comparing paired measurements, CCC ≥ 0.9 indicates strong agreement[60], CCC in the range of (0.8, 0.9) is indicative of good or moderate agreement, and CCC < 0.8 is indicative of poor agreement. Therefore, we also examined whether after GAN application, the agreement of radiomic features has increased enough to meet these criteria an ($CCC_2 \geq$ 0.9). Furthermore, average differences in CCC values before and after GAN application was calculated using Eq. 9.

$$Average\ CCC\ difference = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{CCC_2 - CCC_1}{CCC_1} \times 100\right) \qquad \text{Eq. 9}$$

$$for\ M = 93\ radiomic\ features$$

6.2.5 Model Training and Tuning Configurations:

Following the standard approach for training of GANs[100], the generator and discriminator were trained alternately for up to 1000 epochs. A batch size of 4 and Adam solver[140] momentum parameters of $\beta_1 = 0.5, \beta_2 = 0.999$ were used. The network was trained from scratch and weights were initialized with a normal distribution of mean 0 and standard deviation of 0.05.

The GAN model was implemented by using the TensorFlow interface[141] by using the Keras API. To control memory consumption, instead of feeding all the training images at once, batches of images were fed to the model in real-time. At the end of each epoch, training images were randomly shuffled to change the order of images being fed to the model.

We performed hyperparameter and configuration tuning to obtain the model that best fits our data. In tuning the GAN model, we trained modified GAN models (with different hyperparameter sets and configurations) on each non-reference condition. Then the performance on the validation data was monitored by assessing the learning curves (loss function over time) and improvement of CCC values for the radiomic features in the validation set. For each condition, the GAN model that had the best performance in terms of learning curve convergence and CCC improvement was chosen as the final model for that condition. For example, in assessing CCC improvement, we examined the number of features with $CCC_2 \geq 0.9$ or $CCC_2 \geq \ CCC_1$. In the following sub-sections, we will briefly describe the configuration tuning.

6.2.5.1 Selecting Input Size

In training and testing the GAN model, we only used a sub-volume of chest CT scan that includes the nodule region and its surrounding area. As was previously mentioned, the purpose of the study is improving the agreement of quantitative imaging features between a reference and non-reference image. Therefore, focusing the model’s attention only on the nodule region of interest avoids from confusing the model by exposing it to images from other sections of the lung with different complexity (in terms of shape, intensity, or contrast). Additionally, the usage of smaller-sized data (e.g., the nodule ROI) rather than the large CT slices (512x512) allows for less technical complexity and efficiency in training with respect to memory and computation capacity.

We initially trained a model with an input size of 64x64x64, but the model performed poorly where the training and validation loss for the generator was diverging. However, a smaller input size (32x32x32) resulted in better convergence of GAN during training and validation.

6.2.5.2 Variation of Discriminator and Generator Depth

Increasing or decreasing the number of convolutional layers (or the depth of the network) is correlated with the complexity of the model. Therefore, we trained the models with different depth layers and tested the models on the validation set to find the optimal architecture. For the discriminators we tested 1 to 3 down-sampling blocks (with receptive field sizes of 7 and 22). For the generator we tested two different generators with either 1 or 4 down-sampling and up-sampling convolutional blocks. Our preliminary results showed that using 2 down-sampling and up-sampling convolutional blocks for the generator and 1 down-sampling blocks for the discriminator resulted in less overfitting and better convergence when training and validation losses were compared. Therefore, we selected these as the final choice for discriminator and the generator (also shown in Figure 6-5 and Figure 6-6).

6.2.5.3 Further Configuration Tuning

The training of the GAN model can be difficult as it can often be hard to balance the competition between the discriminator and the generator in their minimax game. In the training procedure of the GAN, maintaining the balance between performance and capacity of discriminator and generator is very important. If one model achieves high performance while the other is poorly performing, the GAN convergence will not be achieved, or gradients will vanish. If the discriminator becomes so powerful that it won't get fooled by any of the generated images in the training iterations, it will not provide good feedback for the generator, and the generator's loss keeps increasing while the discriminator's loss goes to zero. If the discriminator's loss becomes zero, it will not provide any gradient for the generator, and the vanishing gradient is one of the main difficulties in training GAN models.

A variety of regularization techniques, hyperparameter selections, architecture modifications, etc., have been proposed in the literature for non-medical applications[142,143] These techniques have been derived theoretically or through practical experiences. We utilized a selection of these techniques to understand if they can improve the GAN model training and the model performance in improving the reproducibility of radiomic features of transformed images in the validation set.

We trained different variants of a GAN model; in each variant, we applied a combination of these techniques (Though we did not perform an exhaustive search in the space of hyperparameters and configurations). The model convergence and the agreement analysis of radiomic features of the validation set were the basis for judging the efficacy of the configuration tuning.

For example λ_1, the weighting hyperparameter for L1 loss, was set to 100 in the original Pix2Pix. However, we also tuned this hyperparameter (for values of 1,10,50,100) and found $\lambda_1 = 100$ as the optimal choice for our dataset.

As another example, tuning the learning rate is an important part of any deep learning model training. In addition to this, in the context of GANs, this can be used as a trick for handling the balance between the discriminator and the generator. Learning rates can be the same for the two models, or it can be inequal. Heusel *et al.*[144] proposed this trick named as Two Time-Scale Update Ratio (TTUR) and showed that it improved learning and ensured convergence for a variety of

GAN architectures. The rationale for having a higher rate for the generator is to ensure the generator keeps up with the discriminator. The rationale for having a higher rate for the discriminator is to provide a larger and more efficient weight update both for itself and for the generator. The initial version of our GAN model used the same learning rate of 2e-4 for both discriminator and generator. We also explored the use of different learning rates for discriminator and generator. Comparisons were made between models with the exact same architecture and hyperparameter that either used the same learning rate for both discriminator and generator or used different rates (twice or half the rate in the ranges of 1e-5 to 1e-3). The model with a discriminator learning rate of 2e-4 and generator learning rate of 1e-4 was selected as the final choice for the GAN models.

Since in GAN training, the generator's job, especially at the beginning of the training, is much more difficult compared to the discriminator's job. The discriminator's loss can easily be minimal at the beginning. Hence, it can be beneficial to help the generator to pick up its speed in competing with the discriminator by freezing the discriminator from being trained at the beginning of training for a few epochs. We trained and validated variants of a GAN model by freezing the discriminator's training for 0 to 50 epochs and made comparisons based on its impact on model convergence and the discriminator's loss function. For all conditions, the model benefited from the freezing of the discriminator for the first 20 epochs.

Recently, a set of different loss functions have been introduced in the training of GANs that allow the model optimization to capture both high-level and low-level information in an image transformation task. Johnson *et al.*[145] proposed the application of a perceptual loss that, unlike per-pixel loss functions (e.g., L1 loss), is based on loss of high-level features. Since in our problem we care about image content and specifically the high-level imaging features within the nodule image, we explored whether it can be useful to use a loss function that penalizes model optimization based on the preservation of this information. In the literature[146,147], features extracted from a pre-trained convolutional neural network (trained for classification task) as well as the activation maps of the pre-trained network have been used as a basis for measuring loss for the generated image. Ouyang *et al.*[128], in generating standard-dose amyloid PET images using their GAN model, also proposed a similar loss function but instead of using pre-trained VGG[148] used a pre-trained amyloid status classifier to extract features related to the pathology. While the idea is taken from literature, we defined our own perceptual loss to test whether the inclusion of such a loss function can help the GAN model in the generation of images with similar high-level features. Since the purpose of training this GAN is a harmonization of radiomic features of nodule tissue, we decided to penalize the model on the deviations between the radiomic features extracted from the target image (x) and the generated image $G(y)$.

A set of 15 radiomic features (Table B-3 in Appendix B) were calculated from the whole 3D image (of size 32x32x32). Features were scaled to fit in the range of (-1,1). Mean squared error was measured between the radiomic feature vectors $(l_x \; and \; l_{G(y)})$ and the loss multiplied by a hyperparameter λ_2 was added to the total generator loss (see Eq. 10 and Eq. 11 and Figure 6-8). From each different group of radiomic features, a few features were selected by choosing from the features that were initially not reproducible before GAN application (CCC<0.9 when compared to the reference). Also, in the selection of these features, we tried to select features with a variety of definitions such that each feature represents different characteristics of the image.

$$\mathcal{L}_{perceptual} = \mathbb{E}_{x,y}\left\|l_x - l_{G(y)}\right\|_2^2 \qquad \text{Eq. 10}$$

$$\mathcal{L}_{Total} = \lambda_1\mathcal{L}_{L1} + \lambda_2\mathcal{L}_{perceptual} + min_D max_G \mathcal{L}_{Adversarial} \qquad \text{Eq. 11}$$

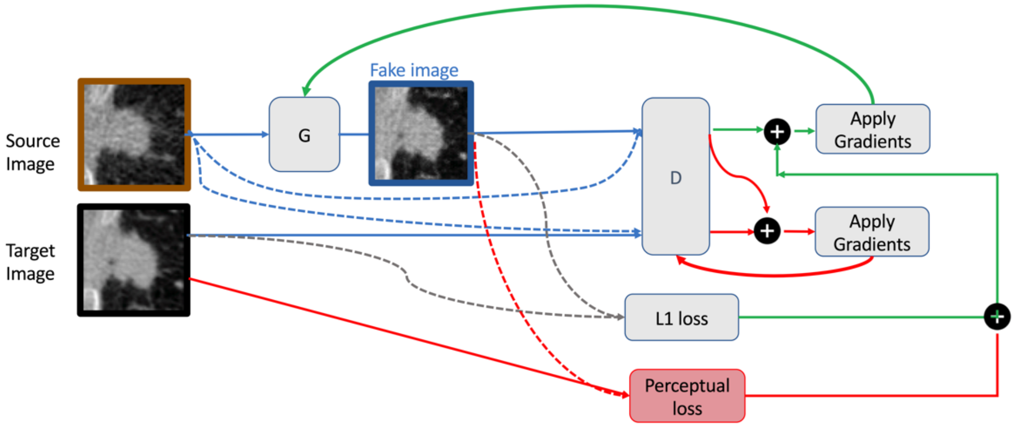


Figure 6-8. G training with an additional cost function for perceptual loss as described in Eq. 10 and Eq. 11

To identify whether the addition of this perceptual loss is beneficial for GAN image generation for each condition, we compared models with the same settings but either with or without the perceptual loss. The hyperparameter λ_2 was also tuned (in the range of [0.1,1,2,2.5,5,10,20]) by making comparisons between training and validation learning curves and CCC value improvement. For all non-reference conditions, the addition of the perceptual loss either resulted in improvements of CCC. However, the impact was the most substantial for harmonization of the conditions with 10% and 25% dose level with $\lambda_2 = 2.5$. While for other conditions, the addition of perceptual loss did result in some improvements, since the current format of this loss function was not implemented with computationally efficient manner, we preferred to not select this loss function for conditions other than the ones with 10% and 25% dose level.

The last configuration tuning that was explored involved handling the overconfidence of the discriminator. In GANs, overconfidence happens when the discriminator becomes powerful in distinguishing real images where it only relies on a few features for classification. The generator in turn will also only focus on faking those features in its image generation. One approach is to increase the complexity of the discriminator's prediction task by adding a small amount of gaussian noise ($\sim\mathcal{N}(0,\sigma^2)$, $\sigma = 0.05$) to the discriminator inputs.

While the implementation of this technique resulted in improvements in the reproducibility of radiomic features in the validation set for some conditions, its impact was not significant. This can suggest that the training did not suffer from the overconfidence of the discriminator. Therefore, we did not include this technique in the finalized GAN models.

The final configuration setting and hyperparameter values are chosen for the GAN model based on the tunings on the validation set are shown in Table 6-2. The models for harmonization of the conditions with dose levels of 10% and 25% were trained with setting 1, and the other models for the five other non-reference conditions were trained with setting 2. In Appendix B, Figure B-1 shows the learning curves for the finalized model for one of the non-reference conditions. Models with setting 1 took between 2 to 8 hours for training based on the determined GPU usage (from 100% to 30%). Models with setting 2, which included perceptual loss function, took from 12 hours to 36 hours of training according to the GPU usage.

Table 6-2. Details of the GAN models settings

Setting	Model Details
1	$\lambda_{L1} = 100, \lambda_2 = 2.5, lr_D = 2e-4, lr_G = 1e-4$ D frozen for 20 epochs in the beginning of the training
2	$\lambda_{L1} = 100, \lambda_2 = 0, lr_D = 2e-4, lr_G = 1e-4$ D frozen for 20 epochs in the beginning of the training

6.3 Results

6.3.1 Example of Images Generated by GAN

Figure 6-10 and Figure 6-9 provide examples of images for a nodule at the reference condition and the non-reference conditions before and after GAN application. The nodule shape in GAN transformation of conditions with varying kernel or dose (Figure 6-10 and Figure 6-9(a)) appears similar to the nodule shape before GAN application. Figure 6-11 shows consecutive slices of a nodule at the condition with 2mm before and after GAN and compares it with the nodule at the reference condition. In the non-reference image (2mm), the nodule appears in a fewer number of slices than the reference condition (1mm) due to volume averaging. However, after GAN, the transformed non-reference image appears in the same number of slices as the reference. It appears that the GAN transformed the non-reference image into an image that appears more similar to the reference condition than the unharmonized non-reference condition.

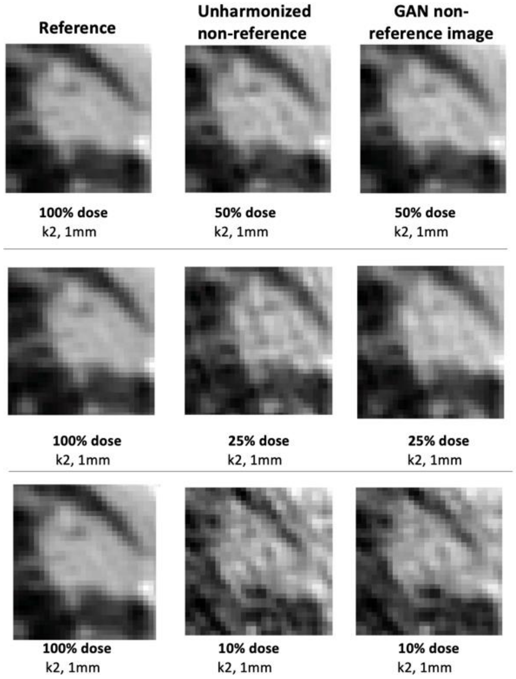


Figure 6-9. Example images of a nodule at reference and non-reference conditions both before and after GAN. Dose is the varying CT parameter

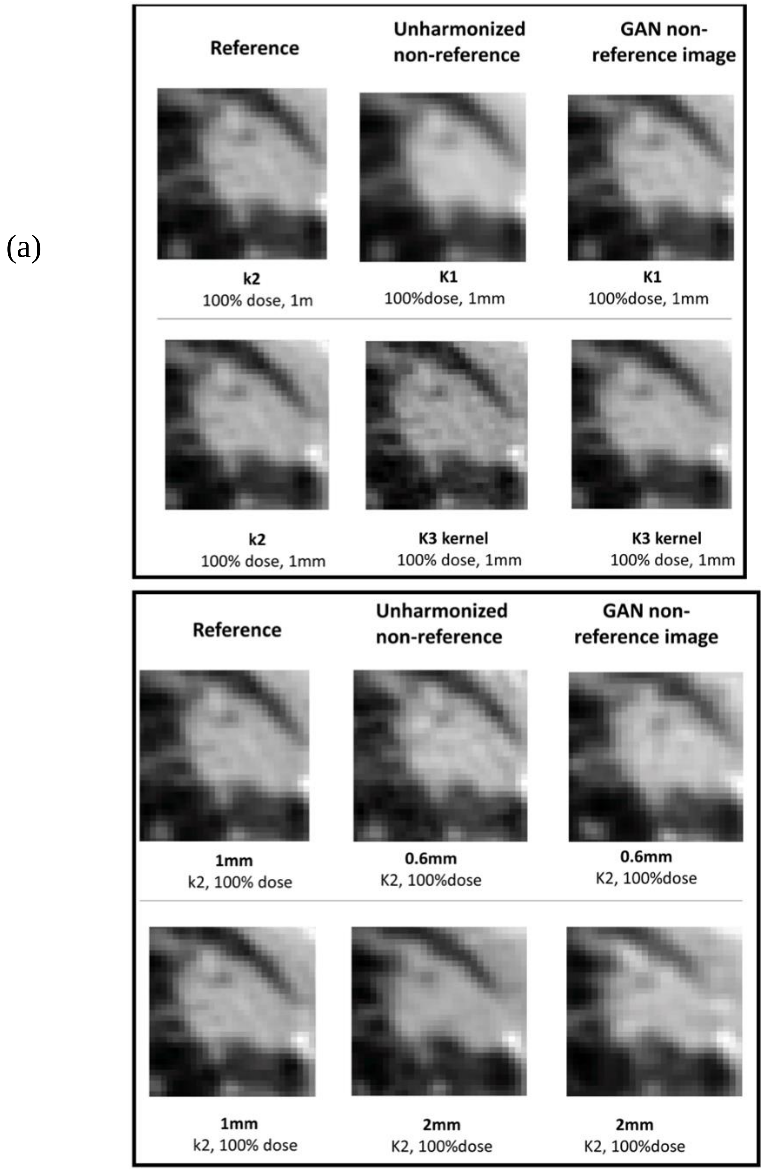


Figure 6-10. Example images of a nodule at reference and non-reference conditions both before and after GAN. (a) kernel is the varying CT parameter, (b) slice thickness is the varying parameter

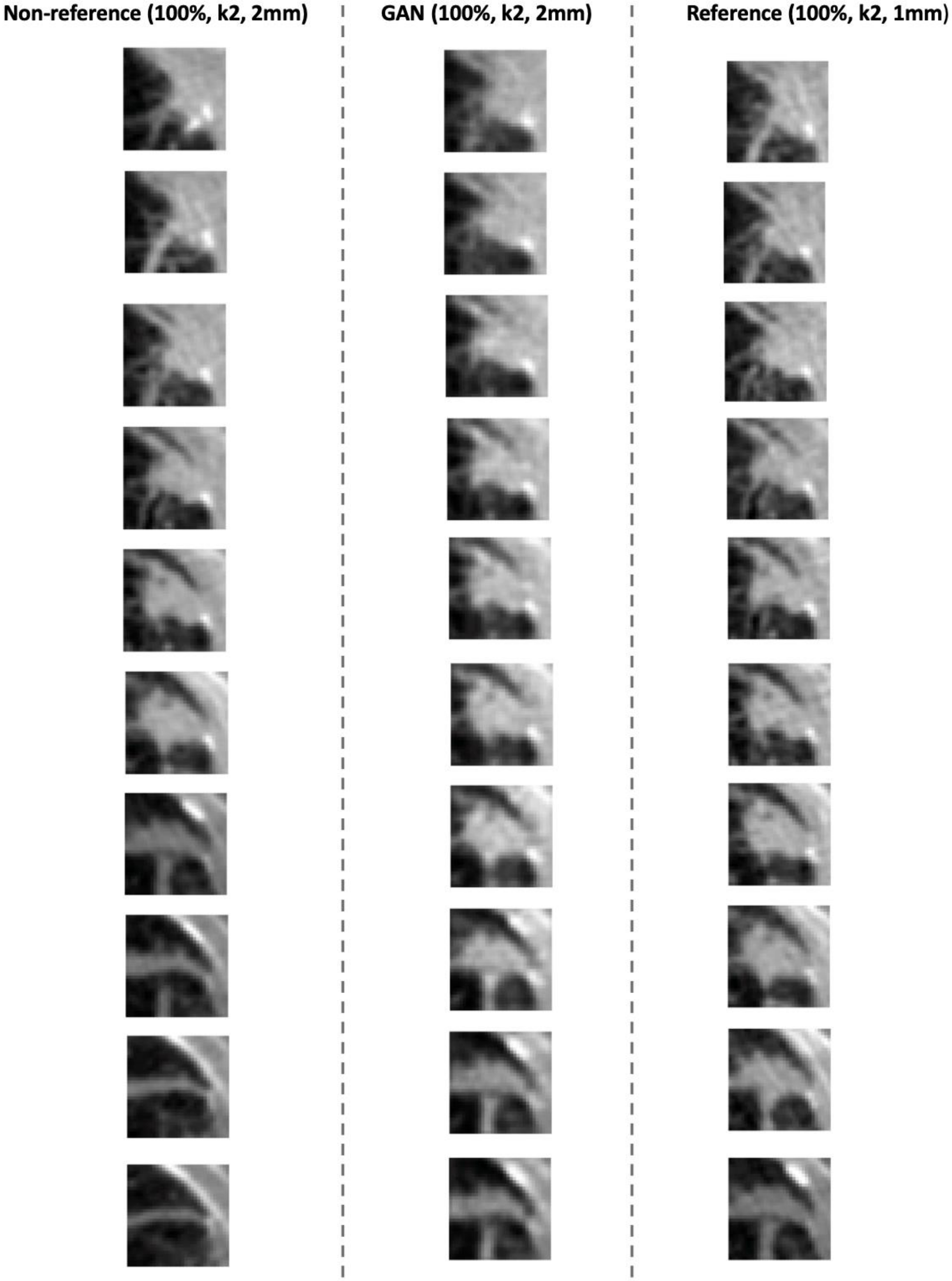


Figure 6-11. Images of a nodule at consecutive slice thicknesses. (a) The non-reference condition with 2mm thickness, (b) the GAN transformation of the non-reference condition with 2mm, (c) the reference condition with 1mm.

6.3.2 Agreement Analysis Before and After GAN Application in the Held-Out Test Set for all radiomic features

Table 6-3 summarizes the impact of the image-to-image translation by the finalized GAN models on 93 radiomic features that were not shape or size descriptors. These radiomic features were calculated over nodule VOIs of GAN-generated images for 55 cases in the held-out test set. This table shows the number of non-reference radiomic features that have improved agreement with reference ($CCC_2 > CCC_1$) and the average difference of CCC values calculated by Eq. 9, as well as the number of radiomic features that meet the cut-off value of 0.9 before or after GAN harmonization.

According to Table 6-3, after GAN application on several non-reference conditions, a large number of features had an increase in their agreements to reference ($CCC_2 > CCC_1$). Additionally, for three of the conditions (with k1, k3, or 10% dose level), the number of features with $CCC_2 \geq 0.9$ was more than twice of the number of features with $CCC_1 \geq 0.9$. The GAN application on images with different slice thickness than reference resulted in a slight improvement of average CCC values for the condition at 0.6mm, and a considerable improvement of average CCC values at the condition with 2mm thickness. After GAN application, the condition with 2mm thickness had 12 more radiomic features with $CCC_2 \geq 0.9$.

At 50% dose level, since the majority of radiomic features were already reproducible before harmonization ($CCC_1 \geq 0.9$), the GAN application did not result in improvements. At 25% dose level agreements to reference and the number of features with $CCC_2 \geq 0.9$ improved moderately.

Table 6-3. The percentage difference between CCC before and after application of ComBat harmonization for 93 radiomic features, the number of features for the specified ranges of CCC values, and the increased number of features with CCC≥0.9

Varying CT parameter	Non-reference Condition	Average CCC difference	$CCC_2 > CCC_1$	$CCC_1 \geq 0.9$	$CCC_2 \geq 0.9$	$C2 - C1$(%)[a]
Kernel	100%, k1, st1	10.6%	85	51	82	31 (60%)
	100%, k3, st1	8.2%	83	50	79	29 (58%)
Dose	50%, k2, st1	-0.1%	31	75	74	-1(-0.01%)
	25%, k2, st1	3.8%	68	40	53	13(32%)
	10%, k2, st1	19.3%	85	20	55	35 (175%)
Slice Thickness	100%, k2, st0.6	5.9%	43	20	15	-5 (-25%)
	100%, k2, st2	60%	64	1	13	12(1200%)

[a] C_1 is the count of features with $CCC_1 \geq 0.9$ and C_2 is the count of features with $CCC_2 \geq 0.9$. The numbers in the parenthesis show a percent increase in the number of radiomic features

Figure B-4, Figure B-5, and Figure B-6 in Appendix B visualize the summarized information of Table 6-3 by showing the CCC_1 and CCC_2 values for each of 93 radiomic features. The improvements of CCC values after GAN application can be visualized by the shift toward higher numbers in the vertical axis.

6.3.3 Results from Hypothesis Testing on 15 Uncorrelated Features

Before testing the hypothesis in Eq. *8*, a group of 15 uncorrelated features (names are given in Appendix B Table B-1) was initially selected. Figure 6-12 to Figure *6-14* visualize both CCC_1 and CCC_2 values for each of these uncorrelated radiomic features at each non-reference condition. The name of the radiomic features corresponding to the index number on the horizontal axis can be found in Appendix B (Table B-1). For conditions that we see a shift of data points to higher values of CCC (vertical axis), harmonization had positive influence on the feature's reproducibility (i.e., $CCC_2 > CCC_1$). These figures demonstrate two CCC cut-off values. 0.8≤ $CCC < 0.9$ indicates moderate/good agreement and $CCC \geq 0.9$ indicates strong agreement.

As was mentioned earlier, the model in Eq. *7* was built by $p = 15$ radiomic features to test significance of the impact of harmonization of conditions with different CT parameters on differences between $CCC_1$and CCC_2. Table 6-4 summarizes the results from the mixed-effect model that tested whether CCC_2 values are significantly higher than CCC_1 as a result of harmonization of each non-reference.

The harmonization of the conditions with different kernels (k1 and k3), the condition with 10 % dose, and the condition with 2mm slice thickness have p-value<0.05. This indicates that the GAN harmonization of these conditions resulted in significant improvement of CCC_2 compared to CCC_1. The estimated fixed effect for the harmonization of these three conditions (Table 6-4) is larger than that of the non-significant conditions, and the corresponding one-sided CI for these three conditions lies above zero. Furthermore, for these three conditions, Figure 6-12 (c), Figure 6-13, and Figure 6-14(b) demonstrate the increase of agreements from the CCC_1 to CCC_2 for the radiomic features included in the mixed-effect model.

As seen in Figure 6-12(a) and (b), for 50% dose level, GAN application did not result in large substantial improvement of CCC values for the set of 15 radiomic features. Also, as demonstrated in Figure 6-12(b), for 25% dose level, the GAN harmonization either kept the reproducibility of the radiomic features similar to before or slightly improved it, so the impact was not significant. Simultaneously, both in Figure 6-14 and Table 6-4 we do not observe significant mitigation of the radiomic features at the condition with 0.6mm slice thickness variability.

Figure 6-15 shows the range of $CCC_2 - CCC_1$ values for the 15 radiomic features via boxplots. The boxplots of the conditions that received significant improvement have more positive mean values and distribution.

The estimated intercept in the mixed-effect model was -1.03e-16 ± 0.03 with CI: [-0.05, 1]. The between-subject variability was low as the random effect had an estimated standard deviation of 0.04, and the confidence intervals for the random effects lay around zero (Figure B-2 in Appendix B). These provide certainty in the results from the model.

In the Quantile-quantile plot (Figure B-3 in Appendix B), the residuals had slight deviations from the theoretical normal quantiles. The assumption of equal variances was also verified with Levene's test as there was no significant difference in the spread of residuals among features (*p*-value >0.05).

Table 6-4. Estimated fixed effects of harmonization of CT parameters on improvement of CCC

Harmonized non-reference Condition	**Estimated Fixed Effect**	**95% One-sided CI**	***p*-value**
100%, **k1**, st1	0.09 ± 0.04	[0.02 , 1]	0.02*
100%, **k3**, st1	0.07 ± 0.04	[0.0008 , 1]	0.04*
50%, k2, st1	-0.002 ± 0.04	[-0.07 , 1]	0.52
25%, k2, st1	0.006 ± 0.04	[-0.06 , 1]	0.43
10%, k2, st1	0.09 ± 0.04	[0.02, 1]	0.01*
100%, k2, **st0.6**	-0.03 ± 0.04	[-0.1, 1]	0.7
100%, k2, **st2**	0.13 ± 0.04	[0.05 , 1]	0.002*

*Significant *p*-values with 95% confidence

The above findings from the 15 uncorrelated features (Table 6-4) are also in accordance with the visualizations of CCC values for all 93 radiomic features shown in Figure B-4, Figure B-5, and Figure B-6 in Appendix B and the summary shown in Table 6-3; those conditions that received significant positive influence by GAN application (Table 6-4) had a larger number of radiomic features with $CCC_2 \geq 0.9$ and larger average CCC difference (Table 6-3).

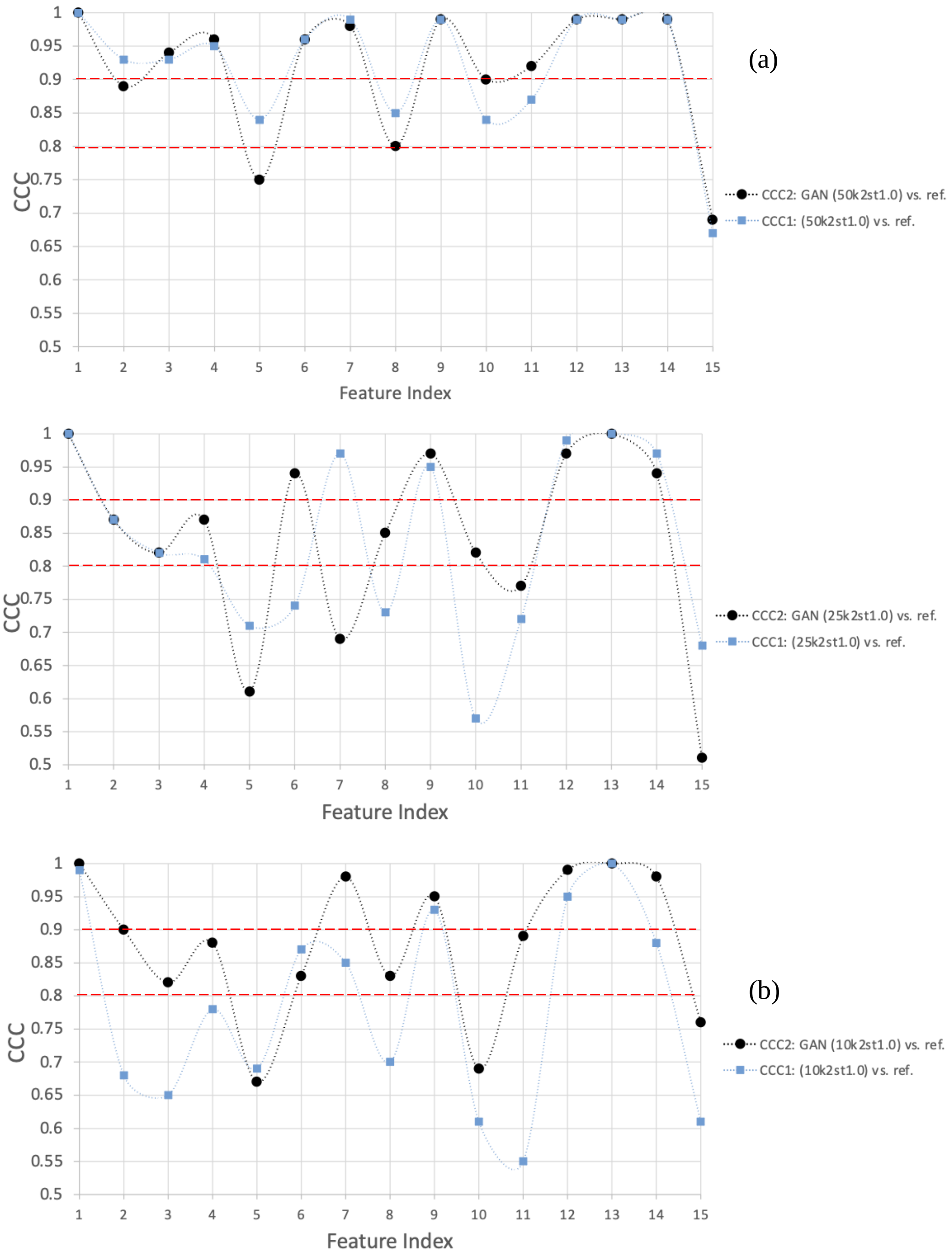


Figure 6-12. CCC_1 for 15 uncorrelated features and the corresponding CCC_2 values for non-reference conditions with different dose level than reference. Harmonization of (a) 50% dose, (b) 25% dose, (c) 10% dose. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

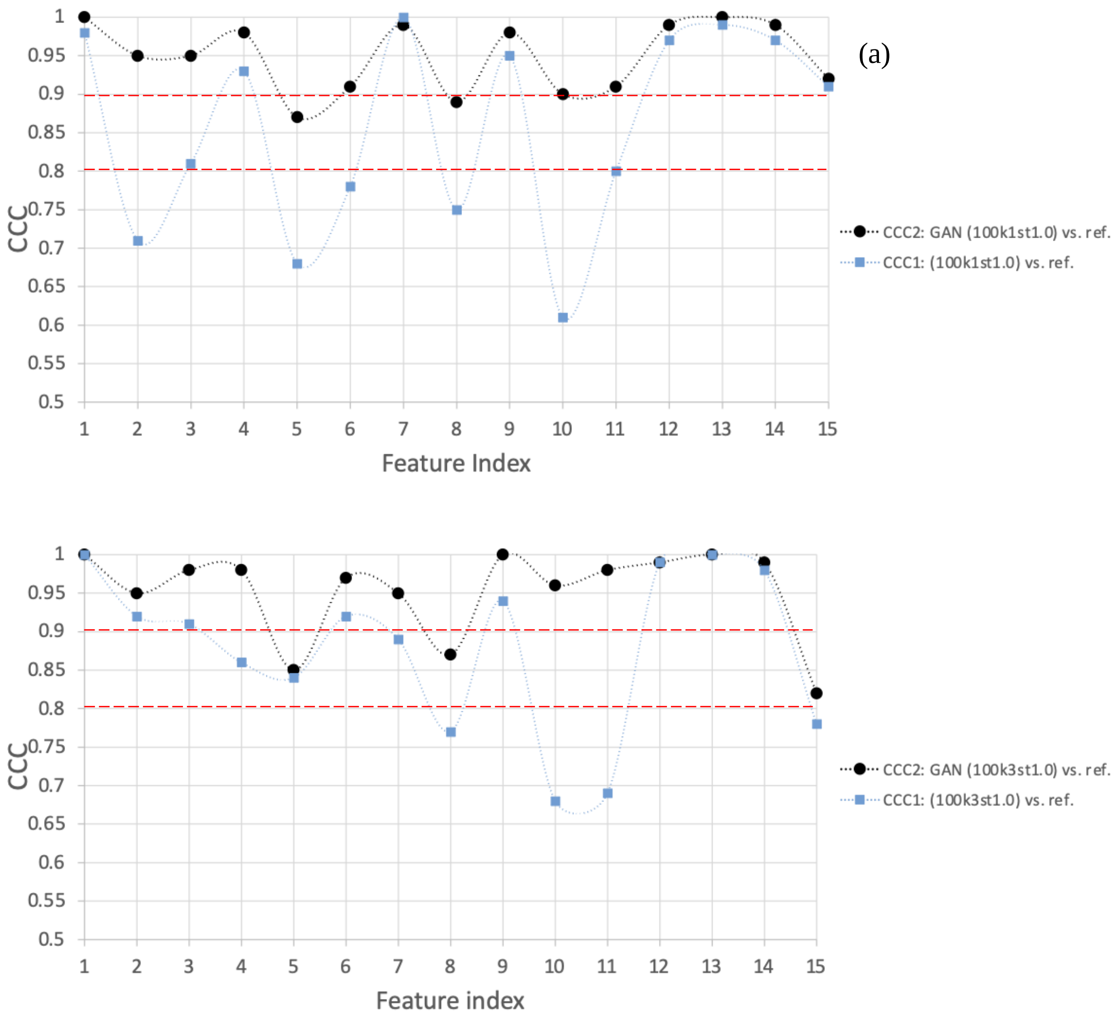


Figure 6-13. CCC_1 and CCC_2 values for conditions with different kernel than reference. Harmonization of (a) smooth kernel, (b) sharp kernel for 15 uncorrelated features. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

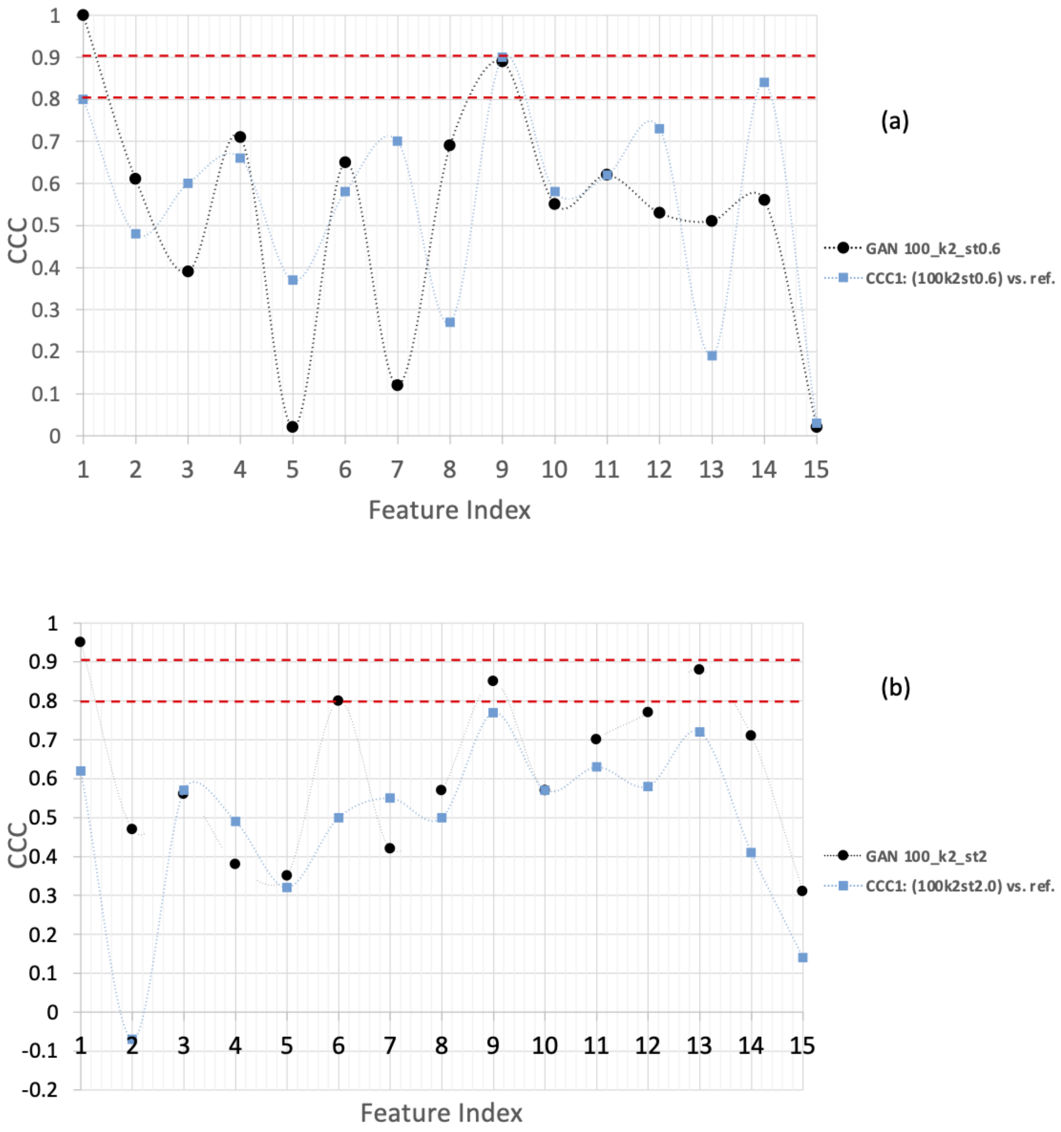


Figure 6-14. CCC_1 and CCC_2 values for conditions with different slice thickness than reference. Harmonization of (a) 0.6mm thickness, (b) 2mm thickness for 15 uncorrelated features. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

Out of 93 radiomic features (excluding shape and size), some features were negatively impacted by GAN application in few non-reference conditions where $CCC_2 < CCC_1$. However, none of these decreases in CCC values were significant (p-value>0.05, for one-sided lower-tailed t-test). Agreements for some radiomic features decreased (i.e., $CCC_2 < CCC_1$) in condition with 0.6 mm and 2mm. Out of 93 radiomic features, 49 and 26 features had $CCC_2 < CCC_1$ at 0.6mm and 2mm conditions, respectively. Though, only a few (<10) radiomic features had $CCC_2 < CCC_1$ in the harmonization of kernel variations or the condition with 10% dose. For condition with 50% and 25% dose, respectively, 30 and 24 out of 93 had $CCC_2 < CCC_1$ but these differences were not significant, and average CCC differences were minimal (-0.1% and 3.8% according to Table 6-3). This can also be visualized in Figure 6-15.

As an example, Figure 6-12(a) shows $CCC_2 < CCC_1$ after harmonization of the condition with 50% for *GLRLM Long Run Low Gray Level Emphasis* feature. Other features that are highly correlated with this feature also had their CCC values decreased after GAN application in 50%, 25%, or 10% dose level. Example of these features include: *GLDM Low Gray Level Emphasis*, *GLRLM Low Gray Level Run Emphasis*, *GLRLM Short Run Low Gray Level Emphasis*.

To give two other examples, CCC_2 is less than CCC_1 for *GLRLM Long Run Low Gray Level Emphasis* at all the three conditions with different dose level than the reference (10%, 2%, 50%), and *GLSZM Large Area Low Gray Level Emphasi*s had poor CCC_2 at conditions with smooth kernel or conditions with 50% and 25% dose. Such radiomic features that don't receive improvement of reproducibility may be considered as non-reproducible features or may be considered as features that require other techniques for their harmonization. Though the number of these radiomic features was not noticeable.

In Appendix B (section B.2), we have provided an additional explanation about the model trained on the non-reference condition with 50% dose level versus few other conditions.

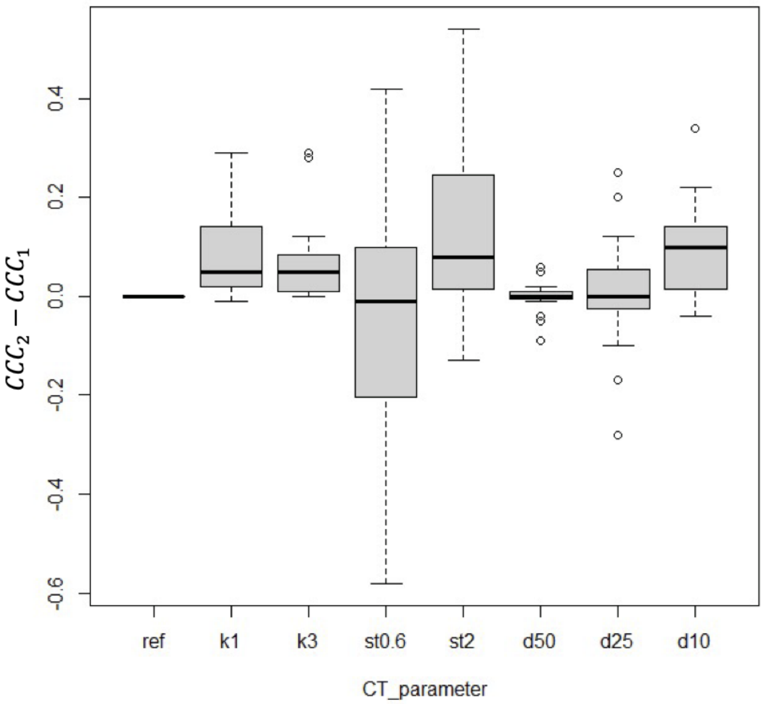


Figure 6-15. Boxplots showing the range of differences in CCC of the 15 uncorrelated features before and after GAN application. Each condition is identified by the level of the varying CT parameter. e.g., d50: (50%, k2, st1), d25: (25%, k2, st1), d10: (10%, k2, st1), etc. The horizontal lines show the median.

6.3.4 Other Image Quality Metrics

Table 6-5 provides the average image quality metrics of PSNR and SSIM between the reference and non-reference images across the cases in the held-out test set. For comparisons, the average is shown before and after the GAN application. The 32x32x32 sub-volume around the nodule was used to measure these quality metrics. According to this table, the GAN application enhanced both quality metrics for all conditions (p-value<0.05 with 95% confidence in Student's t-test). This indicates that the transformed non-reference images were more similar to the reference images compared to unharmonized images. Some conditions already had high-quality metrics (e.g., SSIM for conditions with k1 and k3), though, for other conditions, such as images at a low dose of 10%, GAN harmonization had a larger impact.

For images at the condition with 0.6mm thickness or 25% dose, even though image quality metrics have improved, GAN harmonization did not enhance the reproducibility of radiomic features significantly. This points to the fact that while a general-purpose metric assessment can be informative, only a radiomic task-based metric can provide an understanding of whether the application of the GAN model can be helpful for the harmonization of radiomic features.

Table 6-5. Average of quality metrics of images at each non-reference condition

Non-reference Condition	Before GAN		After GAN	
	PSNR	SSIM	PSNR	SSIM
100%, **k1**, st1	37.43	0.95	40.16	0.98
100%, **k3**, st1	34.94	0.93	40.67	0.98
50%, k2, st1	35.48	0.92	37.05	0.94
25%, k2, st1	30.25	0.80	32.52	0.87
10%, k2, st1	25.9	0.62	31.52	0.87
100%, k2, **st0.6**	20.87	0.45	25.17	0.73
100%, k2, **st2**	18.43	0.37	26.30	0.68

6.4 Discussion:

The goal of this chapter was to investigate GAN as a mitigation technique to remedy the poor reproducibility of CT radiomic features of lung tumors between images that are acquired at different CT protocols (by using a different range of dose levels, slice thicknesses, and kernels).

This study was a proof-of-concept study, and we explored the potential of image-to-image translation by a GAN model in harmonizing variability of lung nodule radiomic feature values between reference and non-reference CT images. After some tuning exercises to select the preferred GAN architecture and hyperparameters, an experiment was performed to test the improvement of reproducibility of lung nodule radiomic features between reference and non-reference CT conditions. This experiment involved training and testing of a slightly modified 3D Pix2Pix[132] GAN model to transform non-reference CT images of a set of lung nodules into images that have similar radiomic features to the radiomic features of their corresponding CT images at reference condition.

We assessed the harmonization of radiomic features between a reference condition with 100% screening dose, medium kernel (k2), 1mm thickness (st1), and seven non-reference conditions with a different dose, kernel, or slice thickness. The reference condition was chosen as a protocol most similar to clinical protocols for lung cancer screening scans. For each set of non-reference images, a separate GAN model was trained and then was used to harmonize the images of the corresponding non-reference condition. Therefore, seven GAN models were trained, and while each model used the same architecture, two of the GAN models (corresponding to the conditions at 25% and 10% dose) had a better performance when a slightly different loss function was used during the training.

The main results from this experiment were that firstly, the transformation of images via the utilized GAN model could remedy the variability of a large number of lung nodule radiomic features due to variation of CT parameters of the kernel, dose, and slice thickness at certain levels when they vary one at a time. Harmonization of the effect of kernel variation (at k1 and k3), the effect of dose reduction (by 10%), and slice thickness variation (from 1mm to 2mm) significantly improved (p <0.05) the agreement of non-reference radiomic features in comparison to the reference CT condition. Secondly, this image-to-image translation enhanced the overall quality of images acquired with different ranges of slice thickness, dose, and kernel in comparison to a reference condition. Though, the improvement of the overall quality of images (as measured by general-purpose metrics of SSIM and PSNR) did not necessarily indicate the efficacy of the technique in the task of reducing the variability of the quantitative measurements provided by radiomic features. Hence, to understand the potential of a GAN model in medical imaging applications, a task-based assessment, such as the approach used in this study, is essential. Finally, we found that tuning and modification of the previously proposed GAN model[132] better fit the data and the scope of our problem.

Both before and after the GAN application on each set of non-reference images, pairwise CCC was used to understand the agreement of radiomic features. In our analysis, we first examined the agreement of all radiomic features before and after the GAN application. But since we found that the CCC values of several features change in correlation with each other, in the testing significance of the impact of GAN mitigation, we extracted a set of uncorrelated features. This allowed us to avoid being biased by the number of colinear radiomic features in the data.

We initially explored two different GAN models of Pix2Pix[132] and CycleGAN[127]. Pix2Pix involves paired image-to-image translation as oppose to CycleGAN, which involves unpaired image-to-image translation. The paired approach benefits from the correspondence between pair of images at the source and target domain, but the unpaired approach can be more difficult since there are no matching subjects in the target and source domain and potentially requires a larger training dataset. On the other hand, the unpaired approach has its own advantage as it gets trained in an unsupervised fashion without the need of matching pairs, and therefore can be useful in scenarios where CT image data of subjects is only available at one CT condition instead of multiple CT conditions. Though, our preliminary results were superior for the Pix2Pix model. For example, visual comparisons between images generated by CycleGAN showed hallucination[6] and misalignments between images of the generated and reference image of the same subject. This reflects the difficulty of the unpaired learning task and the requirement of larger training set.

[6] Hallucination is a term used in the GAN community to refer to generation of surreal objects in image

Another preliminary exploration in building the GAN model indicated the advantage of 3D GAN model over 2D GAN model. As a result, we focused the study on 3D Pix2Pix GAN architecture.

6.4.1 Summary of Results

Overall, the GAN was able to significantly ($p < 0.05$) mitigate the impact of kernel variations from medium to sharp or medium to smooth. For conditions with different kernels than reference, the number of radiomic features that have strong agreement with reference condition improved by more than 50%.

In mitigation of the impact of dose reduction, GAN harmonization was most effective ($p < 0.05$) in remedying variability between reference and the condition with 10% dose level. The condition with 50% dose level was already in harmony with the reference condition, and the condition with 25% dose level received moderate CCC improvements (but not significant). Interestingly, during model tuning, we found that for the two low dose levels of 25% and 10%, the inclusion of the radiomic-based perceptual loss in the cost function better helped the GAN model to preserve the high-level information and the radiomic features in the nodule image. Therefore, the finalized model for these two conditions was penalized by this loss function in addition to L1 loss and adversarial loss functions.

In harmonization of the impact of slice thickness, the GAN application improved the agreement of radiomic features at 2mm slice thickness. In the mixed effect model, we observed a significant improvement of radiomic feature agreements at the condition with 2mm thickness. Accordingly, we observed a large improvement in average CCC values (by 60%) for this condition. Moreover, in Figure 6-11, we can visualize how the appearance of the nodule has changed after the application of GAN on the non-reference condition of 2mm. This figure demonstrates that the GAN model was able to output an image that looks more like the image at thinner slice thickness (i.e., the reference at 1mm). Specifically, it appears that the GAN model has learned the differences between 1mm and 2mm thicknesses due to the volume averaging.

On the other hand, still, the number of radiomic features that became reproducible at 2mm condition (i.e., with CCC_2 ≥0.9) was the least compared to other conditions (Table 6-3). Though this should be noted that the number of features identified as reproducible depends on the choice of CCC threshold, and we have chosen 0.9 as a conservative choice. However, in the future, it should be evaluated to understand what cut-off values are meaningful in the context of the utility of radiomic features. According to the designated and task-based cut-off value, we can decide whether the current GAN model is sufficiently alleviating the variability of radiomic features due to slice thickness.

Furthermore, to compare the results of GAN harmonization on the two conditions with 0.6mm and 2mm with a simple slice thickness resampling, we resampled the non-reference images and their segmentation masks at these two conditions with 0.6mm and 2mm thickness to 1mm and calculated the radiomic features. CCC values were calculated and compared to the results from the GAN model (shown in Table B-4 in Appendix B). The comparisons demonstrated that the resampling did not improve the reproducibility of radiomic features at 2mm. However, at 0.6mm, resampling improved the reproducibility slightly better than GAN. Therefore, it seems that a combination of slice thickness resampling and GAN may result in better improvement of reproducibility of radiomic features.

In assessing the quantitative image scores, we observed that for the conditions at 0.6mm or 25% dose, while the PSNR and SSIM enhanced significantly ($p < 0.05$) the CCC values among radiomic features did not improve. This points to the importance of having a task-specific metric to ensure a fair judgment for the potential of the model based on the task of interest.

As was described earlier in the results section, in the harmonization of CT parameters other than slice thickness, a few radiomic features were negatively impacted by the GAN application. The negative impact was not significant, and therefore the application of GAN was not harmful.

For the condition with 50% dose that already had highly reproducible features, the GAN application may not be necessary.

On a separate note, we made comparisons between the performance of a GAN model and a model that was only consisted of the generator used in the GAN model to understand the effectiveness of the addition of the discriminator in the GAN model. Our results indicated that the GAN model exceeded the harmonization of radiomic features as measured by CCC. Therefore, the addition of the discriminator was identified as useful for the task of radiomic feature harmonization in our study.

6.4.2 Contribution and Comparisons to Literature

This study takes a step toward the identification of effective GAN architectures that can have practical usefulness in the medical imaging task of radiomic feature harmonization. This is important because the availability of a mitigation technique, such as a GAN model, that offers image standardization and/or radiomic feature harmonization across different CT protocols will allow for the integration of multicenter image datasets in radiomic studies without being affected by inconsistencies due to variation of CT image acquisitions. Once images acquired at different CT protocols are standardized, the calculated radiomic features will not be biased or impacted by the acquisition shift. This can further result in an acquisition of large and comprehensive datasets in radiomic studies.

While the utilization of GAN models has been recently introduced in medical imaging applications, further investigations are needed to verify the promise of this technique, specifically in the context of radiomic harmonization. Our findings demonstrate that there is a potential for application of the 3D Pix2Pix GAN model in mitigating variabilities induced by acquisition shift in chest CT scans and in lung nodule radiomic feature values.

Another contribution of this study is that in addition to testing the original framework of the Pi2Pix model in the context of radiomic harmonization, we explored potential modification of the original configuration by trying a number of techniques that can improve the training balance between the discriminator and the generator (such as handling discriminator overconfidence, scheduling training, etc.) and potentially improve radiomic feature harmonization. As a result, we introduced a task-specific perceptual loss by borrowing the idea from Johnson *et al.*[145] and Ouyang *et al.*[128]. While the L1 loss in the cost function penalized the model for structural dissimilarity and misalignment, the perceptual loss function enabled the model to preserve high-level information from nodule tissue. The addition of the task-specific perceptual loss function in the GAN models improved the consistency and similarity of generated non-reference images to reference images for two of the conditions (10% and 25% dose levels) by ensuring the preservation of high-level features. This suggests that the utilization of the quantitative medical imaging biomarkers (radiomics) in the form of a task-specific loss function can improve the generator's performance.

Very few studies have tested GAN application in the context of radiomic harmonization. Liang *et al.*[130] explored the application of a different framework for cGAN models by introducing a different training procedure than Pix2Pix. In their assessment of radiomic feature variability, they explored the impact of variation of reconstruction kernel and slice thickness by assessing eight radiomic features measured from random 2.5D patches of soft tissue in CT images. They observed improvement of relative error of radiomic features between different CT protocols after applying GAN. In our study, we further expanded this exploration on a larger number of 3D radiomic features (instead of 2.5D) and included an investigation of harmonization of CT parameter of dose as well. Another deviation between the studies is the patient population as we expanded the investigations into the potential of the GAN approach into lung cancer screening patient populations that are acquired with low dose scans that result in different image qualities. Our results on the overlapping set of eight features with this study were in agreement for the transformation of the CT condition with a smoother kernel and lower dose level than reference. Though, one important difference is that the low dose levels explored in our study are at much lower CTDIVol (2mGy to 0.2mGy), which results in very different and poor quality compared to the images in that study.

Wei *et al.* [129] explored the application of a different GAN framework and a different cost function (hinge loss function) in reducing the variability of radiomic features across different CT conditions with the different dose levels and slice thickness. Radiomic features were measured over 3D patches centered on nodule and assessed effectiveness of GAN approach via normalized error and paired Wilcoxon signed-rank sum test. One important difference between this study and ours is that for feature calculation, we focused on the nodule region only rather than the whole area depicted in the image patches. The hypothesis for radiomic features is that they can reflect the biological characteristics of tumor tissue. Hence it matters to specifically ensure the reliability of the features calculated from within the nodule tissue.

Finally, in assessing reproducibility and harmonization by GAN, we based our method and conclusions on the metric of CCC, which assesses both precision and accuracy between radiomic features of each nodule. Since this metric measures deviations from the 45° line of identity, it is very sensitive to small deviations and serves as a more accurate agreement metric than other metrics or methodologies[59,149].

6.4.3 Practicality and Future work

While this study suggested a set of improvements to be applied to the Pix2Pix GAN model for the purpose of lung nodule CT image generation and indicated a potential for GAN technique in standardization of CT images to remedy radiomic feature variability, we should consider practical aspects of the model in the context of standardization of CT images acquired with different protocols. First of all, in this study, we only focused on mitigation of impact of dose, kernel, and slice thickness variation within a certain and limited range (seven conditions only). In doing so, we built one model for the transformation of images at each CT condition, such that each condition had only one CT parameter inconsistent with the reference condition. Therefore, we tested the mitigation potential of the GAN model in a predefined and controlled framework. Though, in practice, we will have a different scenario. For example, in a multicenter image dataset, we may have images at the wide variation of the kernel, dose, and slice thicknesses (hence we may have even larger variability in CT conditions). Therefore, for practicality, it may seem that instead of building different GAN models for each CT condition, it will be more optimal to build one general-purpose GAN model that learns data distribution from all CT conditions and generates transformed non-reference images that match with reference. While this is worthy of exploring, two considerations should be made. Firstly, in our preliminary investigations, we explored the building of a general-purpose GAN model to transform images from four CT conditions, and we found out that in doing so, the model performed better when during the training, images were labeled by their CT condition, which means that information (metadata) about dose level, slice thickness, or kernel of the image was fed into the network by adding channels to the input image. Secondly, since our results indicated that for each CT condition, use of a slightly different hyperparameter set or configuration was more optimal, it may be beneficial that instead of having one general-purpose model for all conditions, we build one model for mitigation of variation of each CT parameter; therefore, we build three models for mitigating three CT parameters of dose, kernel, and thickness. By labeling input images with metadata, the model for each CT parameter will be exposed to images from a range of different CT parameter values (e.g., it will be exposed to images at dose levels of 100%, 50%, 25%, 10% as opposed to only receiving 50% dose images). This approach may enable the model to obtain an understanding of the relationship between the variation of the CT parameter within a range and its impact on the image quality rather than being exposed to a discrete data point (e.g., one dose level) in the CT parameter space. So, with such a model, and as opposed to the model trained in this study, we may be able to also achieve optimal mitigation of images with a CT parameter that the model has not been trained on but lies within the range that the model was exposed to (e.g., the dose level of 75%). It should be noted that if GAN model building is implemented through the latter proposed approach (three models for three CT parameters), for the transformation of images that have more than one inconsistent CT parameter compared to the reference, an ensemble solution shall be implemented that combine the GAN models in transforming the input images.

Another important consideration is the selection of the reference condition. We selected 100% screening dose, medium kernel, and 1mm thickness as a reference since it is the most common CT protocol. If a different combination of CT parameters is chosen as a reference, the results from the GAN model may have a different trend. For example, if the reference condition is a condition with a lower dose level, then the mitigating (transforming high dose to low dose) will be reverse to the mitigation performed in this study (transforming low dose to high dose).

Moreover, for the purpose of harmonization of radiomic features in our study, it made sense to limit the image inputs to a sub-volume around the nodule as it reduced the dimensionality of input, the complexity of the network. We also found that both of the discriminator and generator networks were able to achieve acceptable harmonization performance with few convolutional blocks. However, in the standardization of images for other purposes such as nodule detection, choosing small-size input may not be suitable as it may be needed to normalize the whole lung CT scan. Accordingly, the network may need more complexity and more depth (more convolutional blocks). Building a medical image standardization via a paired image-to-image translation GAN (similar to the one used in this study) requires the availability of pairs of corresponding images from the same patients for training and evaluating the model. However, in the standardization of images across a wide variety of conditions, this may not be feasible to acquire patient images at all the variation of CT conditions. So, a paired framework may face a challenge in the lack of availability of representative data. Though, an unpaired image-to-image translation framework, such as CycleGAN[127], can overcome this challenge as it can learn the data distribution and the mapping function between different CT conditions without the need for tightly correlated pairs.

Future steps toward building and accomplishing a high-performing GAN model for standardization and mitigation of radiomic feature variability can include further tuning of model hyperparameters and configurations as well as the implementation of further evaluation strategies. Among potential techniques that can be applied to improve the current state of the GAN model used in this study can be a modification of the perceptual loss by applying a pre-trained CT-based CNN classifier to penalize the model both on the similarity of the extracted features and the similarity of the activation maps between generated images and the target images. Additionally, in scenarios that the discriminator model is outweighed by the generator, it might be useful to also use the perceptual loss function to the discriminator's cost function so that the discriminator also obtains the perception from the radiomic features and high-level information of the image. This may result in more meaningful feedback to the generator, which can, in the end, result in better weigh updates for the generator.

Our prior studies described in Chapter 4 and Chapter 5 showed that the slice thickness variation impacted a large number of radiomic features due to the presence of nodules in fewer slices and due to volume averaging with the background. Although the results of the GAN model tested in this study are not perfect for harmonizing slice thickness variation, we can see the potential in remedying the radiomic feature variabilities, perhaps with future modifications of the GAN model. As was previously mentioned, a potential approach to obtain a GAN model that can better harmonize variability due to variation of slice thickness could be resampling and bilinear interpolation of images to the reference slice thickness before feeding the images into the GAN model. In our experiment (Appendix B, Table B-4), resampling alone did not result in a considerable improvement of reproducibility, but in combination with GAN, we may be able to better control the variability of radiomic features due to variation of slice thickness.

Another potential approach could be the usage of super-resolution techniques, such as the super-resolution GAN (SRGAN)[115], for mitigation of slice thickness variation. Since this model has shown potential in recovering fine texture in images, it may be useful to match the content in images with different slice thicknesses. Karnewar *et al.*[150] proposed a model called MSG-GAN in which the generator generates images at multiscale with different resolutions, and the discriminator becomes a function of multiscale images. While the model is based on a 2D task, this concept can be modified and applied in a manner to get trained and handle multi-thickness problems. Exploration of this idea is, therefore, an interesting question to be investigated in the future.

Other regularization techniques that have been proposed in regularizing the training of the GAN, such as spectral normalization[151] in which the convolutional kernels are normalized, may also help improve the performance of the GAN model in generating images with better quality compared to the reference images.

Further evaluation techniques can be derived in a task-based manner to not only ensure the quality of image or reproducibility of radiomic features but also ensure that mitigations and improvement of reproducibility are reflected in the clinical task of interest. For example, if a diagnosis is available for a dataset of patients with lung nodules scanned at different CT protocols, and a pre-trained classifier is available that detects cancerous nodules, we can evaluate the impact of GAN application by assessing the prediction performance of the classifier before and after training. One other approach that may provide task-based evaluation is a clustering of images before and after application of GAN to understand whether clustering metrics improve or whether non-reference conditions get grouped with reference condition images after GAN transformation.

Another important consideration in this study is the choice of CCC threshold in interpreting strong vs. moderate agreement. We initially chose a CCC threshold of 0.9 according to the findings of McBride[60] and suggestions in the radiomics research[61,152]. Currently, there are no means to obtain a meaningful and task-based cut-off value that suggests that if CCC values fall below this threshold, the predictability of the radiomic feature will be impacted negatively. While this remains as future work, we have chosen a conservative scenario (by enforcing a cut-off value of 0.9). Though, it may be that meeting a CCC of 0.8 is also sufficient for the radiomics model to behave consistently on CT images with different protocols. Hence, with the variation of the choice of threshold, the results in Table 6-3 may change. We provided visualization of the CCC values (Figure 6-12 to Figure 6-14) to enable comparisons and assessments of the impact of GAN on agreement of radiomic features without being constrained by the cut-off value. Furthermore, the CCC metric itself serves as a conservative and individualized metric. CCC is very sensitive to deviation of radiomic feature values of the reference and non-reference conditions. Hence, the choice of the metric in this context can also be important. In situations where the reproducibility of radiomic features needs to be restricted to ensure consistency of model predictions, we can choose the more conservative metric, while in situations that reproducibility requirement is less stringent, we can rely on other metrics.

6.4.4 Limitations

One of the limitations of this study was the lack of availability of a large held-out test set with nodule segmentation that allows for evaluating of mitigation performance of the GAN model in a more representative dataset. In addition, as was discussed earlier, the models trained in this study were limited to a sub-volume of CT images around the lung nodule. While this has its own benefits, it may not be practical to use in situations where the location of the nodule is unknown, and we need to perform image standardization before detection of the nodules.

CT image datasets may suffer from inconsistencies in various parameters and acquisition techniques such as field of view, resolution, pitch, scanner type, etc. Our study only focused on three CT parameters of dose, kernel, and slice thickness. Moreover, due to time consumption and computational expenses, we only acquired images and trained GANs at a limited range of these three CT parameters. However, we tried to keep other variables consistent in our dataset.

Another limitation is that the current study does not address the harmonization of CT conditions that have more than one varying CT parameter compared to the reference.

The radiomic feature quantifications for the perceptual loss calculations were set to use CPU cores instead of GPU cores. Hence during the model training, for each update, the GPU and CPU were in communication to transfer results. This resulted in a bottleneck that causes the models trained with perceptual loss to become inefficient. The training time for these models at least took more than 12 hours were the training of the models without perceptual loss took up to 8 hours. If the radiomic feature quantifications were implemented in a way that they use GPU cores, this bottleneck could be removed.

Finally, while we performed task-specific evaluation by assessing the similarity of radiomic features between reference and non-reference images, we were unable to address whether the improvements yielded by the application of GAN will impact the robustness of a prediction task such as diagnosis.

6.5 Conclusion

In this study, we trained and tested the 3D version of a popular GAN model, Pix2Pix, to assess its potential in transforming seven non-reference CT images of lung nodules into images that have radiomic features similar to reference CT images. Seven GAN models were trained for the harmonization of seven non-reference conditions.

The GAN models were able to reduce the variability of radiomic features at either smoother or sharper kernel, the low dose level of 10%, or thick slice thickness of 2mm compared to a designated reference CT condition.

In harmonization of dose reduction, the inclusion of a radiomic-based perceptual loss function into the generator’s cost function ensured better preservation of high-level content and better radiomic feature reproducibility.

We found that it is important to evaluate the image generation performance based on task-specific metrics rather than relying on general-purpose image quality scores. This study suggests a potential for GAN models in remedying the poor reproducibility of radiomic features. However, previously proposed non-medical GAN architectures can require further configuration tuning to better fit the data and the radiomic context to achieve more optimal harmonization.

Chapter 7 Application of ComBat Technique in Harmonization of Lung Nodule CT Radiomic Features

7.1 Introduction and Rational for Using ComBat Technique

In the previous chapter, we discussed the application of Generative Adversarial Networks (GAN) in the standardization of lung nodule CT images to mitigate radiomic features. In this chapter, we discuss applying a different technique in harmonizing radiomic features that, unlike the GAN technique, is directly applied to radiomic features instead of the images. We first performed a proof-of-concept experiment with ComBat on data from patient cohort 2 (described in Chapter 3) that is different and larger than the data used for making conclusions about GAN harmonization in the previous Chapter. We will then compare the two techniques of ComBat and GAN on the same dataset (i.e., patient cohort 6) in Chapter 8.

ComBat technique is a data-driven technique. Therefore, we can apply it to a radiomic feature dataset that has been calculated from images acquired with different scan protocols to estimate and correct the CT effects and acquisition shift effect on radiomic features.

Johnson *et al.*[101] proposed the ComBat (Compensating for Batch Effect) method to correct the variabilities induced by differences in equipment, time of the experiment, labs, etc., between different batches of samples (i.e., batch effect). Batch effects that arise due to non-biological conditions cause the data not to be directly comparable. Therefore, it is required to adjust the data for these batch effects before combining multiple batches. The ComBat technique is a method from the family of location and scale (L/S) adjustment techniques, though Johnson *et al.* [101]

extended the application of Empirical Bayes[153] in building this method. Since then, the method has been widely used in genomics studies. ComBat has been shown to perform well in datasets with a small sample size and high dimension[154]. For example, one study[155] used synthetic expression microarray data with simulated batch effect to compare ComBat to different adjustment methods (surrogate variable analysis, mean centering, etc.) by assessing the precision and accuracy between identical samples and classification performance before and after harmonization. Their study showed that ComBat not only outperformed other methods, but it also performed robustly in small size data, removed variations due to batch effects, and increased the classification performance of a prediction model. Robustness in small size data is an essential characteristic of this technique for studies with limited data.

ComBat harmonization has a clear difference to the normalization approach, which divides data points by a reference value or the z-score normalization approach; these normalization techniques aim to harmonize deviations from an overall mean or standard deviation. In contrast, ComBat estimates and targets removal of variances due to differences in batches. In addition, ComBat seeks to maintain the meaningful biological differences between samples.

Variability due to the batch effect in genomic data is similar to the variability of quantitative medical imaging methods due to the variation of image acquisition protocols. Hence, ComBat has also been used in medical imaging research studies to mitigate bias and variability due to unwanted and non-biological conditions such as variation of imaging protocols, scanners, imaging centers, etc.[108]. For example, researchers successfully used ComBat for combining and harmonizing cortical thickness measurements in MR images across different scanners[156], or for adjusting diffusion tensor imaging data[157], or for removing site effects on functional connectivity measures in fMRI dataset[158].

Application of the ComBat technique has been recently introduced in the harmonization of radiomic features and has shown potential in mitigating variability of radiomic features in multicenter datasets. For example, in PET imaging, ComBat removed the multicenter effect in texture radiomic features and standardized uptake values (SUV)[109]. Similarly, it has shown promise in mitigating the variability of radiomic features due to variation of image acquisition protocols or across scanners. For example, this has been explored in MRI for brain tumors[159] and breast lesions[70], or in CT radiomic features of phantom data[110,160], liver[81], and lung cancer patients[110,160].

In removing unwanted batch effects, while it is essential to have high performance in adjusting unwanted effects, it is also crucial to avoid removing important information regarding the biological variability of samples. Various studies that have used ComBat have shown that the application of this technique avoids the removal of biological heterogeneity among samples; For example, Fortin *et al.* [156], in the harmonization of cortical thickness measurements, confirmed that biological variability is associated with age is well preserved with ComBat. Interestingly, studies have also shown that harmonization removed variabilities and yielded more discriminant predictors/models. In breast DCE-MR images, the application of ComBat successfully harmonized radiomic features across different field strengths and obtained an improved classification of benign versus malignant lesions[98]. Orlhac *et al.* [159] achieved a better classification performance in distinguishing prostate cancer grade scores (low risk vs. high risk) after using ComBat to harmonize MRI radiomic features.

Furthermore, by randomly assigning sham labels of prostate tumor Gleason score, the study[159] tested whether the application of ComBat on radiomic features of prostate cancer patients can result in false-positive predictions by a classifier. Their analysis showed that the ComBat application did not result in incorrect positive predictions, indicating preservation of data integrity after the ComBat application. These findings ensure that the ComBat method maintains data integrity and prediction power after harmonization.

Contrary to machine learning techniques (e.g., GAN), ComBat is a simple technique that, due to the application of Empirical Bayes, can be efficient even with a small sample of data. Moreover, ComBat does not require original medical images to harmonize multicenter datasets and works directly with the radiomic feature data. In summary, this tool holds some promise for the standardization of quantitative medical imaging techniques against adverse impacts of scan acquisition and equipment differences.

Since the ComBat technique has shown promising results in mitigating variability of radiomic features, we aimed to verify further its ability in improving the reproducibility of CT radiomic features of lung nodules across different CT acquisition and reconstruction parameters. To this aim, we tested the harmonization of a set of well-known radiomic features calculated from lung cancer screening scans across eight different CT conditions consisting of varying dose, weighted Filter back Projection (wFBP) reconstruction kernel, and slice thickness. Agreement of radiomic features between seven non-reference CT conditions and a reference CT condition were assessed before and after ComBat harmonization.

Our goal was to test the ComBat method in harmonizing radiomic features of lung cancer screening patient population, which has not been investigated before.

Other studies have also investigated the application of ComBat in CT radiomic features; however, there are apparent differences between our study and these studies. For example, Foy *et al.* [81] investigated ComBat application in a different anatomic setting (normal liver scan). While other studies have investigated harmonization in diagnostic lung CT scans of lung cancer patients, we have expanded this investigation into the different populations of lung cancer screening patients acquired at a different CT image protocol. The low-dose protocol of screening scans results in different image quality and noisier images than high-dose diagnostic scans. Hence, in addition to the CT parameters investigated in the works by Orlhac *et al.* [160] and Mahon *et al.* [110], we have assessed the removal of impact of dose level variation on radiomic features. Dose level variation is simulated in terms of tube current variation (refer to Chapter 3 for more details). Another difference between the present study and the studies mentioned above is that by using a modified version of the ComBat technique (proposed by Stein *et al.* [161]), we aimed to test the alignment of non-reference radiomic features with a gold standard (i.e., reference radiomic features) rather than with a pooled grand mean. Alignment of radiomic features to a pooled grand mean involves adjusting both non-reference and reference (the gold standard) radiomic features. The modification of the gold standard is not desired in situations where we need to harmonize the radiomic features to a specific reference CT condition.

Unlike Mahon *et al.* [110], our patient dataset had scans at eight different CT protocols in question. Hence our study has a balanced batch-group design. Furthermore, we only assessed mitigation of variability due to variation of one CT parameter at a time, which means that each of the seven non-reference conditions has all CT parameters similar to the reference condition except for one (kernel, dose, or slice thickness). This approach removes any confounding impact in our analysis. Additionally, our goal was to only test the potential of ComBat in the reproducibility of nodule radiomic features; hence unlike the study by Mahon *et al.* [110] that has studied radiomic features of regions of lung or vertebra tissue, our study was focused on radiomic feature measurements over segmented lung nodule regions.

We will describe the ComBat technique theory, our experiment, and the study results in the following sections.

7.2 Materials and Methods

7.2.1 Theory of ComBat

ComBat is an adjustment technique from the family of location and shift (L/S) methods, and it assumes that data points in each batch can be modeled by a location (mean) shift and a scale (variance). Therefore, by standardization of the mean and variances of batches, the data will be adjusted, and unwanted batch effects will be removed[101]. In our analysis, each batch represents a CT condition identified by scan dose level, reconstruction kernel, and slice thickness. Thus, for each radiomic feature (f) and its calculated value, Y_{ijf}, for sample j in condition (batch) i the L/S model is in the form of:

$$Y_{ijf} = a_f + X\beta_f + \gamma_{if} + \delta_{if}\varepsilon_{ijf} \qquad \text{Eq. 12}$$

a_f is the overall (mean) value for radiomic feature f. X is the design matrix for biological covariates of interest (to maintain biological variability), and β_f are the coefficients corresponding to X. γ_{if} and δ_{if} are the location (additive) and scale (multiplicative) parameters for the batch effect for batch i and feature f. The error term ε_{ijf} is assumed as $\sim N\left(0, {\sigma_f}^2\right)$.

The ComBat technique assumes that the conditions or batch effects have a similar effect on all features (e.g., higher variability, increased or decreased values, etc.); hence it pools information from the features in each batch to estimate the parameters of the model via Empirical Bayes (EB)[153]. The estimated parameters are then used to remove the batch effect and adjust the data.

The initial step includes removing features with missing values and standardization of feature values so that all features have similar overall mean and variance (i.e. they lie in the same range).

Standardization is performed by estimating a feature-wise estimate of mean ($\hat{\alpha}_{i,f}$) and standard deviation ($\hat{\sigma}_{i,f}$) for each condition using ordinary least square approach. It is then assumed that standardized data, Z_{ijf}, satisfies the following $Z_{ijf} \sim N\left(\gamma_{if}, {\delta_{if}}^2\right)$. Then the prior distributions on the batch effect parameters $(\gamma_{if}, {\delta_{if}}^2)$ are assumed to follow the forms of $\gamma_{if} \sim N(\gamma_i, {\tau_i}^2)$ and ${\delta_{if}}^2 \sim Inverse\ Gamma(\lambda_i, \theta_i)$. The hyperparameters $(\gamma_i, \tau_i, \lambda_i, \theta_i)$ are estimated via EB using the method of moments[162]. Johnson *et al.* [101] recommend that if these priors don't fit the data, a non-parametric approach without these priors can be applied to estimate the parameters. EB estimates of the batch effect parameters $({\gamma_{if}}^*, {\delta_{if}}^{2^*})$ are then used to adjust the data. To match the data with the reference condition (as a gold standard), we used the Combat version proposed by Stein *et al.*[161], by using the estimated feature-wise mean and standard deviation for the reference condition $(i = r)$.

$$
{Y_{ijf}}^* = \frac{\hat{\sigma}_{i=r,f}}{{\hat{\delta}_{if}}^*}\left(Z_{ijf} - {\hat{\gamma}_{if}}^*\right) + \hat{\alpha}_{i=r,f} + X\widehat{\beta_f} \tag*{Eq. 13}
$$

Where ${Y_{ijf}}^*$ will contain harmonized radiomic feature data.

The implementation of the ComBat harmonization tool was obtained from Fortin et al. [157] (https://github.com/Jfortin1/ComBatHarmonization). R programming language[163] with RStudio[164] software was used for harmonization.

To check whether the parametric prior distributions reasonably fit the data, the empirical distributions and the priors for the batch effect parameters $(\gamma_{if}, {\delta_{if}}^2)$ were plotted and were compared.

Since our study design has images from the same subjects at different CT conditions (batches), there was no biological variability between data at different batches; therefore, we set $X = 0$.

7.2.2 Data Used for Analysis in this Chapter

Chest CT scans of 134 lung cancer screening patients (cohort 2 described in Chapter 3) were used for this study. The dataset used for this study was described in Chapter 3. For each patient, images reconstructed at eight conditions were included in the study. Table 7-1 describes the range of these conditions. Like the previous Chapter, the condition with 100% screening dose, medium kernel (k2), and 1mm slice thickness was considered the reference condition.

Table 7-1. Description of CT conditions included in the study. (Values in bold show the reference condition)

% of Screening Dose	wFBP Reconstruction Kernel	Slice Thickness	Condition Identifier	Varying CT parameter
100%	**Medium**	**1mm**	**100%, k2, st1**	**-**
100%	Smooth	1mm	100%, k1, st1	Kernel
100%	Sharp	1mm	100%, k3, st1	
50%	Medium	1mm	50%, k2, st1	Dose
25%	Medium	1mm	25%, k2, st1	
10%	Medium	1mm	10%, k2, st1	
100%	Medium	0.6mm	100%, k2, st0.6	Slice Thickness
100%	Medium	2mm	100%, k2, st2	

7.2.3 Experiments

For all 134 cases, a single nodule was identified and contoured to form a region of interest (ROI). Similar to previous studies described in Chapter 4 and 6, for each case, three segmentation ROI were used for the three different slice thickness reconstructions. So, for each subject, the images with 1mm slice thickness had the same segmented ROI, but the images with 0.6mm and 2mm thickness had different ROIs. From each ROI, 93 radiomic features (names available in Table B-2) were calculated. This was performed for images for each case at the baseline condition and then repeated for images at each of the conditions in Table 7-1. The radiomic features that represented size and shape were excluded, as ComBat does not harmonize or modify nodule shape or size.

We made paired comparisons between the radiomic features calculated at non-reference and reference conditions before and after harmonization. Concordance Correlation Coefficient was calculated for each radiomic feature to measure the agreement of the radiomic feature between each non-reference condition an the baseline. The agreements before harmonization (CCC_1) and after harmonization (CCC_2) were compared.

ComBat was applied to harmonize radiomic features from non-reference conditions to match the reference condition (100% dose, medium kernel (k2), 1mm thickness). For each non-reference condition, only one CT parameter was varied, as shown in Table 7-1, and other parameters were kept fixed. The number of features that obtained higher agreement to the reference condition ($CCC_2 > CCC_1$) was counted. To understand how many features with an improved agreement meet the cut-off value of 0.9 and become reproducible, the number of reproducible features before and after harmonization was counted as follows: C_1 is the number of features with CCC_1≥0.9 and C_2 is the number of features with CCC_2≥0.9. Furthermore, the percentage difference of CCC for a condition with M=93 radiomic features was calculated by:

$$Average\ CCC\ difference = \frac{1}{M}\sum_{i=1}^{M}(\frac{CCC_2 - CCC_1}{CCC_1} \times 100) \qquad \text{Eq. 14}$$

7.2.4 Evaluation of ComBat Harmonization Performance

Similar to the previous Chapter, we hypothesized that the reproducibility of radiomic features improves after harmonization. We tested this hypothesis (Eq. 8 in Chapter 6) by fitting a mixed-effect model (Eq. 7 in Chapter 6) to $CCC_2 - CCC_1$ to understand whether ComBat has resulted in significant improvement of the agreements of radiomic features with baseline. As was mentioned in Chapter 6, due to an existing correlation among several radiomic features, the model was only fitted to the data from 15 uncorrelated radiomic features (described in Table B-1 of Appendix B).

7.3 Results

7.3.1 Data Distribution

Figure 7-1 shows the prior distributions for ComBat batch effect parameters as well as the empirical distributions (described in section 7.2.1). Visual comparisons show that the empirical distributions from data fit the priors reasonably well, especially for the γ parameter. Since some distributions of δ slightly deviate from the priors, we also tested the non-parametric ComBat method without these priors. However, the harmonization results did not change substantially. Therefore we report the results from the parametric ComBat harmonization. γ and δ refer to the location shift and scale parameter in Eq. 12. For visualization purposes, we have divided the plot of data distributions for seven condition into two rows in Figure 7-1.

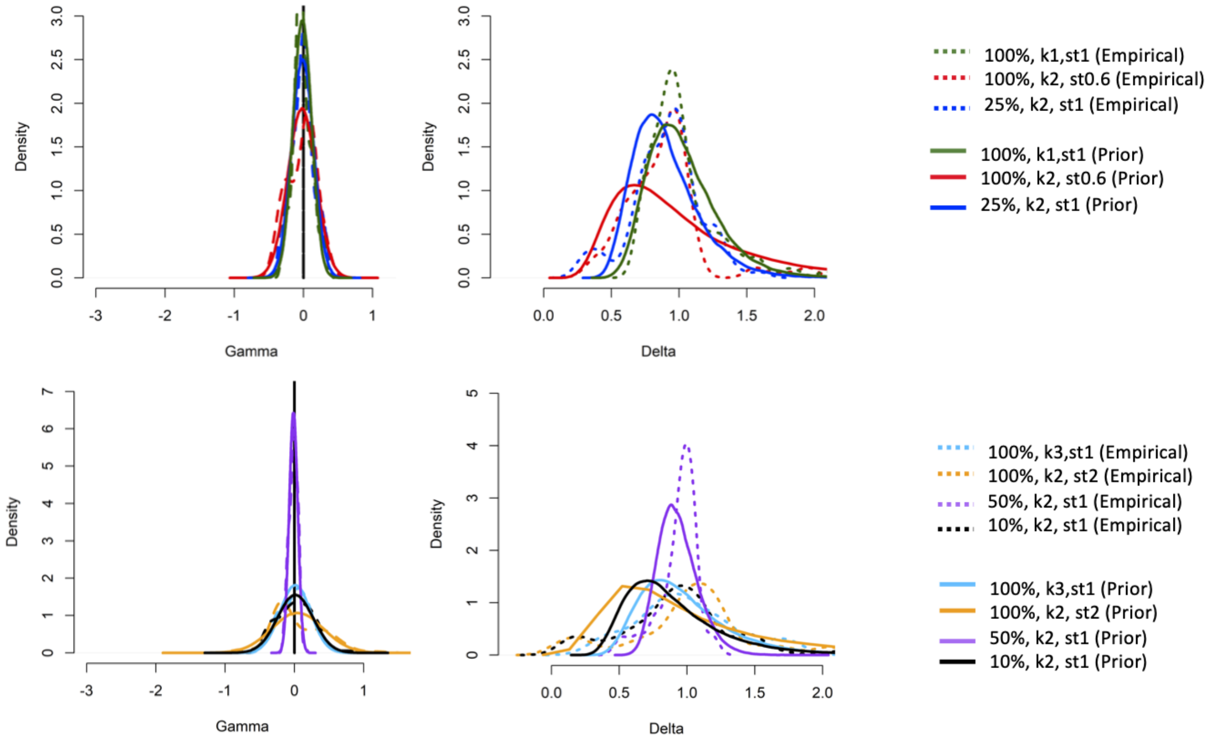


Figure 7-1. Prior distribution and empirical data distribution. Dotted lines are density plots from empirical values for features at each non-reference condition. Solid lines are the EB-based prior distribution that was used in data harmonization of each non-reference condition. For visualization purposes, we have divided the plot of data distributions for seven conditions into two rows.

7.3.2 Agreement Analysis Before and After ComBat Harmonization for all 93 radiomic features

Table 7-2 shows the results from CCC analysis of 93 radiomic features before and after ComBat harmonization on all of the 134 cases. According to Table *7-2*, after harmonizing kernel variation, the number of radiomic features that meet the cut-off (CCC_2≥0.9) increases compared to before harmonization (CCC_1≥0.9). Between the smooth and sharp kernel, the unharmonized condition with the smooth kernel (100%, k1, st1) had slightly more features with a strong agreement. But, after harmonization, the condition with a sharper kernel has more radiomic features harmonized (number of features with CCC_2≥0.9 increased by 42%) and a more considerable average CCC difference.

In harmonizing radiomic features in the condition with 50% dose, harmonization did not substantially affect the radiomic features; at the condition with 50% dose, many radiomic features were already in strong agreement with reference before the harmonization. According to Table 7-2, both of the conditions with 25% and 10% dose have more radiomic features with $CCC_2 \geq 0.9$ after the harmonization. Although the condition with 10% dose level has few reproducible features after harmonization, we see a considerable improvement in its CCC values.

In harmonizing variation of slice thickness, the condition with thick slice thickness had no reproducible features before harmonization and has only six features with CCC_2 ≥0.9 after harmonization. For the thin slice thickness, the count of radiomic features with CCC_2 ≥0.9 increased slightly. Although fewer features met the cut-off value (CCC_2 ≥0.9), there was a noticeable improvement in the CCC values after harmonization: 5.4% increase for 0.6mm and 16.4% for 2mm thickness.

Table 7-2. The percentage difference CCC before and after application of ComBat harmonization, the number of radiomic features (out of 93) for the specified ranges of CCC values, and the increased number of features with CCC≥0.9

Varying CT parameter	Non-reference Condition	Average CCC difference	$CCC_2 > CCC_1$	CCC_1 ≥0.9	CCC_2 ≥0.9	C_2 -C_1 *(%)*[a]
Kernel	100%, k1, st1	1.4%	62	58	68	10 (17)
	100%, k3, st1	4.3%	82	49	70	21(42)
Dose	50%, k2, st1	0.5%	24	81	82	1 (1.2)
	25%, k2, st1	2.9%	69	46	58	12 (26)
	10%, k2, st1	9.2%	75	20	35	15 (75)
Slice Thickness	100%, k2, st0.6	5.4%	66	31	40	9 (29)
	100%, k2, st2	16.4%	84	0	6	6

[a] C_1 is the count of features with CCC_1 ≥0.9 and C_2 is the count of features with CCC_2 ≥0.9.

Figure C-1, Figure C-2, and Figure C-3 in Appendix C visualize CCC_1 and CCC_2 values (vertical axis) before and after harmonization of each of the CT parameters for 93 radiomic features (horizontal axis). The data points that shift to higher values in the vertical axis indicate that the harmonization has improved the agreement of radiomic features with the reference condition.

7.3.3 Results from Hypothesis Testing on 15 uncorrelated features

Similar to the analysis in the previous Chapter, 15 uncorrelated features (names given in Appendix B Table B-1) were used for the hypothesis testing using a mixed-effect model (Eq. 7 in Chapter 6).

Table 7-3 summarizes the results from the tests on the coefficients (estimated fixed effects) from the harmonization of each non-reference condition.

The estimated intercept was -9.4e-17 (CI: [-0.01, 1], *p*-value = 0.5). The between-subject variability was low as the random effect had an estimated standard deviation of 0.03, and the confidence intervals for the random effects lay around zero (Figure C-4 in Appendix C). So there was a low uncertainty in the results of the mixed-effect model. In the Quantile-quantile plot (Figure C-5 Appendix C), the residuals had only a few deviations from the theoretical normal quantiles. The assumption of equal variances was also verified with Levene's test as there was no significant difference in the spread of residuals among features (*p*-value >0.05). Figure 7-2Figure 7-4 show the range of CCC_1 and CCC_2 values for the 15 uncorrelated features used in the mixed effect model.

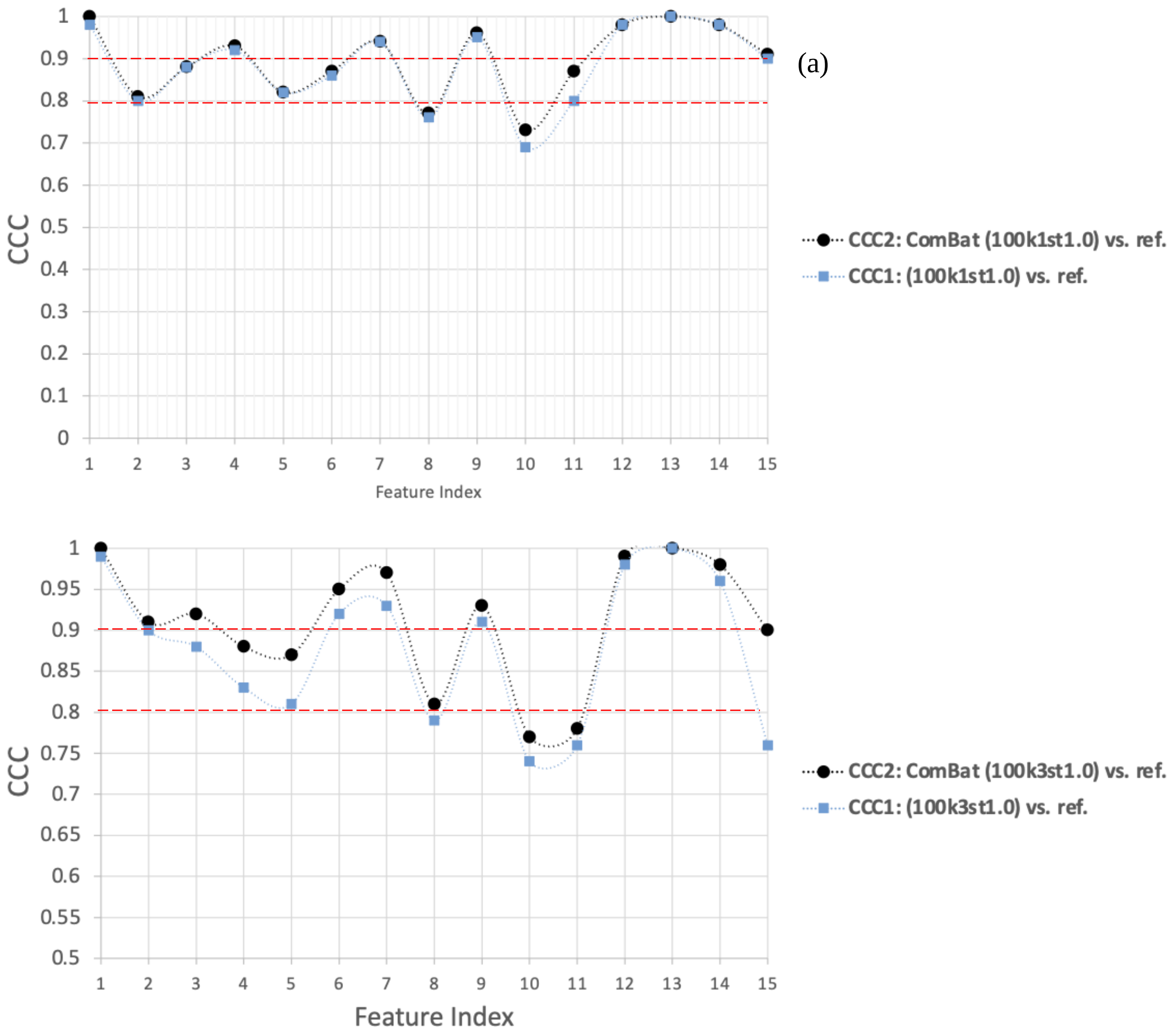


Figure 7-2. ComBat harmonization of kernel variations for (a) smooth kernel and (b) sharp kernel in the 15 uncorrelated features. CCC_1 and CCC_2 values for conditions with different kernels than reference are shown. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

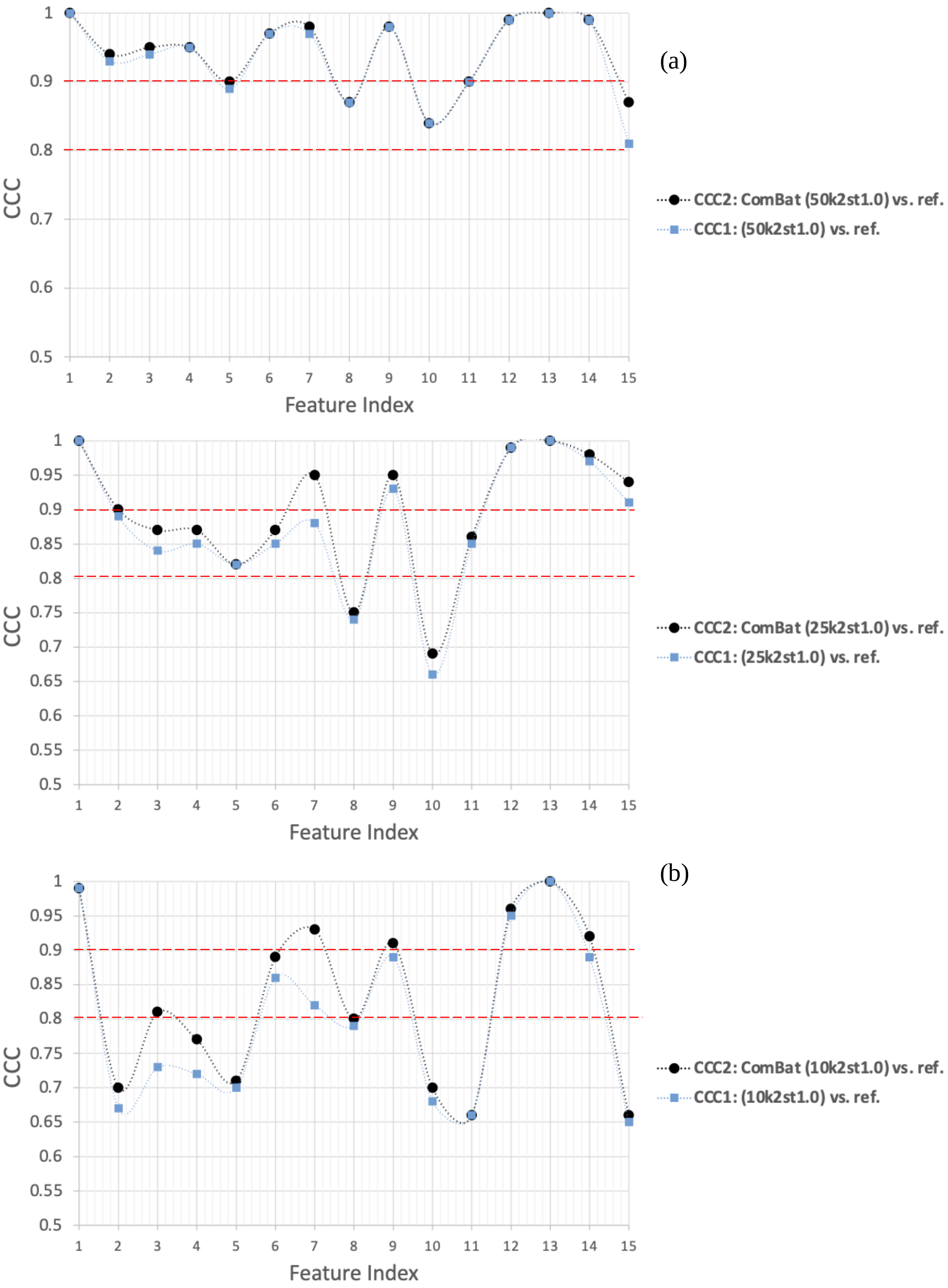


Figure 7-3. ComBat harmonization of 15 uncorrelated features (a) 50% dose, (b) 25% dose, and (c) 10% dose. CCC_1 for each radiomic feature and the corresponding CCC_2 values for non-reference conditions with different dose levels than reference are shown. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

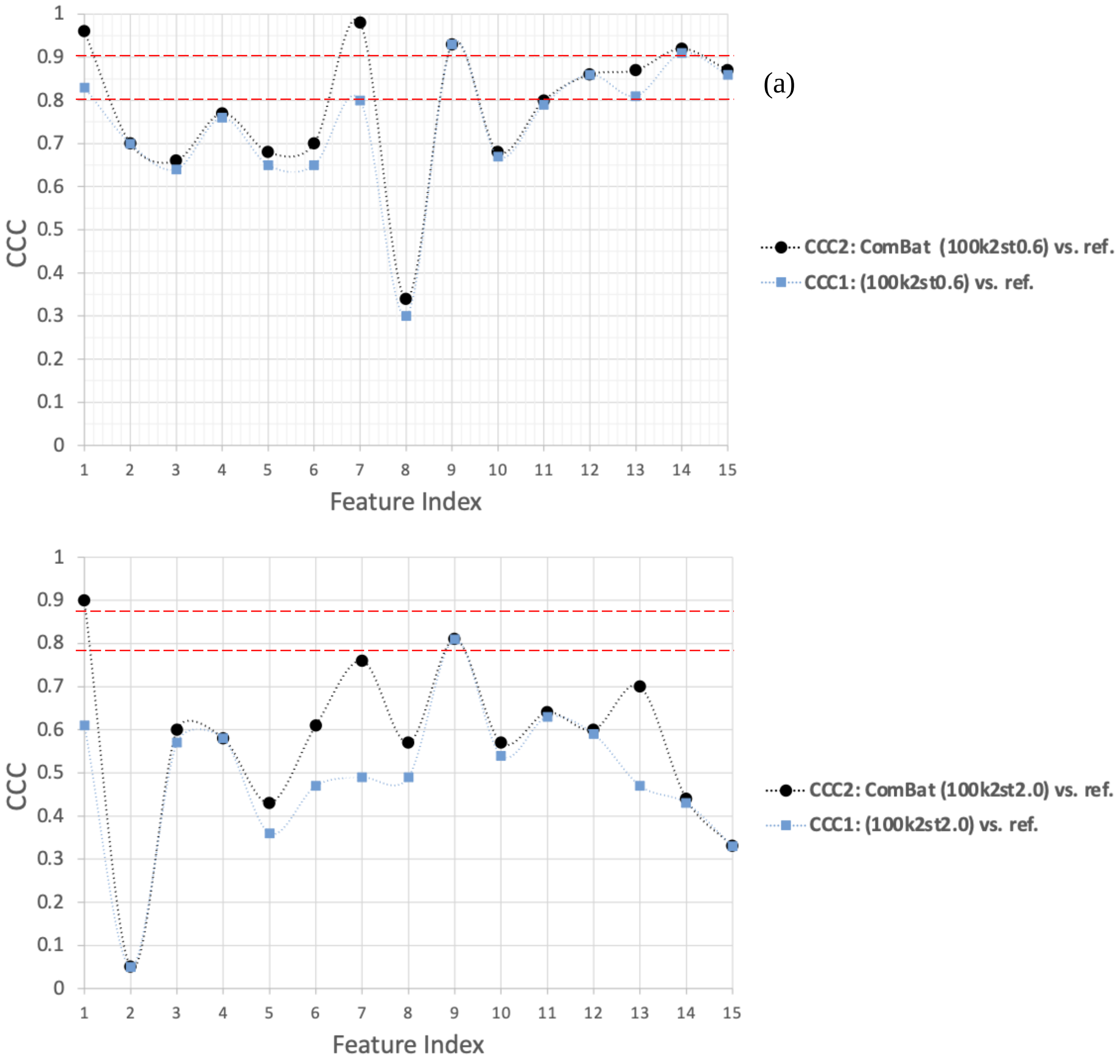


Figure 7-4. ComBat harmonization results of 15 uncorrelated features (a) 0.6mm thickness and (b) 2mm thickness. CCC_1 and CCC_2 values for conditions with different slice thickness than reference are shown. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

Table 7-3. Estimated fixed effects of harmonization of CT parameters on improvement of CCC

Harmonized non-reference Condition	Estimated Fixed Effect	95% One-sided CI	p-value
100%, **k1**, st1	0.01 ± 0.01	[-0.012, 1]	0.2
100%, **k3**, st1	0.03 ± 0.01	[0.008, 1]	0.01^*
50%, k2, st1	0.006 ± 0.01	[-0.018, 1]	0.3
25%, k2, st1	0.02 ± 0.01	[-0.007, 1]	0.1
10%, k2, st1	0.03 ± 0.01	[0.002, 1]	0.03^*
100%, k2, **st0.6**	0.04 ± 0.01	[0.012, 1]	$7.8e\text{-}3^*$
100%, k2, **st2**	0.08 ± 0.01	[0.05, 1]	$6.4e\text{-}7^*$

*Significant p-values with 95% confidence

According to *Table 7-3* for the set of 15 uncorrelated features, the harmonization of the conditions with different slice thicknesses than reference, the condition with the sharp kernel (k3), and the condition with a dose level of 10% are significant. This is also visualized in Figure 7-3 and Figure *7-4*, where CCC values of several features have increased after harmonization.

Harmonizing the condition with the smooth kernel was not significant, and it can be visualized in the comparison between Figure 7-2 (a) and (b).

As was expected, harmonization of 50% dose is not significant. For this condition, the reproducibility of radiomic features was already strong before harmonization (Figure 7-3 (a)). In the harmonization of dose variations, the application of ComBat was only significant for 10% dose level (Figure 7-3 (c)).

Figure 7-5 plots the range of $CCC_2 - CCC_1$ for the 15 uncorrelated features via boxplots. Data for conditions that did not receive significant harmonization are distributed close to zero. But the boxplots for those conditions with a significant impact of harmonization are distributed around more positive values.

The test results on the 15 uncorrelated features were consistent with observations made from CCC values for all 93 radiomic features (summarized in Table 7-2 and shown in Figure C-1, Figure C-2, and Figure C-3 in Appendix C). Those conditions that were significantly influenced by the ComBat application (in Table 7-3) had more radiomic features with $CCC_2 > CCC_1$, higher C_2-C_1, or better improvement in average CCC across all 93 radiomic features.

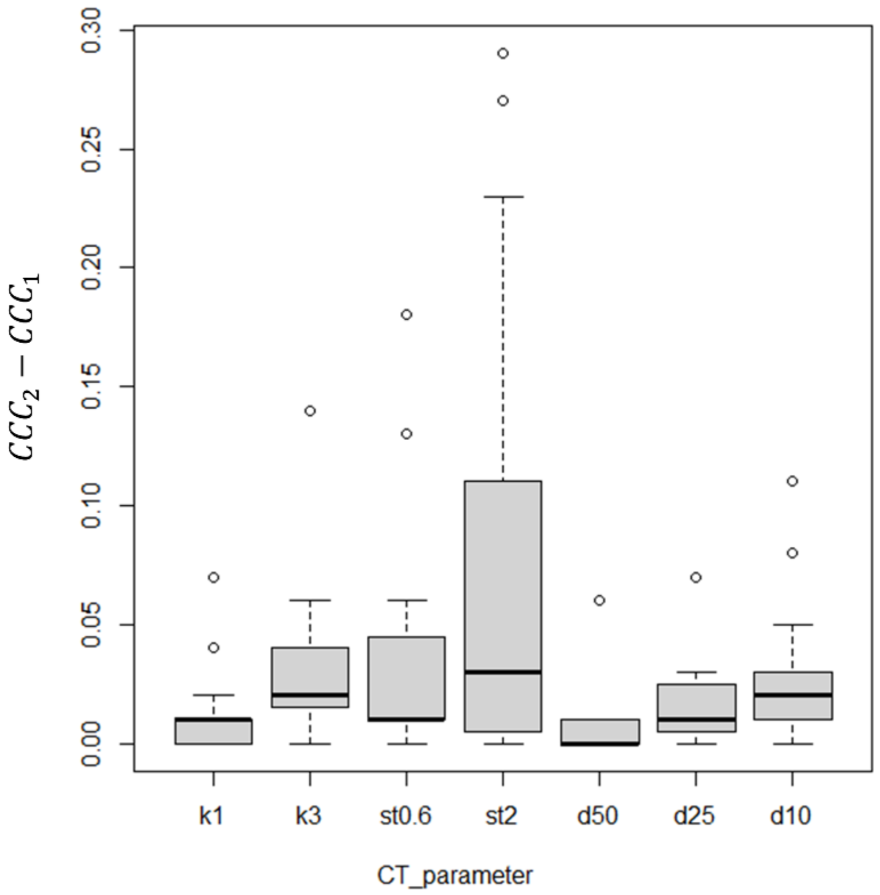


Figure 7-5. Boxplots showing the range of differences in CCC values of the 15 radiomic features before and after ComBat application. Each condition is identified by the level of the varying CT parameter. e.g., d50: (50%, k2, st1), d25: (25%, k2, st1), d10: (10%, k2, st1), etc. The horizontal lines show the median

7.4 Discussion

We investigated the ComBat technique to mitigate the impact of variation of slice thickness, kernel, and dose variation within a specific range on lung nodule radiomic features. Our results indicated that this technique could compensate for changes in the previously mentioned CT parameters in the explored ranges. After the application of ComBat, the radiomic features calculated from images at non-reference conditions were more closely aligned with the radiomic features of images at reference condition. We observed that for almost all non-reference conditions, there were more reproducible radiomic features after the ComBat application, except for the condition at 50% dose level, which already had several reproducible radiomic features. While dose reduction up to 50% has not impacted radiomic feature reproducibility, dose reduction up to 25% and 10% has affected the count of reproducible features. ComBat method increased the number of reproducible features for these conditions, and its impact was significant (p <0.05) in the harmonization of the condition with a 10% dose level.

Better average CCC improvement was observed across all 93 radiomic features in conditions that initially had very few reproducible radiomic features (e.g., (100%, k2, st2), (100%, k2, st0.6)). In addition, the mixed-effect model built by the 15 uncorrelated features indicated the significant (p <0.05) impact of harmonization on these conditions.

These findings can suggest that when the effect of a CT parameter variation was more evident, ComBat helped improve the agreement better than when a CT parameter had less impact on the reproducibility of radiomic features. So, the technique was more sensitive to more pronounced batch effects.

Although the average improvements for conditions like (100%, k2, st2) and (10%, k2, st0.6) were more than other conditions, fewer features were identified as reproducible. This can be partly due to the choice of a cut-off value of CCC = 0.9 as well.

In examining the variation of CCC values for all 93 radiomic features, according to Table 7-2 and Figure C-1, Figure C-2, and Figure C-3 in Appendix C, a group of radiomic features that had very poor reproducibility compared to the reference ($CCC_1 < 0.8$), did receive improvement after ComBat application, but the improvement did not result in the agreement going above the cut-off value of 0.9. For example, some features were consistently in poor agreement (CCC<0.8) with the reference condition even after harmonization at any of the seven non-reference conditions. For example, *GLSZM Normalized Gray Level Non-Uniformity*, *GLSZM Low Gray Level Zone Emphasis*, *GLSZM Small Area Low Gray Level Emphasis*, *GLCM MCC* always had low CCC values except that only a few had $0.8 < CCC < 0.9$ at a few conditions. Hence, in a radiomic study, these features that don't meet the cut-off values even after harmonization may not be recommended for being selected as a radiomic signature.

A few radiomic features changed from poor reproducibility to high reproducibility in a few conditions. *GLRLM Gray Level Non-uniformity* in (100%, k2, st0.6) and *GLRLM Run Variance* in (25%, k2, st1) condition were among those features that became highly reproducible. Another group of initially moderately reproducible features (0.8<CCC_2<0.9) received slight improvements by application of ComBat, and they became reproducible. Finally, for a group of already reproducible features compared to the reference (CCC_2≥0.9), their reproducibility remained the same or improved slightly.

7.4.1 Contributions

Although our findings were also in agreement with the results of prior studies[110,160], in this study, we sought to evaluate the ComBat technique further and verify findings from previous ComBat studies in the harmonization of radiomic features. The contribution of this study is that we expanded the explorations in the previous studies[110,160] by analyzing the reproducibility of radiomic features of lung nodules in a lung cancer screening cohort. Our analysis explored the harmonization impact of ComBat on the radiomic features that were only from the nodule region. Like Mahon *et al.* [110] our work explored harmonization of the effect of different CT protocols on radiomic features. However, in our study, either the investigated parameters were different, or the explored ranges were different. For example, we investigated the impact of the dose in terms of the variation of tube current instead of the variation of kVp. Also, we explored the ComBat technique's performance in mitigating a lower range of dose levels (2 to 0.2mGy) that degrades image quality more than the diagnostic dose levels and makes the task of mitigation more difficult. Our results showed that it is possible to correct the variability of some radiomic features due to dose reduction. Moreover, we explored a wider range of reconstruction kernels (from smooth to sharp).

Furthermore, unlike the studies mentioned above, we tested the utility of the ComBat technique in adjusting the non-reference radiomic features with respect to a reference (a gold standard). In contrast, those studies had aligned all radiomic features to a pooled grand mean. The latter approach involves modification of the reference radiomic features as well, which may not be desired. For example, suppose a model has been pre-trained on a reference CT condition to apply the model on non-reference radiomic features. In that case, we need the data adjusted to the training set instead of a pooled grand mean.

In understanding the robustness of radiomic feature values and the impact of a harmonization technique, it is important to use a sensitive metric that thoroughly assesses the robustness. Therefore, we investigated the reproducibility of radiomic features by CCC metric, which can detect lack of agreements between repeated measurements better than other metrics such as correlation coefficient, paired t-test, or least square analysis[59].

We assessed the impact of ComBat harmonization in two manners; First, by making comparisons between the agreement of all 93 radiomic features against a threshold of 0.9. Second, by taking into account the correlation of the radiomic features in making conclusions about the results. The former is valuable as applying a hard cut-off can be suitable for implementing a feature selection strategy based on radiomic feature robustness. Furthermore, in future radiomic studies that use image datasets with an inconsistent set of CT acquisition protocols, such thresholds can be used to filter features that are either inherently reproducible or can become reproducible by using ComBat. The latter aspect is also consequential since we found that many radiomic features are highly correlated (section 3.5 in Chapter 6). Therefore, we tested the significance of the impact of ComBat harmonization only on a set of uncorrelated features.

7.4.2 Practicality and Future Work

Implementation of the technique of ComBat in mitigating variability of radiomic features has multiple advantages. ComBat is a simple method, so it does not require computational power, and it is quick. On the other hand, one disadvantage of this technique is that calibrating radiomic feature variability requires a statistically representative sample data to estimate batch effects via the EB method. Furthermore, this technique's harmonization may not generalize well to radiomic features from unseen CT conditions, and verifying this fact remains as future work. In our study and other ComBat studies, harmonization has been performed in a data-driven manner in which batch effects are estimated and adjusted for the same data without the need for training of the ComBat algorithm on a separate training set. Nevertheless, it can be further explored to evaluate Combat to understand if batch effect estimates learned from previous data (i.e., a pre-trained ComBat) can reasonably harmonize a new dataset. This can avoid future runs of ComBat, as when new data arrives, it is not required to re-establish the harmonization and the estimation.

In exploring mitigation techniques that eliminate variability due to potential inconsistencies in image data, two other aspects are crucial: robustness and preservation of biological variability.

While the effectiveness of a mitigation technique is vital, we have to ensure that the meaningful biological information, hypothetically provided by the radiomic features, is not removed. With the ComBat method, it is possible to avoid accidental removal of biological variability by providing the information in the form of biological covariates to the model. By adding these biological covariates (β_f in Eq. 12), the algorithm estimates the variability due to the biological covariates, and then during the adjustment and removal of batch effects, it adds these estimates ($\widehat{\beta_f}$) to keep them protected (Eq. 13). Fortin *et al.* [157] has also verified this in the harmonization of multi-site diffusion tensor imaging data. Moreover, this was also confirmed in phantom studies[160] where the capability of radiomic features in distinguishing between two different texture patterns of the Credence Cartridge Radiomics (CCR) phantom[24] was not impaired after harmonization. Also, among other patient studies[28], it was shown that the technique maintained biological variability.

Our study did not check for this because we had the same subjects for images at different conditions (batches). Having images of the same subjects for all conditions allowed us to fully assess the impact and effectiveness of ComBat through pairwise comparisons between the harmonized features and a reference ground truth. Nonetheless, it remains as future work to further determine the preservation of biological variability and effectiveness of ComBat in a study designed so that each batch (CT condition) has a different and independent set of subjects. One such study design would be a multicenter study in which each center has patient datasets different from other centers.

The minimum sample size required for the ComBat technique to achieve satisfactory performance is another essential concept. Johnson *et al.*[101] originally proposed the ComBat technique to harmonize the batch effects in genomic studies as it was effective for small-size datasets. However, the minimum sample size requirement can be different for radiomic features. Orlhac *et al.* [160] investigated the robustness of the ComBat harmonization across different sample sizes and found that by gradually decreasing the sample size, ComBat performed satisfactorily with at least 20 patients per CT condition. In Appendix C, we have provided results from a similar experiment where we gradually decreased the sample size and performed bootstrapping to obtain a confidence interval for the measured CCC_2 corresponding to each sample size. We saw that the confidence interval for CCC_2 was wider for the sample size of 20, indicating more uncertainty. However, sample size reduction of up to around 50 to 60 samples maintained CCC_2 values and resulted in a narrower confidence interval.

The batch effect correction via ComBat may involve several runs of the algorithm with different design and setups; for example, in correcting the effect of the CT parameters, we also tested whether we would get different results if we had a data set that is consisted of batch effect due to variation of only one CT parameter instead of three. In our case, we did not see disparities between the algorithm's performance compared to the results reported above.

An important consideration in using ComBat is that data normalization before the application of ComBat is necessary. Additionally, if batch sizes are not equal or batch-group design is not balanced, meaning that the number of sample classes or sub-types is different in each batch, the approach will not work well. However, our study did not have this issue since we have measurements from the same set of subjects in each batch[154].

7.4.3 Limitation

Our study assessed the reproducibility of radiomic features within a cohort from our institute. Therefore, our findings require validations in a larger multicenter cohort that includes data from different CT scanners.

A limitation of this study is that the range of CT conditions explored in the current study was not wide enough to represent a comprehensive range of possible CT scans acquired in the clinic. However, our findings can have implications for other ranges of the values of the three CT parameters of dose, kernel, and slice thickness. Furthermore, since we applied a hard cut-off value to discriminate between reproducible and nonreproducible features, the choice of the threshold itself can impact the number of features identified as reproducible. In the context of radiomic features, it is unclear which CCC cut-off value can affect the radiomic feature predictivity or utility in a prediction model. Hence, we relied on a conservative threshold of CCC = 0.9.

It may not always be possible to know the exact batch effects that can cause inconsistency in the radiomic feature values in practice. For this study, we experimented with the scenario that we are aware of the batch effects, and we controlled for other CT parameters. If the sources for batch effects are unknown, it might be helpful to use other methods such as Surrogate Variable Analysis[166] to adjust for unknown sources of variability.

Finally, another limitation of this study is that, due to lack of availability of patient diagnosis, patient labels, or outcomes, we did not assess how the harmonization impacts the utility or predictivity of the radiomic features in a classification task

7.5 Conclusion

We investigated the mitigation of variability of CT radiomic features of lung nodules in a lung cancer screening patient population by application of the ComBat technique. Our results demonstrated that while radiomic features were impacted by variations of CT parameters of dose, kernel, and slice thickness, the ComBat technique could correct some of these variations under certain CT conditions.

Since the Combat technique is a simple technique that can effectively remove inconsistencies between radiomic features acquired with different protocols, it can facilitate future radiomic studies that use image datasets with a heterogeneous set of CT protocols.

Chapter 8 Comparisons Between the Techniques of GAN and Combat In Mitigating Variability Of Radiomic Features

8.1 Introduction

As has been discussed in the previous chapters, mitigating variability of radiomic features can primarily be performed in two main ways: in the image domain (before radiomic feature calculation) or in the feature domain (after feature calculation)[165].

In Chapter 6, we trained and evaluated the performance of a GAN model in the mitigation of variability of lung nodule radiomic features. The GAN model transforms images of lung nodules acquired with different CT conditions (non-reference conditions). Therefore, this technique functions in the image domain. The image transformation is aimed to align non-reference radiomic features with radiomic features of images of the same subject at a reference condition.

In Chapter 7, we evaluated a different harmonization technique, ComBat, that functions in the feature domain. This technique is directly applied to the measured radiomic features.

Both of these techniques have their strengths and weaknesses. As has been previously discussed, ComBat is a simple method that does not require the original image data for harmonization. Combat uses the Empirical Bayes (EB) method to estimate radiomic feature variability caused by the variation of CT parameters. Therefore, ComBat can be robust to sample size. On the other hand, GAN is computationally expensive, requires a large image dataset for training and evaluation, and is time-consuming. Though once a model with satisfactory performance is built, the task of image transformation itself takes a few seconds only. Other than the effectiveness of these techniques, the aforementioned practical aspects can also have a role in determining and selecting a valuable technique for the mitigation of radiomic feature variability in practice.

The unique differences between these two mitigation techniques make it compelling to compare their performance and effectiveness. Currently, there are no existing studies that directly compare the GAN technique with the ComBat technique. Therefore, in this chapter, we aimed to compare the performance of these two techniques in the mitigation of variability of lung nodule radiomic features in CT images acquired with different dose levels, wFBP reconstruction kernels, and slice thicknesses.

The inter-algorithm comparisons in this study can be helpful for future multicenter radiomic studies that need to identify an appropriate harmonization method. For example, results from our study can provide insight on either of the techniques that has a higher potential in eliminating inconsistencies due to variations of any of the CT parameters of slice thickness, dose, and kernel.

In Chapter 6, the GAN model was trained on cohort 4 and was evaluated on cohort 6 as explained in Chapter 3. However, in Chapter 7, we tested the effectiveness of the ComBat technique on a different dataset (cohort 2) than the data used in Chapter 6. Therefore, in this chapter, in order to more directly compare the performances of the two models on the same set of data, we applied the ComBat technique on the cohort 6 with 55 samples. We then compared the results obtained in Chapter 6 to the results from ComBat harmonization. Figure 8-1 shows the comparisons made in the current chapter.

The results from experiments on GAN shown in Chapter 6 will be shown once more in this chapter to compare with findings from ComBat application on cohort 6.

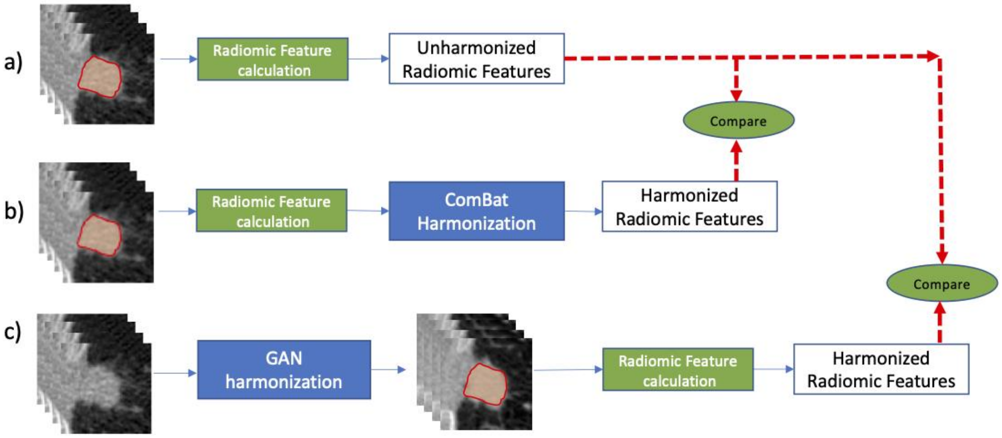

a)
Radiomic Feature calculation
Unharmonized Radiomic Features
Compare
b)
Radiomic Feature calculation
ComBat Harmonization
Harmonized Radiomic Features
Compare
c)
GAN harmonization
Radiomic Feature calculation
Harmonized Radiomic Features

Figure 8-1. Comparison of two mitigation techniques. a) Radiomic feature calculation without any harmonization step. b) Harmonization through applying ComBat on feature data. c) Harmonization through applying GAN on image data. Dashed red line show the comparisons made between unharmonized and harmonized data.

8.2 Methods

8.2.1 Data

The cohort 6 with 55 samples was used to test and compare ComBat harmonization performance versus GAN harmonization. From each case, one nodule was selected for analysis in the study. As shown in Table 8-1, images from the same set of conditions described in Chapters 6 and 7 were used. The condition (100% dose, medium kernel, 1mm thickness) was considered the reference CT protocol and was compared to other non-reference conditions to assess the reproducibility of radiomic features.

Similar to our experiments in previous chapters, for each case, a single nodule was identified and contoured to form a region of interest (ROI) for radiomic feature calculation. Three segmentation ROIs were used for the three different slice thickness reconstructions. So, for each subject, the images with 1mm slice thickness had the same segmented ROI, but the images with 0.6mm and 2mm thickness had different ROIs.

The same set of 93 (non-shape and non-size) radiomic features described in Chapters 6 and 7 were calculated with the same settings (see name of all features in Table B-2 of Appendix B).

Table 8-1. Description of CT conditions included in the study. (Values in bold show the reference condition)

Screening Dose	wFBP Reconstruction Kernel	Slice Thickness	Condition Identifier	Varying CT parameter
100%	**Medium**	**1mm**	**100%, k2, st1**	**-**
100%	Smooth	1mm	100%, k1, st1	Kernel
100%	Sharp	1mm	100%, k3, st1	
50%	Medium	1mm	50%, k2, st1	Dose
25%	Medium	1mm	25%, k2, st1	
10%	Medium	1mm	10%, k2, st1	
100%	Medium	0.6mm	100%, k2, st0.6	Slice Thickness
100%	Medium	2mm	100%, k2, st2	

8.2.2 ComBat Harmonization

The ComBat harmonization technique (described in Chapter 7) was applied to the radiomic features extracted from each case under each condition to estimate and adjust the batch effects caused by the variation of CT parameters of kernel, slice thickness, and dose to the reference conditions. The ComBat harmonization toolbox developed by Fortin *et al.* [157] was used in RStudio software[164]. The parametric approach, which assumes prior distributions of Gaussian for location shift parameter (γ) and inverse Gamma for the scale parameter (δ^2) was applied, and Empirical Bayes (EB)[153] was used for estimation. To check whether the assumptions (i.e., the parametric prior distributions) reasonably fit the data, the empirical distributions and the priors for the batch effect parameters (γ, δ^2) were plotted and were compared. Since our study design has images from the same subjects at different CT conditions (batches), there was no biological variability between data at different batches; therefore, we did not include biological covariates.

8.2.3 GAN Harmonization

In this Chapter we will use the results obtained from application of the trained GAN model in Chapter 6. Specifically, the results related to Concordance Correlation Coefficient (CCC) assessments from Table 3 of Chapter 6 will be shown here again to compare with the CCC results from the experiment with ComBat. Additionally, we will also use the results from the mixed effect model fitted to the difference of CCC before and after GAN harmonization (Table 4 in Chapter 6) to compare with results from ComBat.

8.2.4 Harmonization Evaluation

We made paired comparisons between radiomic features calculated at non-reference and reference conditions before and after each of the harmonization techniques. Similar to Chapter 6 and 7, the CCC was calculated to measure the agreement between the unharmonized data (CCC_U), after the GAN application (CCC_{GAN}), and after the ComBat harmonization (CCC_{ComBat}).

We initially examined the set of 15 uncorrelated radiomic features (described in Table B-1 of Appendix B). From these features, we compared the number of radiomic features that meet the cut-off value of 0.9 (i.e., $CCC_{GAN} \geq 0.9$ and $CCC_{ComBat} \geq 0.9$). We also compared the average differences between CCC values of unharmonized radiomic features (CCC_U) and harmonized radiomic features ($CCC_{GAN}\ or\ \ CCC_{ComBat}$) within each non-reference condition:

$$Average\ CCC\ difference\ for\ a\ non-reference\ condition = \frac{1}{M}\sum_{i=1}^{M=15}\left(\frac{CCC_2 - CCC_U}{CCC_U} \times 100\right) \qquad \text{Eq. 15}$$

$$where\ CCC_2\ is\ either\ CCC_{GAN}\ or\ \ CCC_{ComBat}$$

Similar to Chapter 6, our hypothesis (Eq. 8 in Chapter 6) was that the reproducibility of radiomic features improves after harmonization. We further compared results from two mixed-effect models that were separately fit to harmonized data acquired by applying the two techniques of GAN and ComBat. The mixed-effect model was fit to ($CCC_{GAN} - CCC_U$) by using the 15 radiomic features has been already demonstrated in Chapter 6. Results from the mixed-effect model fit to

($CCC_{ComBat} - CCC_U$) by using the 15 radiomic features in the dataset with 55 samples was performed in this study. Details of the mixed effect models are described in Eq. 7 and Eq. 8 in Chapter 6.

A summary of CCC values for all 93 radiomic features are also presented in Appendix D (Table D-1). In comparing the two harmonization techniques in this chapter, we focused on the set of 15 uncorrelated features to alleviate the bias that may be introduced when several features vary in correlation with each other.

8.3 Results

8.3.1 Comparison of CCC values before and after each harmonization technique

Table 8-2 summarizes the agreement analysis before harmonization (CCC_U) and after application of GAN (CCC_{GAN}) and ComBat (CCC_{ComBat}) for the selected 15 radiomic features. Also, Figure 8-2, Figure 8-3, and Figure *8-4*, visualize these values for each group of CT parameters of dose, kernel, and slice thickness. In these plots, CCC_1 refers to CCC_U, $CCC_2{:}\,GAN$ refers to CCC_{GAN}, and $CCC_2{:}\,Combat$ referes to CCC_{ComBat}.

Table 8-2. Summary of results from harmonization of 15 uncorrelated radiomic features using GAN or ComBat. The percentage difference between CCC before and after harmonization for 15 uncorrelated radiomic features, the number of features for the specified ranges of CCC values, and the increased number of features with CCC ≥ 0.9

		100%, st1		**K2, st1**			**100%, k2**	
		k1	**k3**	**50%**	**25%**	**10%**	**st0.6**	**st2**
Unharmonized	$\boldsymbol{CCC_U \geq 0.9}$	8	8	10	6	4	1	0
GAN **Harmonization**	$\boldsymbol{CCC_{GAN} > CCC_U}$	14	12	5	6	12	6	11
	$\boldsymbol{CCC_{GAN} \geq 0.9}$	13	12	11	6	7	1	1
	Average $\boldsymbol{CCC}$ **difference**	12.8%	9.8%	-0.2%	2%	14.4%	5.52%	75.4%
ComBat **Harmonization**	$\boldsymbol{CCC_{ComBat} > CCC_U}$	11	12	9	12	13	13	11
	$\boldsymbol{CCC_{ComBat} \geq 0.9}$	8	11	10	6	6	3	2
	Average $\boldsymbol{CCC}$ **difference**	2.3%	3.7%	1.9%	3.9%	3.9%	18.6%	12.9%

In harmonizing the effect of variation of dose on the 15 radiomic features, the GAN application in harmonizing 10% dose level resulted in higher average CCC difference and higher CCC values in Figure 8-2 (c). However, for 25% or 50%, the difference between GAN or ComBat results is minimal, though ComBat was slightly more effective. While the GAN application on these two conditions has negatively impacted a few features (where $CCC_{GAN} < CCC_U$), after Combat application, all the 15 radiomic features had $CCC_{Combat} > CCC_U$ (Figure 8-2 (a) and (b)).

In harmonizing smooth and sharp kernels, the GAN application has resulted in a more noticeable average CCC difference and more radiomic features that meet the 0.9 cut-off value compared to the ComBat application. This result is also visible in Figure 8-3, where for both non-reference kernels (k1 and k3), the GAN application (black and circular data points) have higher CCC values than the ComBat application (orange triangle data points).

In harmonizing the effect of slice thickness variation on the 15 radiomic features, the GAN considerably improved the average CCC for the condition with 2mm thickness. But only one radiomic features met the cut-off value of 0.9 at either of the two different slice thickness. On the other hand, in harmonization of slice thickness, the ComBat application slightly increased the average CCC difference, and two to three radiomic features reached strong agreement with the reference ($CCC_{Combat} \geq 0.9$). *Figure 8-4* compares the CCC values for the GAN and ComBat harmonization with unharmonized CCC values. For some features the GAN harmonization is better and for some other features, the ComBat has resulted in higher agreements with reference. For example, at the condition with 2mm slice thickness, the feature *GLRLM Normalized Gray Level Non-Uniformity* has larger CCC improvement by GAN.

Thus, while ComBat and GAN harmonization were impactful, only a few radiomic features met the cut-off value (i.e., $CCC_{ComBat} \geq 0.9$ or $CCC_{GAN} \geq 0.9$). This finding was also observed in Chapters 6 and 7, where the average CCC increased but was not sufficient to increase the number of features that meet the cut-off value.

To understand the improvement of CCC values for all the 93 radiomic features, we provided the results in the Table D-1 of Appendix D. As a summary, we observed that CCC assessments for all 93 radiomic features, in the Table D-1of Appendix D, are consistent with the results summarized in Table 8-2 of the current chapter. The GAN application had a more positive influence in remedying radiomic feature variability due to kernel changes across all radiomic features (i.e., more features with $CCC_{GAN} \geq 0.9$ and greater average CCC difference). We also saw similar results for harmonizing dose variations where GAN had a more positive influence on the 10% dose level, and ComBat was slightly better for 50% and 25%. Similarly and consistent with Table 8-2, in the harmonization of slice thickness variation across all 93 radiomic features, GAN application strengthened the agreements of non-reference condition with 2mm with reference more than ComBat.

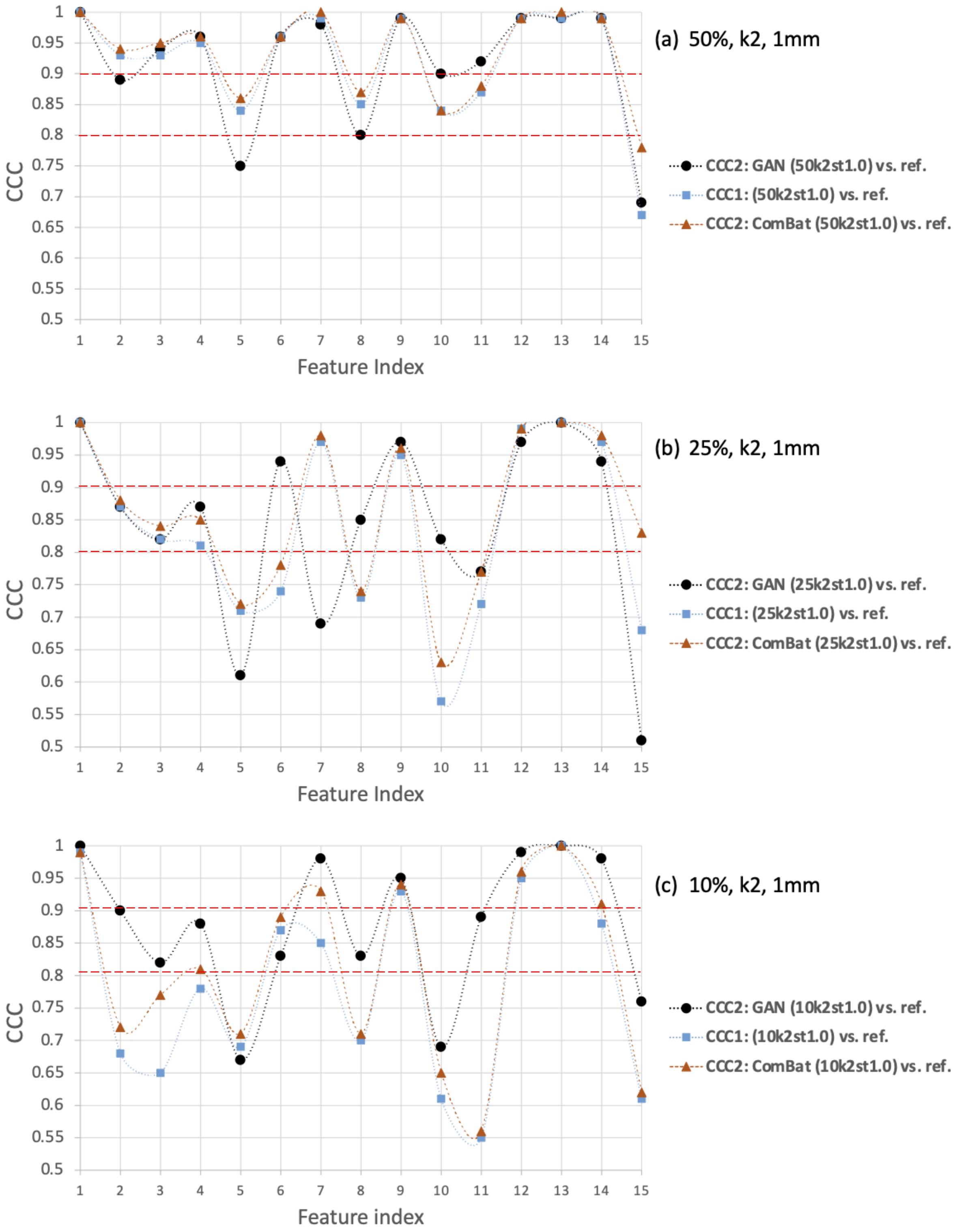


Figure 8-2. $CCC_1(or\ CCC_U\ )$ for 15 uncorrelated features and the corresponding GAN or ComBat harmonized CCC values for non-reference conditions with different dose levels than reference. Harmonization of (a) 50% dose, (b) 25% dose, (c) 10% dose. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

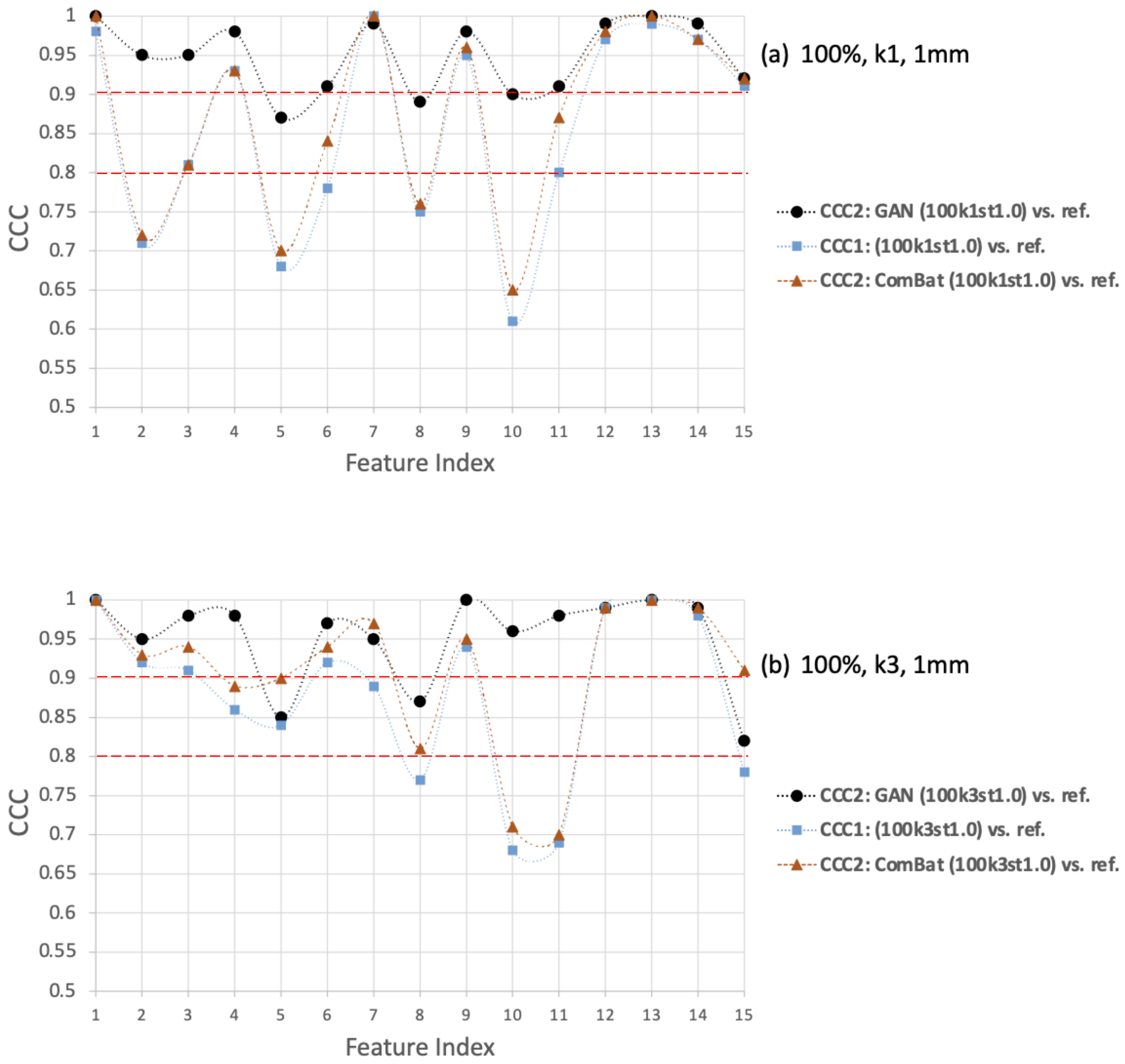


Figure 8-3. $CCC_1 (or\ CCC_U\ )$ and CCC_2 after ComBat or GAN harmonization of conditions with different kernel than reference. Harmonization results of (a) smooth kernel, (b) sharp kernel for 15 uncorrelated features. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

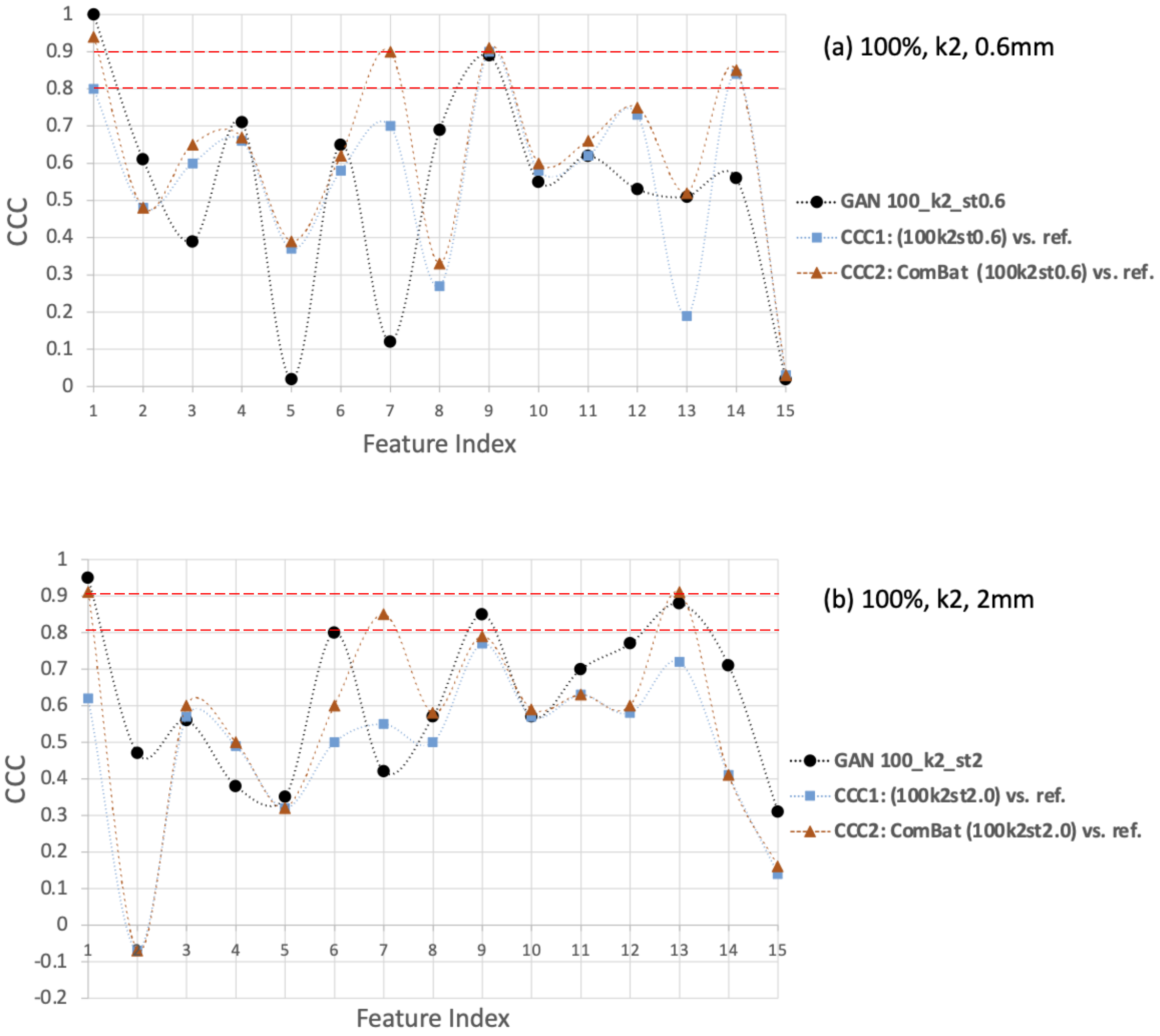


Figure 8-4. $CCC_1(or\ CCC_U\ )$ and CCC_2 (after ComBat or GAN harmonization of conditions with different slice thickness than reference. Harmonization results of (a) 0.6mm thickness, (b) 2mm thickness for 15 uncorrelated features via GAN or ComBat. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

8.3.2 Results from fixed-effect models

Table 8-3 shows results from the two mixed-effect models that tested the significance of the harmonization impact of each of the GAN and ComBat techniques on the 15 radiomic features for each non-reference condition. Specifically, in Chapter 6, for GAN harmonization, we fit the model to $CCC_{GAN} - CCC_U$. For ComBat, in this chapter, we fit the mixed-effect model to $CCC_{ComBat} - CCC_U$.

Table 8-3. Estimated fixed effects for the impact of harmonization of CT parameters on the improvement of CCC for the 15 radiomic features

Non-reference Conditions	**GAN Harmonization**			**ComBat Harmonization**		
	Estimated Fixed Effect	**95% One-sided CI**	**p-value**	**Estimated Fixed Effect**	**95% One-sided CI**	**p-value**
100%, **k1**, st1	0.09 ± 0.04	[0.02 , 1]	0.02*	0.02 ± 0.02	[-0.01, 1]	0.2
100%, **k3**, st1	0.07 ± 0.04	[0.0008 , 1]	0.04*	0.03± 0.02	[-0.0004, 1]	0.05
50%, k2, st1	-0.002 ± 0.04	[-0.07 , 1]	0.52	0.01 ± 0.02	[-0.01, 1]	0.2
25%, k2, st1	0.006 ± 0.04	[-0.06 , 1]	0.43	0.03± 0.02	[-0.003, 1]	0.07
10%, k2, st1	0.09 ± 0.04	[0.02, 1]	0.01*	0.03± 0.02	[-0.002, 1]	0.06
100%, k2, **st0.6**	-0.03 ± 0.04	[-0.1, 1]	0.7	0.06± 0.02	[0.03, 1]	0.0005*
100%, k2, **st2**	0.13 ± 0.04	[0.05 , 1]	0.002*	0.07± 0.02	[0.04, 1]	0.0001*

*Significant p-values with 95% confidence

Table 8-3 demonstrates that harmonization of non-reference conditions with smooth (k1) or sharp kernel (k3) was only significant when the GAN technique was used. Similarly, in harmonizing the condition with a 10% dose level, only the GAN technique was significant.

On the other hand, only the ComBat application resulted in a significant increase of CCC values in conditions with 0.6mm slice thickness than the reference. The condition with 2mm slice thickness received significant improvements by both of the harmonization techniques.

The estimated intercept in the fixed-effect model fit to ComBat harmonized data was -3.9e-16 (CI: [-0.02, 1], p-value = 0.5). The between-subject variability was low as the random effect had an estimated standard deviation of 0.01, and the confidence intervals for the random effects lay around zero (Figure D-1 in Appendix D). So, there was low uncertainty in the results.

The results from the fixed-effect models are consistent with the results shown in the previous section for the 15 radiomic features. Those conditions that received significant enhancement by applying any of the harmonization techniques have greater CCC values in their corresponding plot in Figure 8-2, Figure 8-3, and Figure 8-4.

8.4 Discussion

In this study, we aimed to make a comparison between the GAN and the ComBat techniques to understand their differences in mitigating variability of lung nodule radiomic features in response to variation of CT parameters of dose, kernel, and slice thickness.

8.4.1 Summary of Results

The main finding from this study was that the harmonization performance of the GAN and ComBat techniques could be complementary for some non-reference conditions; for some non-reference conditions, the application of only one or two of the two techniques resulted in substantial or significant harmonization impact. Simultaneously, for some conditions (e.g., conditions with 50% dose or 25% dose), applying either of the techniques resulted in slight to moderate improvements. Another finding of this study is that for some conditions (e.g., conditions with 0.6mm and 2mm), even though both of the methods increased the agreement (CCC) of non-reference radiomic features with the reference radiomic features, the CCC values did not meet the cut-off value of 0.9.Additionally, while the mixed-effect model indicated that the improvements of CCC values at the condition with 2mm slice thickness was significant, the average CCC improvements was much higher by the GAN harmonization (see Table 8-2).

Furthermore, contrary to Chapter 7, ComBat harmonization of conditions with k3 kernel and 10% dose was not significant. However, for these conditions, the impact of ComBat harmonization was substantial as we obtained a very small (but not significant) *p*-value in Table 8-3, and we observed the increase in the count of radiomic features with $CCC \geq 0.9$ in Table 8-2. One reason for obtaining different harmonization results for conditions with k3 kernel and 10% dose can be the variation of dataset size. In Appendix C (Figure C-6, Figure C-7, Figure C-8), we showed that with a dataset of size around 60 samples, ComBat's performance did not change noticeably.

Nevertheless, for datasets with sizes below 60, the performance dropped slightly. Considering that the dataset used in this study contains 55 samples, it might be expected to see a drop in performance when harmonizing conditions with 10% dose and sharp kernel.

8.4.2 Differences in GAN and ComBat Harmonization

GAN and ComBat techniques perform in two different fashions; GAN works with original images, and the GAN approach used in this study involved training and applying the seven condition-specific GAN models for harmonizing each of the seven non-reference conditions. Each condition-specific GAN model is trained using a pair of images, one at the reference condition and one at the corresponding non-reference condition. Variation of each CT parameter causes different amounts or types of image quality deviations from the reference. The generator's and the discriminator's weights are trained or penalized according to these deviations within each pair of images. Therefore, the GAN training aims to generate images without those deviations so that the generated images look similar to the image at reference condition.

On the other hand, ComBat directly compares the non-reference radiomic features with the reference radiomic features and estimates the location shift and scale parameter required to align non-reference data with the reference. For those non-reference conditions (such as the condition with 50% dose) with more radiomic features similar to the reference (i.e., more radiomic features with $CCC \geq 0.9$), the location shift and scale that is required for adjustment is minimal. Therefore, ComBat applies a slight adjustment for such radiomic features (e.g., see $CCC_U\ and\ CCC_{Combat}$ values for the condition with 50% dose).

Another difference in the performances of GAN and ComBat techniques is observed in the harmonization of the condition with 0.6mm slice thickness. The GAN model sees the differences in the shape of nodules between thin or thick slices compared to the 1mm images. As has been shown in Chapter 6, the model has tried to modify the image such that the nodule appears similar to the reference. However, it was not able to accomplish this perfectly. Hence, this phenomenon for non-reference conditions with 0.6mm slice thicknesses has resulted in a poor agreement between the generated images (or their radiomic features) with the reference condition. However, the ComBat technique, again, is directly exposed to the deviation of the radiomic features from the reference condition, and by estimating the location shift and scale for the deviation it was able to mitigate radiomic feature variations at different slice thicknesses.

8.4.3 Practicality and Future Directions

According to the results of this study, both GAN and ComBat techniques have their trade-offs. These two techniques can complement each other in terms of both effectiveness and practical aspects. When only one technique, from these two techniques, provides impactful harmonization, we have a limited choice. However, if both techniques positively influence the radiomic features, we will have the flexibility to choose the technique that best suits the practical aspects of a radiomic study of interest.

For example, in a multicenter study, where it is not possible to share image data due to privacy or storage concerns (e.g., radiomic features), we will be limited to a mitigation technique that does not require image processing. In such situations, a technique such as ComBat can be helpful.

However, when it is possible to share the image data, applying a technique, such as GAN, that can harmonize both the image and the quantitative radiomic feature values may be desired. Nevertheless, even if the image data is available, but it is acquired at a CT condition that the GAN model was not trained on, the GAN model will not be helpful. Whereas, if the data is a statistically representative dataset consisting of a range of different CT parameters, ComBat can be applied to adjust the radiomic feature variabilities without the need for a pre-trained model.

If it is feasible to use any of these two harmonization techniques in terms of effectiveness and practicality, ComBat may be preferred as it is fast, simple, and not computationally expensive. While the ComBat technique is simple compared to other basic methods, such as basic scaling, it adjusts the fraction of total variance due to variations caused by CT parameter inconsistencies.

Additionally, since the GAN model only transforms the non-reference images, additional radiomic feature calculations are required after image transformation. The radiomic feature recalculation also requires the availability of the nodule mask. If nodule masks are not already available from original non-reference images, nodule segmentations shall be performed on the transformed images to obtain the masks for feature calculation. On the other hand, re-segmentation may result in variation of segmented mask compared to the mask used for radiomic feature calculation before GAN transformation. Variation of the segmentation can itself result in inconsistencies in radiomic feature values. However, the variability of segmentations before and after the GAN application remains a future investigation.

An advantage of using the GAN model is the possibility of obtaining and keeping a standardized image that can be used in the future, for example, for calculating different groups of radiomic features or for other purposes such as augmentation data for future studies.

Both techniques have particular requirements to achieve a reasonably good performance. For example, they require a statistically representative dataset or a sufficiently large dataset. In Chapter 7, we found that the performance of ComBat was maintained with at least 60 samples. However, it remains as future work to estimate minimum required training samples for the GAN model.

Moreover, in situations where the batch effect (the source of inconsistency or variability) is unknown, we cannot use ComBat. However, it may be possible to train a general-purpose GAN that transforms images to look similar to a reference condition without knowing the specific CT parameter that deviates from the reference condition.

ComBat performance can be unreliable if classes (categories) of samples are not balanced across batches (non-reference conditions)[154]. If the proportion of classes in each batch is not balanced, the batch-effect estimates will depend on the proportion of classes within each batch. Hence in an unbalanced batch-group design, it is not recommended to use ComBat.

In Chapter 4 of this dissertation, we showed that slice thickness was an important factor affecting the reproducibility of radiomic features. In this chapter, we found that even though a technique such as ComBat can mitigate the variability of radiomic features, it may not be practically sufficient; we found that several radiomic features still had $CCC_2 < 0.9$. For future studies, it is worthwhile to explore whether a resampling (or interpolating) of images to the reference slice thickness before the ComBat harmonization helps improve the producibility of radiomic features. In Appendix B (Table B-4), we showed that resampling alone does not substantially improve the reproducibility of radiomic features. However, the combination of the resampling technique with another harmonization technique may result in better mitigation performance.

Among other approaches that can be combined with ComBat or with resampling to improve the reproducibility of radiomic features is the approach proposed by Gang *et al.*[105] that operates in both image and feature domain by performing two deconvolution steps to remove blur and noise. Since the volume averaging impacts the image blurriness, the recovery technique proposed by the authors can adjust the radiomic features for variations due to slice thickness.

8.4.4 Contribution

Some previous studies have explored the potential of various mitigation techniques, either in the image domain or feature domain. The application of ComBat and GAN has been previously explored independently in a few studies (we have discussed their distinction to the approach in this dissertation in Chapters 6 and 7). Nevertheless, few studies have directly compared the performance of two or more mitigation techniques from different domains. For instance, Foy *et al*.[81] compared the ComBat technique with histogram normalization and Butterworth filtering. However, the harmonization performance of ComBat has not been directly compared to the GAN technique. We performed this analogy in a consistent setting (with the same scan protocols) and the same dataset. So, we were able to inform and compare the limitations or the strengths of the harmonization performance of each of the two techniques.

Moreover, we discussed the practical scenarios of radiomic studies in which we can utilize either of the techniques. Additionally, our results identify the range of non-reference CT conditions that can receive effective harmonization via any two techniques.

The inter-algorithm analysis in this chapter can provide a basis for future research in radiomic feature stabilization.

8.4.5 Limitation

A limitation of this study is that we did not compare the effect of harmonizing each of these techniques on the predictivity or utility of radiomic features. Therefore, it is vital to understand how these harmonization techniques affect the downstream performance of radiomic features when used in a prediction model.

While we have already investigated the potential of the ComBat technique in a larger dataset, the comparisons in this chapter were limited to smaller size data. Therefore, it will be beneficial to expand further the size of the held-out test set used to evaluate the performance of the ComBat technique compared to the GAN technique.

8.4.6 Conclusion

In this chapter, we compared and contrasted two mitigation techniques of GAN and ComBat to stabilize CT radiomic features. These two techniques can have complimentary performance. In harmonizing radiomic features acquired from conditions with different kernel or dose levels than the reference, either of these techniques provided acceptable or significant harmonization. This result proves the possibility of correction of the effect of kernel and dose on CT radiomic feature values. Future explorations are required to achieve better radiomic feature harmonization in response to variation of slice thickness.

Chapter 9 Summary and Conclusion

9.1 Summary

Radiomic features are quantitative data calculated from medical images to characterize properties of regions of interest within the image, such as a suspicious lesion or tumor tissue. Recent research has shown the promise of radiomic features in describing tumor phenotype and biology. This suggests that radiomic feature values change accordingly in response to variations of tissue biology.

To ensure the reliability of radiomic features, we have to understand whether these quantitative measurements can be affected by the variation of factors unrelated to tissue biology. Therefore, this dissertation's goal was to determine whether CT radiomic feature values are impacted by the variation of factors related to CT image acquisition. Secondly, we aimed to assess the possibility of correcting the deviation of radiomic features due to changes in non-biological factors. The non-biological factors investigated in this dissertation were CT dose level, reconstruction slice thickness, and kernel.

In Specific Aim (SA) 1 of this dissertation, we first collected the required data to accomplish these goals. We then developed a modular framework for image data analysis as part of an in-house computational pipeline that enabled us to analyze image datasets and test the reproducibility of radiomic features. In SA 2, we used the computational pipeline to determine whether a set of well-known radiomic features are reproducible in response to variation of dose, kernel, and slice thickness (in both reconstruction algorithms of wFBP and iterative). Finally, in SA 3, we tested the potential of the two mitigation techniques of ComBat and Generative Adversarial Networks (GAN) in remedying the lack of reproducibility of radiomic features in response to the variation of CT parameters of dose, kernel, and slice thickness (in wFBP reconstruction algorithm).

The in-house computational pipeline described in Chapter 2 consisted of two main components: 1) simulation and reconstruction of CT projection data over a wide variation of CT conditions, 2) image data analysis. The modules developed in SA2 were part of the image data analysis and performed evaluation of lung nodule detection as well as radiomic feature calculation.

First, as was described in Chapter 3, under IRB approval, raw projection data of chest CT scans of lung cancer screening cases were collected. Then, the computational pipeline was used to create a dataset consisting of images at various CT conditions with nodule segmentations and radiomic feature calculations.

For SA 2, we initially investigated the robustness of radiomic features of patient lung nodules across 36 different CT conditions at four different dose levels, three different slice thicknesses, and three different kernels. Univariable and multivariable analyses of this study not only showed the lack of reproducibility of the majority of radiomic features but also revealed interactions among CT parameters, meaning that the effect of individual CT parameters on radiomic features can be conditional upon other CT acquisition and reconstruction parameters. This finding suggests that in designing scan protocols for future radiomic studies, specific combinations of CT parameters can result in groups of CT conditions that have reproducible radiomic features among themselves (e.g., OCCC≥0.9). Hence by carefully defining such conditions they can be considered as a group of 'safe' set of protocols.

Among the CT parameters investigated, slice thickness had the largest, and dose had the least noticeable impact on lung nodule radiomic feature values. Furthermore, in SA 2, to better understand the underlying process in the variation of radiomic features, we assessed radiomic features calculated from images of three different phantoms with known texture and intensity characteristics. CT scans from three phantom datasets of the FDA anthropomorphic phantom with a homogeneous synthetic nodule and two other phantoms without synthetic nodules: the Credence Cartridge Radiomics (CCR) phantom and a homogeneous water phantom were used. We found that the impact of slice thickness is only noticeable in a phantom with synthetic nodule, and we showed that this is due to volume averaging. This observation explained the large variability caused by slice thickness variation in patient lung nodules. On the other hand, we observed that kernel and dose variations had a more considerable impact on the radiomic features of homogenous material than non-homogenous materials. Therefore, we also demonstrated that the effect of kernel and dose variations is related to the inherent tissue texture.

For SA 3, we applied the two mitigation techniques of GAN and ComBat to harmonize the radiomic features of seven non-reference CT conditions and align the features with a supposed reference condition. We showed that the dependence of radiomic features on non-biological factors of CT dose, slice thickness, and kernel (in the wFBP algorithm) can be corrected to some extent after applying either of these techniques. Furthermore, our results demonstrated that either of these techniques has the potential to remedy the poor reproducibility of radiomic features in some CT conditions.

9.2 Clinical Implication and Contribution

The quantitative medical imaging field is an emerging field. Radiomic features are quantitative imaging data that can be used in clinical practice and enable precision medicine. Although several studies have shown the promise of radiomic features to serve as imaging biomarkers, there is still a need to carry out multicenter trials to test the clinical utility of radiomics in large and diverse datasets. Obtaining a large, multicenter dataset will increase the power of the study in evaluating the clinical relevance and robustness of radiomics. However, currently, there are no standardizations in scan protocols used for radiomic studies and there are no guidelines for gathering large multicenter datasets. This dissertation showed that if image data is not acquired with consistent CT protocols across different centers, there will be inconsistencies in radiomic feature values. These inconsistencies may limit the utility of radiomic features. Therefore, our study warns about this serious problem if the sources of variability of radiomic features are not being considered.

Furthermore, one essential step in validating radiomics as an image biomarker is the technical validation of radiomics to ensure its precision. Therefore, another key contribution of the work presented in this dissertation is the assessment of the reproducibility of radiomic features that provides insight into the precision of radiomics and the potential strategies that can handle the related uncertainties.

While similar efforts in the literature are available that have analyzed phantom data, we performed a comprehensive patient study that systematically explored the multidimensional space of CT parameters in affecting lung nodule radiomic features. Our study provided insight into the impact of three CT parameters on a set of well-known radiomic features. Thus, our findings identify the importance of careful radiomic feature selection and attention to the inclusion criteria for CT image acquisition protocols within multicenter radiomic studies.

Despite showing that the radiomic features may convey information related to the scan protocol, we also showed that via performing a harmonization step, we can mitigate the dependence of the radiomic features on CT parameters and alleviate the severe problem of lack of robustness.

While we did not use a multicenter dataset, we utilized an-in house computational tool that simulated image data at various CT conditions as if they were images acquired at different centers with different scan protocols. Therefore, our findings suggest that inclusion of a harmonization step before prediction modeling, has the potential to make the radiomic features of multicenter studies consistent and comparable. Furthermore, we illustrated that a deep learning-based GAN technique or a data-driven technique such as ComBat are potential mitigation techniques can serve as potential mitigation techniques for radiomics. Although further evaluation is needed to ensure performance across a wide variation of CT parameter ranges, our findings highlight a clear possible path for future multicenter studies to use either of these harmonization techniques and extend the evaluation of radiomic features in large and inclusive datasets, and hopefully enable the application of radiomics in clinical practice.

Although a number of recent works[129,131,167] have explored the potential of GAN models in the biomedical imaging field, only a few applied limited explorations of impact of GAN on radiomic features. However, we have evaluated the potential of a specific and a different variant of GAN model (Pix2Pix[132]) in the particular task of mitigating variabilities of large number of well-known CT radiomic features of lung nodules. Importantly, unlike other studies that have focused on general radiomic features, our focus was only on the radiomic features extracted from an ROI of nodule rather than a random patch in the chest scan or the whole 3D volume surrounding the nodule. Additionally, unlike those mentioned studies, we showed that it is possible to train an effective GAN model that serves as a harmonization tool, even when it is only trained on a sub-volume image around the nodule region instead of being trained on the image of the whole lung. This yields a less computationally expensive GAN training. Furthermore, the effect of dose variation on radiomic features and mitigation of dose impact is rarely studied in the literature, as it is infeasible to acquire repeated CT scans with different dose levels. However, we examined the dose impact and its mitigation by utilizing our in-house pipeline that simulated low-dose levels and avoided repeated scans.

We foresee that, for the future, in building predictive radiomic models, we should either perform a robustness-based feature selection or apply a harmonization technique that aligns the radiomic features. The methods presented in this dissertation lays the groundwork for research direction related to any of the two approaches above. In the implementation of the former approach, the contribution of this dissertation is that, in Chapter 4, we showed that through assessment of agreement of radiomic features via calculation of Overall Concordance Correlation Coefficient (OCCC), we can identify and filter radiomic features with poor reproducibility. To implement the latter approach, in Chapters 6 to 8, we tested and demonstrated the potential of two different harmonization techniques that either perform in the image domain or the feature domain. Notably, the methodology applied in testing the algorithms is universal and can be used for analyzing robustness or mitigation of other radiomic features or other CT parameters. For example, our approach in testing the efficiency of mitigation algorithms can serve as a framework for future inter-algorithm studies for organizations such as Quantitative Imaging Biomarkers Alliance (QIBA) that seek to enable the practicality of quantitative imaging biomarkers.

Furthermore, the developed pipeline tool that handles the evaluation of lung nodule segmentation and radiomic feature calculation can be transferred to other tasks. For example, via the development of few customized modules, the tool can be transferred to evaluate segmentation and detection of tumors in other organs or to calculate and assess other radiomic features across various CT parameters.

9.3 Limitation and Future Work

A limitation of the work presented in this dissertation is that our data was only composed of patients scanned only on Siemens scanners in the UCLA medical centers. Although we have incorporated a wide variation of CT conditions in our studies, a larger patient cohort is needed to capture a more comprehensive range of possible image reconstructions acquired with scanners from vendors other than Siemens and other regions. Availability and obtaining large patient cohorts can facilitate extensive validation of the harmonization methods and performance of inter-algorithm studies that minimize inconsistencies among radiomic features. In addition, exploring data from other clinical centers acquired with different scanners can reveal other potential sources of variability that we need to tackle. Finally, considering that in this dissertation, we have only assessed the impact of three CT parameters on the reliability of radiomic features, a multicenter dataset can allow for investigation of effects of other CT parameters (e.g., kVp, pixel size, etc.).

The dataset used in this dissertation was only composed of CT scans of the lung cancer screening population. The screening CT scan protocol is low-dose (~2 mGy CTDIVol) compared to routine diagnostic scans (~10-15 mGy CTDIVol). Therefore, beyond exploring differences among different scanners, acquiring cohorts from slightly different populations such as diagnostic CT scans will also allow for further validations of the findings of this study. We assessed radiomic feature reproducibility at very low dose levels (up to 10% of screening) to obtain an understanding of the lower bound of reproducibility for the radiomic features. Interestingly we found that although the images at 10% dose are not visually of high quality, there were situations where we could correct the variation of radiomic features at this low dose level. Hence, such a dose level may not be entirely out of the question in quantitative imaging when a harmonization step is included.

Furthermore, the majority of the work in this dissertation was based on the wFBP reconstruction algorithm only. For iterative reconstruction images, we only assessed the reproducibility of radiomic features. Thus, it remains a future work to determine the efficiency of ComBat and GAN harmonization for radiomic features of images with iterative reconstruction algorithm.

Furthermore, considering that more radiomic features had poor reproducibility in iterative reconstructions, it will be useful to understand which technique will be more effective in remedying the poor reproducibility.

We foresee that future radiomic studies shall include a standardization step before discovering clinical relevance (outcome prediction, etc.) of radiomic features. Nonetheless, the effect of variability of radiomic features or their mitigation on subsequent predictive performance remains an open area for future research. While it is essential to ensure the reliability of radiomic features, it is also vital to maintain their capability in distinguishing tumor biology. Therefore, this remains a potential future research direction.

Another potential future direction in the field of quantitative imaging biomarkers is the deployment of deep features. In this dissertation, we focused on hand-crafted radiomic features. However, deep features, extracted from final layers of deep learning models, can allow for the extraction of high-level information that is not dependent on the user definitions and relies on the intuition of an artificial intelligence model[168]. Hence, the study structure presented in this work can also serve as a framework for the assessment of robustness of deep features.

Interestingly, in SA 3, we found that the two techniques of GAN and ComBat have complementary performance. Therefore, the two techniques may be used in combination in future multicenter radiomic studies with datasets acquired with various CT conditions. Furthermore, each method may serve as an appropriate tool for harmonization of inconsistencies resulting from specific CT parameters. Although it may be more practical to have a single method with which we can harmonize various CT parameters, we could not achieve that, and currently, there are no such methods introduced in literature. Therefore, this remains as a future work to identify a solution or a framework that encompasses standardization of several radiomic features against wide variation of sources of variability.

A limitation in exploring mitigation techniques was that we did not assess the harmonization performance of the GAN and the ComBat techniques on scans from conditions (or scanners) previously unseen by the model. For example, it is valuable to understand whether these harmonization techniques can harmonize non-reference conditions other than the seven conditions in our study. This future work matters as it indicates whether we will need a statistically representative sample and separate runs of ComBat or GAN training for harmonization of any other non-reference condition.

As was described in Chapter 6, we trained seven condition-specific GAN models for transformation and normalization of each non-reference condition. While our study provides a proof of concept for potential of the GAN application, further work is needed to make it practical. For example, a limitation here is that each of the GAN models in this work only perform harmonization of impact of one CT parameter. Therefore, future work should include either combination of harmonization of several non-reference conditions (ideally with more than one non-reference CT parameter) in one model or in an ensemble of different GAN models.

Finally, in assessing agreement of radiomic features, we made an assumption regarding an acceptable and strong agreement to determine reproducibility of radiomic features (i.e. threshold of 0.9). This assumption has been made according to the choice of threshold of other similar research in literature[60,62,169]. However, there are no recommendations available regarding an optimal threshold for CCC or OCCC. Since the metrics of CCC and OCCC measure the similarity of two or more measurements by calculating deviations from the line of identity, the selected threshold of 0.9 was a very conservative threshold. Therefore, it is also of interest to perform a sensitivity analysis of the agreement threshold. To come up with an optimal threshold, it will be helpful if the sensitivity analysis is performed with respect to the diagnostic task of interest. For example, to determine a meaningful reproducibility threshold we can select a threshold that maintains both prediction performance (e.g., in terms of area under the ROC curve) and an agreement of more than moderate (i.e. with CCC of at least 0.8).

9.4 Conclusion

As a summary, in this dissertation, we explored the variability and the correctability of a set of well-known lung nodule radiomic features in response to variation of three CT image acquisition and reconstruction parameters of dose, kernel, and slice thickness. The work presented here constitutes a widely applicable experimental technique and methodology for assessing the robustness of radiomic features before or after application of a mitigation technique.

According to the findings of the work presented, we advise against pooling of the radiomic features acquired from images with inconsistent scan protocols. We recommend inclusion of a standardization step, either in form of robustness-based feature selection or in form of application of harmonization algorithms in future multicenter efforts. With further validations, we foresee that techniques such as GAN or ComBat can, separately or jointly, play a role in future efforts to minimize inconsistencies of radiomic features either in the image domain or in the feature domain. Our study contributes to the development of a framework that handles reliability issues of radiomic features. Once such a framework is adopted in future radiomic studies, deployment of multicenter clinical trials allows for extensive validation of clinical relevance of radiomic features. Such trials can eventually lead to expansion of the use of radiomic features into clinical practice, and ultimately enabling precision medicine through disease detection, stratification, and outcome prediction.

Appendix A Additional Theory and Results for Chapter 4

A.1 Settings for radiomic feature calculation

A.1.1 Settings used for Gray Level Co-Occurrence Matrix (GLCM) features:

The matrix was acquired at 16 quantized gray levels through using an 8-connected neighborhood in 2D with the corresponding direction vectors at distance 1 of: (1, 0, 0), (1, 1, 0), (0, 1, 0) and (−1, 1, 0).

A.2 Calculation of the Metric of Reproducibility: Overall Concordance Correlation Coefficient

The Concordance Correlation Coefficient (CCC) between values of a radiomic feature (Y) of n samples at two different CT image conditions $((Y_{i1}, Y_{i2}), i = 1, 2, \dots, n)$ with means of $\mu_1 = E[Y_1], \mu_2 = E[Y_2]$ and variances of $\sigma_1{}^2, \sigma_2{}^2$ and correlation coefficient of ρ is:

$$\rho_c = 1 - \frac{\mathrm{E}[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

or,

$$\rho_c = CCC = \frac{2\rho\sigma_1\sigma_2}{\sigma_1{}^2 + \sigma_2{}^2 + (\mu_1 - \mu_2)^2}$$

Hence CCC is the agreement of feature (Y) within a pair of CT image conditions. The Overall Concordance Correlation Coefficient (OCCC) of a radiomic feature (Y) of n samples across M different CT image conditions $(Y_1, Y_2, \dots, Y_M)$ is the weighted average of all pairwise CCCs where higher weights are given to the pairs of conditions whose feature values have higher variances and larger mean differences:

$$OCCC = \frac{\sum_{x=1}^{M-1}\sum_{y=x+1}^{M}\xi_{xy}\rho_{xy}\chi_{xy}}{\sum_{x=1}^{M-1}\sum_{k=x+1}^{M}\xi_{xy}}$$

Where $\xi_{xy} = \sigma_x{}^2 + \sigma_y{}^2 + (\mu_x - \mu_y)^2$ and $\chi_{xy} = \frac{2\sigma_x\sigma_y}{\sigma_x{}^2 + \sigma_y{}^2 + (\mu_x - \mu_y)^2}$.

Hence agreement (OCCC) of a radiomic feature against variation of dose (across M=4 levels) in a subset (e.g. at fixed kernel $k_1$and slice thickness st_2) is:

$$OCCC_{d.k_1\_st_2} = \frac{\sum_{x=1}^{3}\sum_{y=x+1}^{4}\xi_{xy}\rho_{xy}\chi_{xy}}{\sum_{x=1}^{M-1}\sum_{k=x+1}^{M}\xi_{xy}}$$

Agreement (OCCC) of a radiomic feature against variation of kernel (across M=3 levels) in a subset (e.g. at fixed dose $d_{100}$and slice thickness $st_{0.6}$) is:

$$OCCC_{k.d_{100}\_st_{0.6}} = \frac{\sum_{x=1}^{2}\sum_{y=x+1}^{3}\xi_{xy}\rho_{xy}\chi_{xy}}{\sum_{x=1}^{M-1}\sum_{k=x+1}^{M}\xi_{xy}}$$

Agreement (OCCC) of a radiomic feature against variation of slice thickness (across M=3 levels) in a subset (e.g. at fixed dose $d_{50}$and kernel k_2) is:

$$OCCC_{st.d_{50}\_k_2} = \frac{\sum_{x=1}^{2} \sum_{y=x+1}^{3} \xi_{xy} \rho_{xy} \chi_{xy}}{\sum_{x=1}^{M-1} \sum_{k=x+1}^{M} \xi_{xy}}$$

A.3 Assessing Reproducibility of Radiomic Features across Variation of CT Parameters in Images with Iterative Reconstruction

To acquire scans of subjects with an iterative algorithm, the simulated low dose raw data was taken back to the scanner, and images were reconstructed with the CT parameters in Table A-1. CT parameters used for iterative (SAFIRE) reconstructions at the scanner using the scanner reconstruction software's SAFIRE ("Sinogram Affirmed Iterative Reconstruction," Siemens Healthineers, Forchheim, Germany) algorithm.

Table A-1. CT parameters used for iterative (SAFIRE) reconstructions at the scanner

	Dose Level	**Slice Thickness**	**Reconstruction Kernel**[b]
CT parameter ranges	100%[a], 50%, 25%, 10%	2mm, 1mm,0.6mm	Smooth (k1), Medium (k2), Sharp (k3)

[a] 100% dose level represents the standard lung cancer screening dose with CTDIvol ≅2mGy

[b] k1: I26/3, k2: I44/3), k3: I50/3

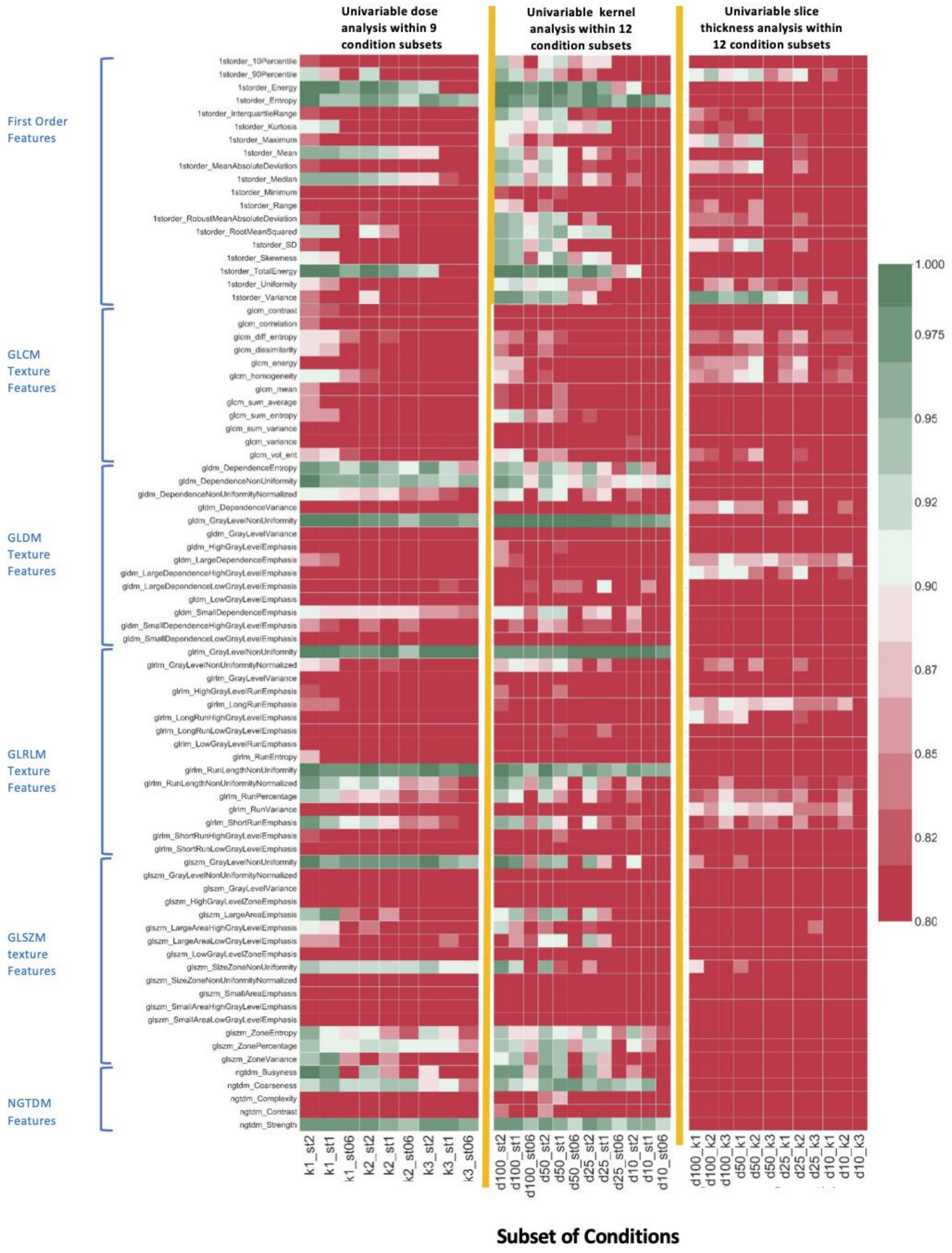


Figure A-1. Agreement (OCCC) of radiomic features in iterative reconstructions within condition subsets. a) impact of dose variation, b) impact of kernel variation, c) impact of slice thickness variation as shown by colors defined by the colormap. Colors in each column show agreements of radiomic features within the subset that is identified on the horizontal axis (e.g. $k_1\_st_2$ shows impact of dose variation at k_1 kernel and 2mm thickness). OCCC≤0.8 values were cut off at dark red color as it indicates very poor agreements.

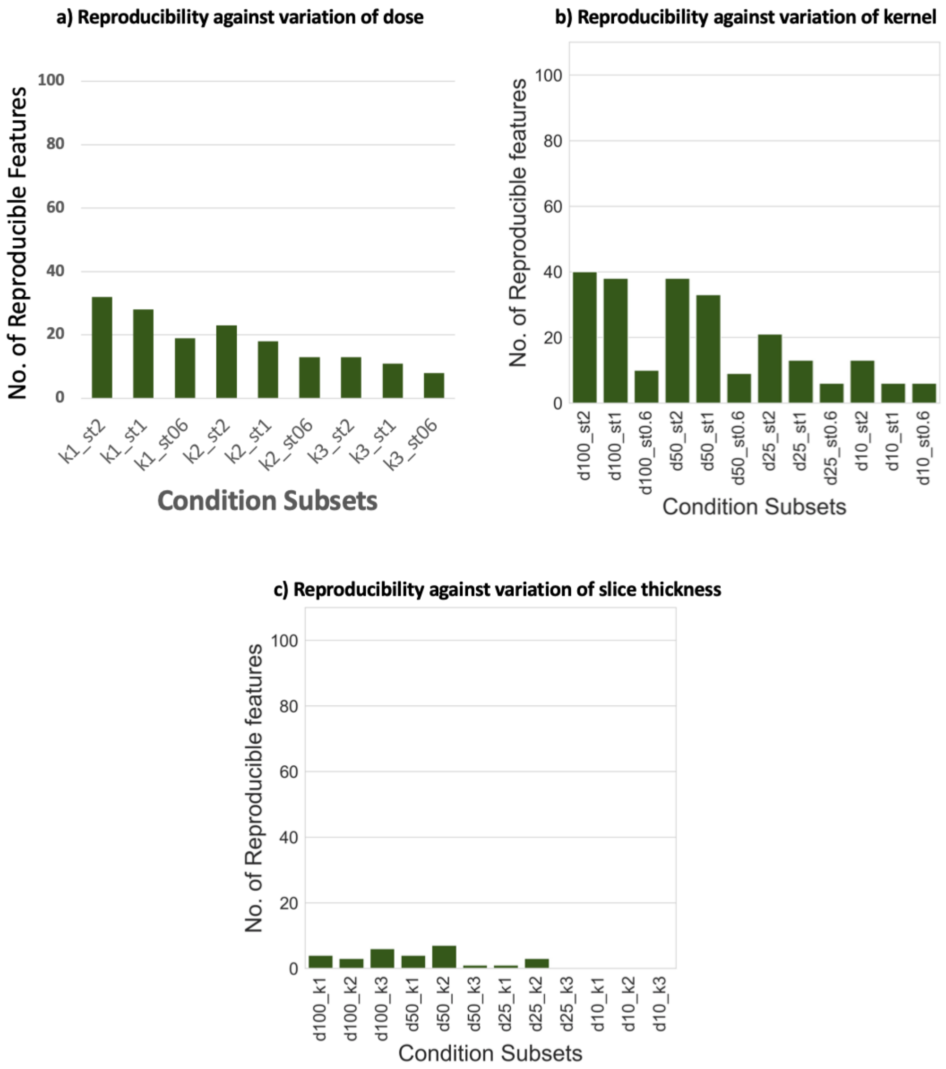


Figure A-2. Number of reproducible features within each condition subset in iterative reconstructions. (a) variation of dose in subsets with constant kernel and slice thickness, (b) variation of kernel in subsets with constant dose and slice thickness, (c): variation of slice thickness in subsets with constant dose and kernel.

Table A-2. Radiomic features of iterative reconstructions: features that were reproducible after dose and kernel variations in all the corresponding subsets

Feature type	**Reproducible against dose variations in all $k_i\_st_j$ subsets**	**Reproducible against kernel variations in all $d_i\_st_j$ subsets**
First order	Entropy	Entropy
GLRLM	Gray Level Non-Uniformity	Gray Level Non-Uniformity
	Run Length Non-Uniformity	Run Length Non-Uniformity
GLDM	Dependence Non-Uniformity	Gray Level Non-Uniformity
	Gray Level Non-Uniformity	-
NGTDM	Strength	Strength
GLSZM	Gray Level Non-Uniformity	-
	Size Zone Non-Uniformity	-

Appendix B Additional Results and Explanations for Chapter 6

B.1 Additional Tables and Plots

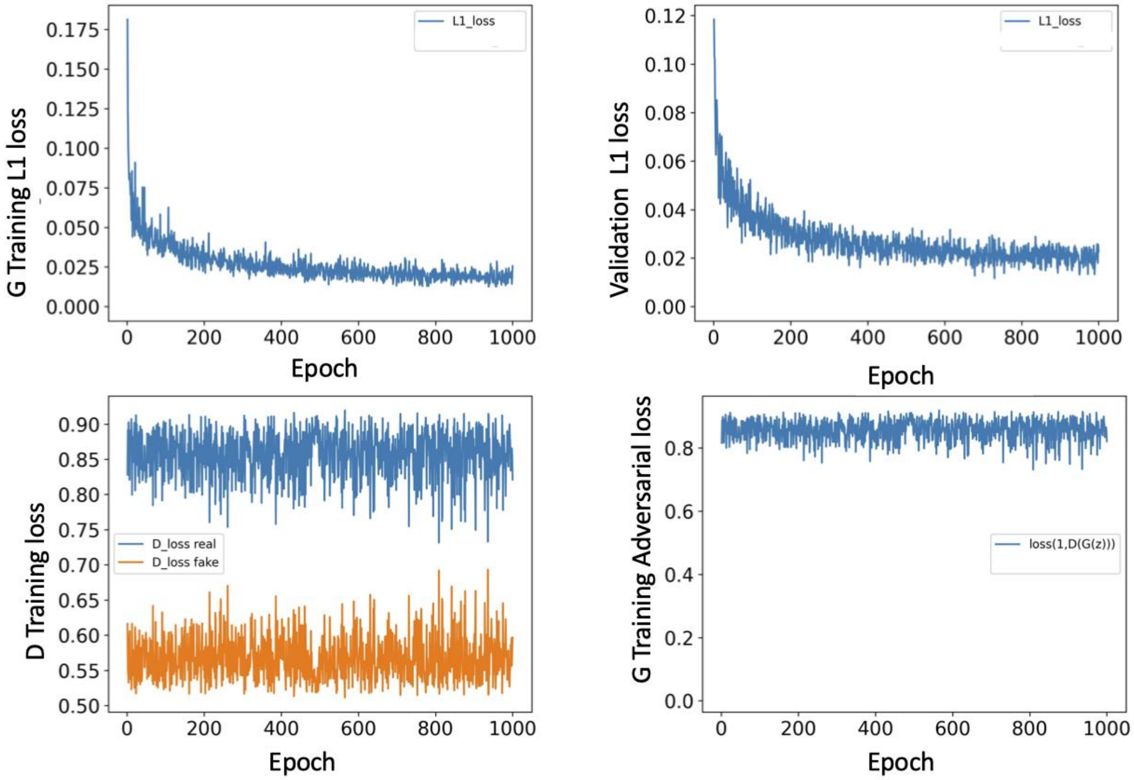


Figure B-1. Learning curves for the discriminator and generator and the loss curve of the validation cases during the training

Table B-1. Name of the selected uncorrelated features

Feature Group	Feature Name	Feature Index
GLRLM	Gray Level Non-Uniformity	1
	Gray Level Non-Uniformity Normalized	2
	Gray Level Variance	3
	High Gray Level Run Emphasis	4
	Long Run Low Gray Level Emphasis	5
	Run Entropy	6
GLSZM	Large Area Low Gray Level Emphasis	7
	Size Zone Non-Uniformity Normalized	8
GLCM	Autocorrelation	9
	Correlation	10
	Imc1	11
First order	10Percentile	12
	Energy	13
	Interquartile Range	14
GLDM	Large Dependence Low Gray Level Emphasis	15

Table B-2. Name of features for each index corresponding to the horizontal axis in Figure B-4 and Figure B-6

	Radiomic Feature		Radiomic Feature		Radiomic Feature
1	GLRLM GrayLevelNonUniformity	**32**	GLSZM ZoneVariance	**63**	1st order Maximum
2	GLRLM GrayLevelNonUniformityNormalized	**33**	GLCM Autocorrelation	**64**	1st order MeanAbsoluteDeviation
3	GLRLM GrayLevelVariance	**34**	GLCM ClusterProminence	**65**	1st order Mean
4	GLRLM HighGrayLevelRunEmphasis	**35**	GLCM ClusterShade	**66**	1st order Median
5	GLRLM LongRunEmphasis	**36**	GLCM ClusterTendency	**67**	1st order Minimum
6	GLRLM LongRunHighGrayLevelEmphasis	**37**	GLCM Contrast	**68**	1st order Range
7	GLRLM LongRunLowGrayLevelEmphasis	**38**	GLCM Correlation	**69**	1st order RobustMeanAbsoluteDeviation
8	GLRLM LowGrayLevelRunEmphasis	**39**	GLCM DifferenceAverage	**70**	1st order RootMeanSquared
9	GLRLM RunEntropy	**40**	GLCM DifferenceEntropy	**71**	1st order Skewness
10	GLRLM RunLengthNonUniformity	**41**	GLCM DifferenceVariance	**72**	1st order TotalEnergy
11	GLRLM RunLengthNonUniformityNormalized	**42**	GLCM Id	**73**	1st order Uniformity
12	GLRLM RunPercentage	**43**	GLCM Idm	**74**	1st order Variance
13	GLRLM RunVariance	**44**	GLCM Idmn	**75**	NGTDM Busyness
14	GLRLM ShortRunEmphasis	**45**	GLCM Idn	**76**	NGTDM Coarseness
15	GLRLM ShortRunHighGrayLevelEmphasis	**46**	GLCM Imc1	**77**	NGTDM Complexity
16	GLRLM ShortRunLowGrayLevelEmphasis	**47**	GLCM Imc2	**78**	NGTDM Contrast
17	GLSZM GrayLevelNonUniformity	**48**	GLCM InverseVariance	**79**	NGTDM Strength
18	GLSZM GrayLevelNonUniformityNormalized	**49**	GLCM JointAverage	**80**	GLDM DependenceEntropy
19	GLSZM GrayLevelVariance	**50**	GLCM JointEnergy	**81**	GLDM DependenceNonUniformity
20	GLSZM HighGrayLevelZoneEmphasis	**51**	GLCM JointEntropy	**82**	GLDM DependenceNonUniformityNormalized
21	GLSZM LargeAreaEmphasis	**52**	GLCM MCC	**83**	GLDM DependenceVariance
22	GLSZM LargeAreaHighGrayLevelEmphasis	**53**	GLCM MaximumProbability	**84**	GLDM GrayLevelNonUniformity
23	GLSZM LargeAreaLowGrayLevelEmphasis	**54**	GLCM SumAverage	**85**	GLDM GrayLevelVariance
24	GLSZM LowGrayLevelZoneEmphasis	**55**	GLCM SumEntropy	**86**	GLDM HighGrayLevelEmphasis
25	GLSZM SizeZoneNonUniformity	**56**	GLCM SumSquares	**87**	GLDM LargeDependenceEmphasis
26	GLSZM SizeZoneNonUniformityNormalized	**57**	1st order 10Percentile	**88**	GLDM LargeDependenceHighGrayLevelEmphasis
27	GLSZM SmallAreaEmphasis	**58**	1st order 90Percentile	**89**	GLDM LargeDependenceLowGrayLevelEmphasis
28	GLSZM SmallAreaHighGrayLevelEmphasis	**59**	1st order Energy	**90**	GLDM LowGrayLevelEmphasis
29	GLSZM SmallAreaLowGrayLevelEmphasis	**60**	1st order Entropy	**91**	GLDM SmallDependenceEmphasis
30	GLSZM ZoneEntropy	**61**	1st order InterquartileRange	**92**	GLDM SmallDependenceHighGrayLevelEmphasis
31	GLSZM ZonePercentage	**62**	1st order Kurtosis	**93**	GLDM SmallDependenceLowGrayLevelEmphasis

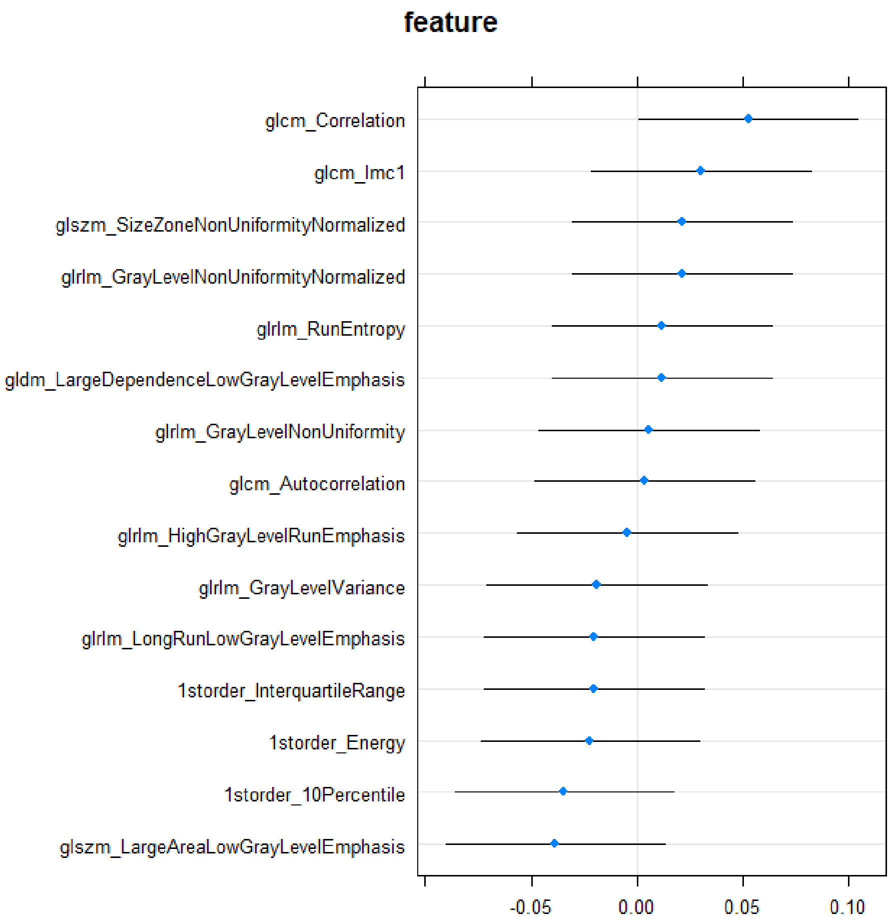


Figure B-2. 95% confidence interval for the estimated random effects for each feature in the mixed-effect model after GAN harmonization

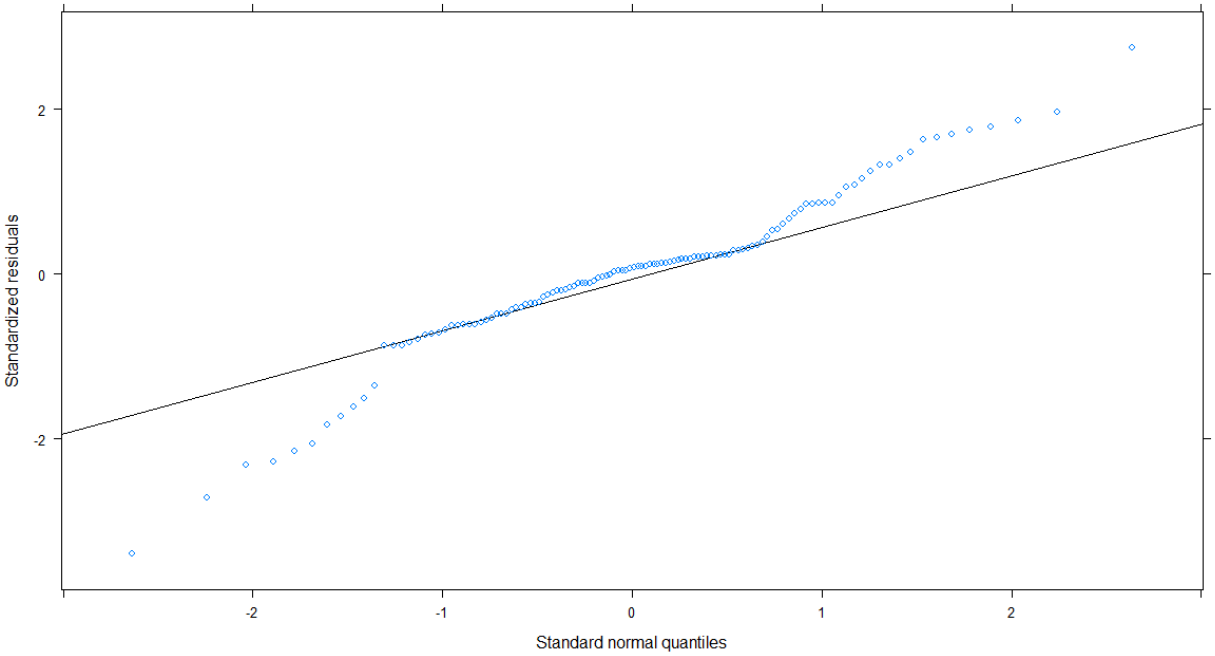


Figure B-3. Quantile-quantile plot to check the assumption of normality of residuals in the linear mixed-effect model

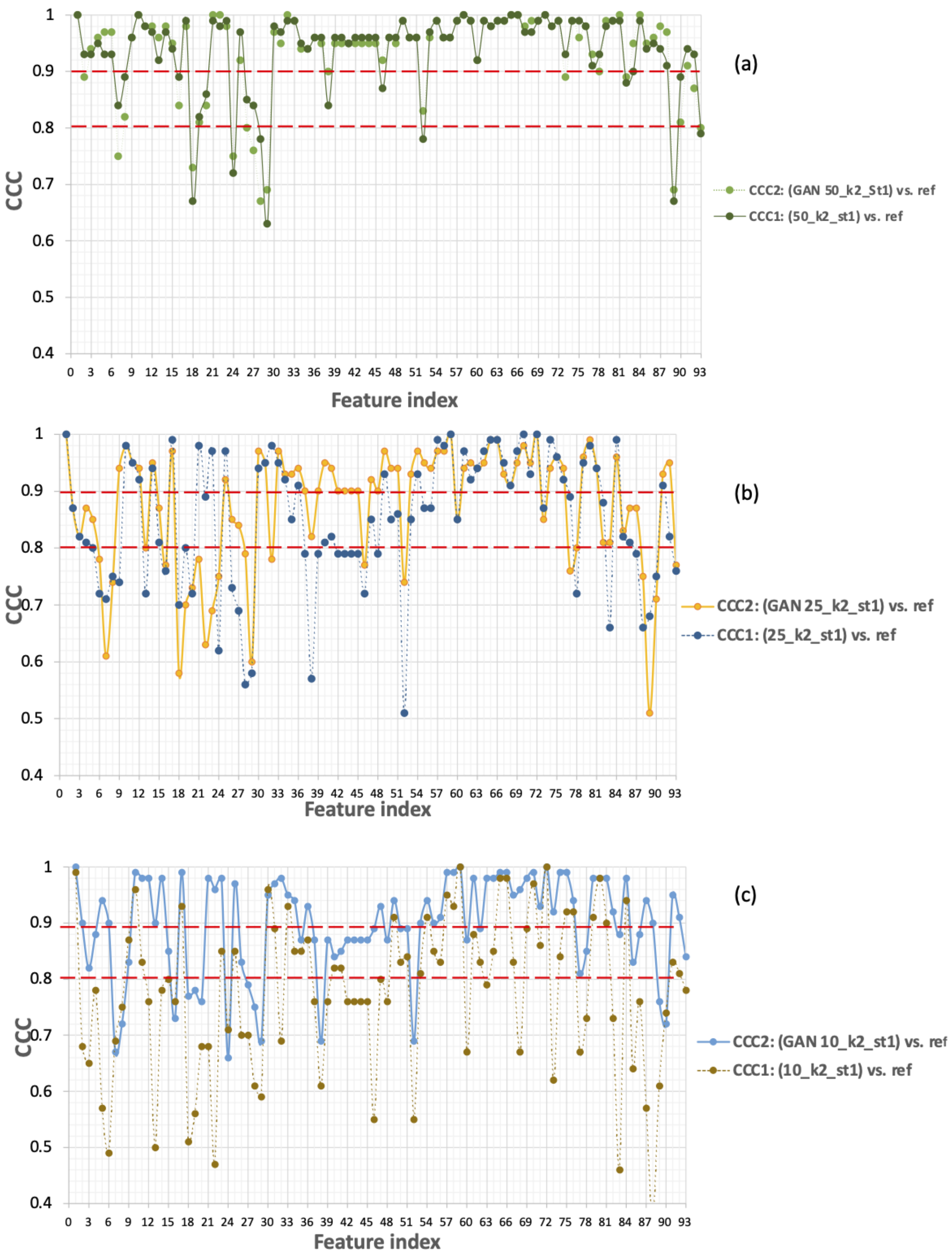


Figure B-4. CCC_1 for 15 uncorrelated features and the corresponding CCC_2 values for non-reference conditions with different dose level than reference. Harmonization of (a) 50% dose, (b) 25% dose, (c) 10% dose. The red dashed lines show the CCC cut-off values for moderate and strong agreements. Name of features are indicated in Table 2.

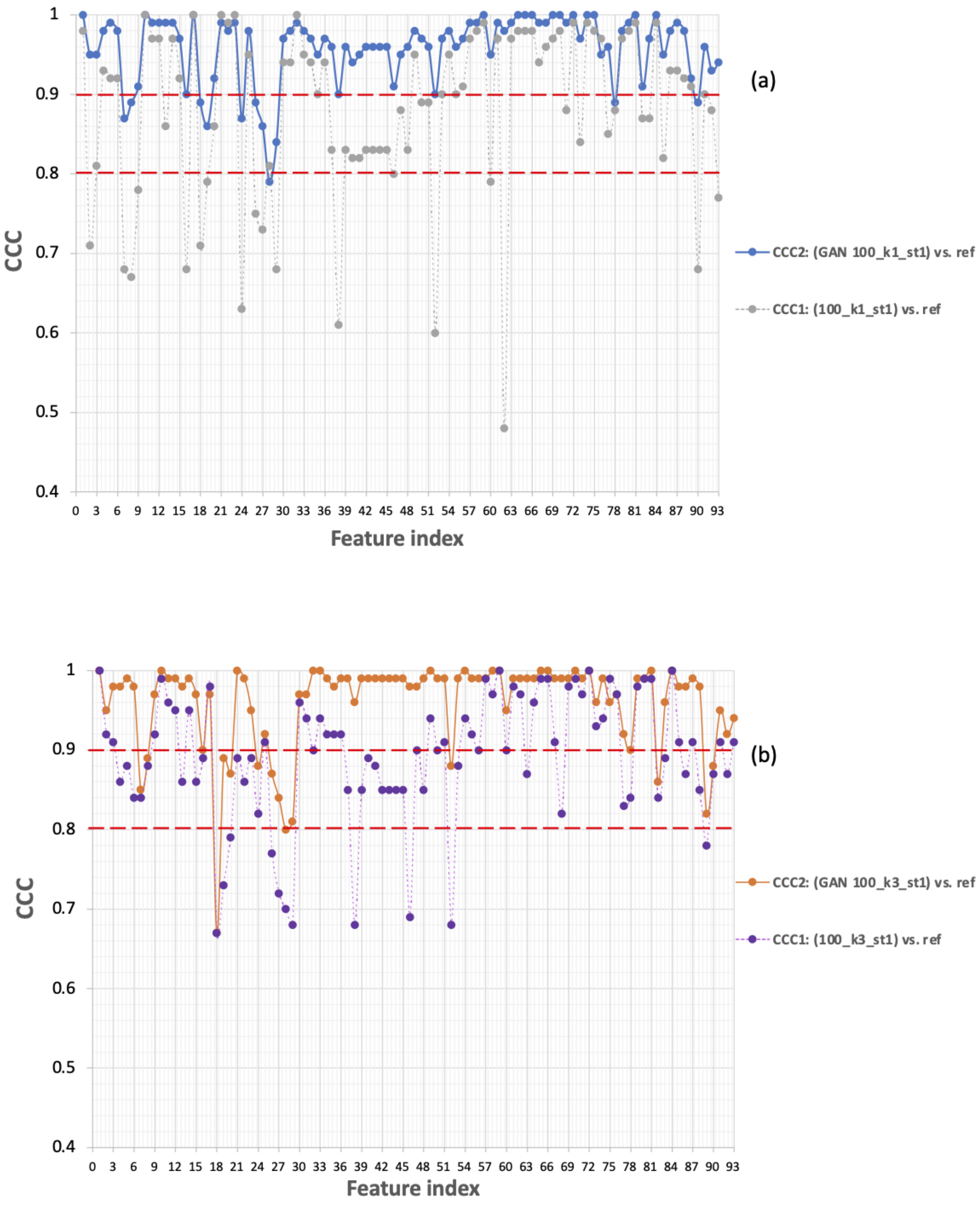


Figure B-5. CCC_1 and CCC_2 values for conditions with different kernel than reference. Harmonization of (a) smooth kernel, (b) sharp kernel. The red dashed lines show the CCC cut-off values for moderate and strong agreements. Name of features are indicated in Table 2.

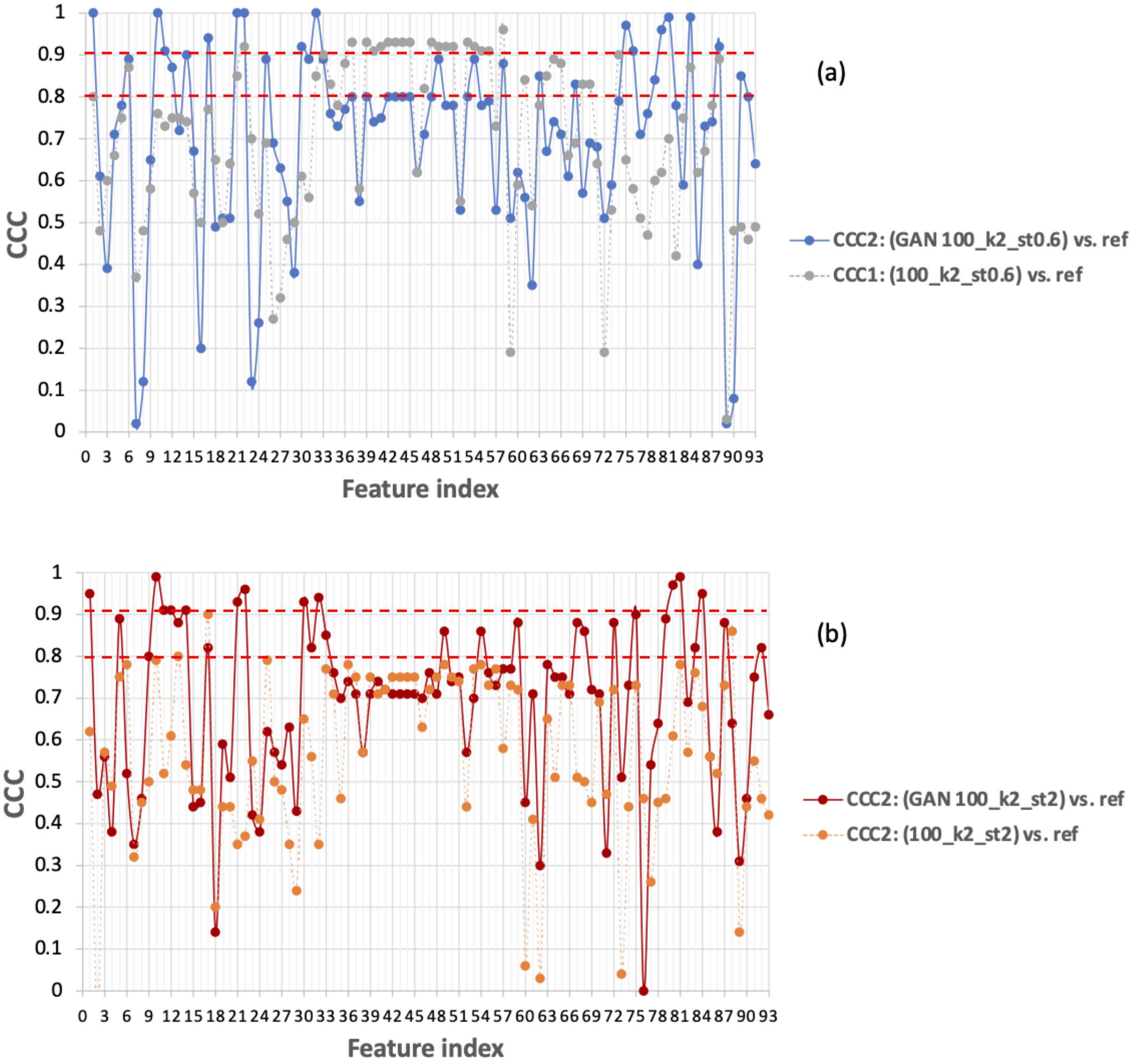


Figure B-6. CCC_1 and CCC_2 values for conditions with different slice thickness than reference. Harmonization of (a) 0.6mm thickness, (b) 2mm thickness. The red dashed lines show the CCC cut-off values for moderate and strong agreements. Name of features are indicated in Table 2.

Table B-3. Radiomic features that were used for calculation of perceptual loss function during the training of GAN models

First-order	Mean, Entropy, Kurtosis, Energy, Range
GLRLM	Gray Level Variance, Run Entropy, Long Run Low Gray Level Emphasis
GLDM	Large Dependence Low Gray Level Emphasis Low Gray Level Emphasis
GLSZM	Small Area Emphasis, Size Zone Non-Uniformity Normalized, Zone Variance
NGTDM	Complexity, Contrast

Table B-4. Comparisons between agreements of radiomic features of the two conditions with different slice thickness: with resampling to 1mm and with application of GAN

	Non-reference Condition	**Average CCC difference**	$CCC_2 > CCC_1$	$CCC_1 \geq 0.9$	$CCC_2 \geq 0.9$	$C2 - C1(\%)^a$
Resampled Non-reference	100%, k2, st0.6	16.5%	57	20	24	4 (20%)
	100%, k2, st2	39.3%	53	1	1	0
Non-reference After GAN	100%, k2, st0.6	5.9%	43	20	15	-5 (-25%)
	100%, k2, st2	60%	64	1	13	12(1200%)

B.2 Additional Explanation and Visualization for GAN Model Training

In Chapter 6 we found that for some conditions (e.g., condition with sharp kernel k3), the trained GAN model was able to significantly improve agreement of radiomic features with the reference condition. For some conditions (e.g., conditions with 25% or 50% dose) we saw minimal impact of GAN harmonization on radiomic features.

In this section we show the loss values (or the learning curves) for the generator (G) and the discriminator (D) of the GAN model that was trained on three non-reference conditions (Figure B-7 to Figure B-8). This provides an insight into the training process for each GAN model.

Details of the trained models have been presented in Chapter 6, Table 6-2. As shown in Figure B-7, the adversarial loss for the model trained on the condition with 50% dose level is the largest and for example the adversarial loss for the model trained on the condition with k3 kernel is the less than the condition with 50%. The lower the adversarial loss, the higher is the generator's ability in generating realistic images that match the target image (i.e., image at reference condition). This is in correlation with our finding about the performance of GAN in harmonization of radiomic features; when the GAN model trained on condition with k3 was applied on the test data with k3 kernel, we observed significant improvement of agreement (CCC) of radiomic features with the reference condition. However, for 50% dose we saw minimal improvements of CCC values. Additionally, Figure B-8, demonstrates the discriminator curves for real and fake loss. It is visible that the fake loss for condition with 50% dose is the least compared to the two other conditions. This indicates that the discriminator of this GAN model was better able to distinguish the fake images. This again suggest the inferiority of the generator in condition with 50% compared to the two other conditions.

One observation that can justify the inferiority of the generator in Figure B-7(a) and Figure B-8(a) is that the discriminator loss in Figure B-8(a) is quite consistent and very close to an average value of $\cong 1.4$. When the discriminator's loss is constant or is not changing from epoch to epoch, it indicates that the discriminator is not being challenged enough. When the discriminator is not challenged, it will not provide sufficient (or useful) feedback to the generator, so the generator does not learn to improve itself.

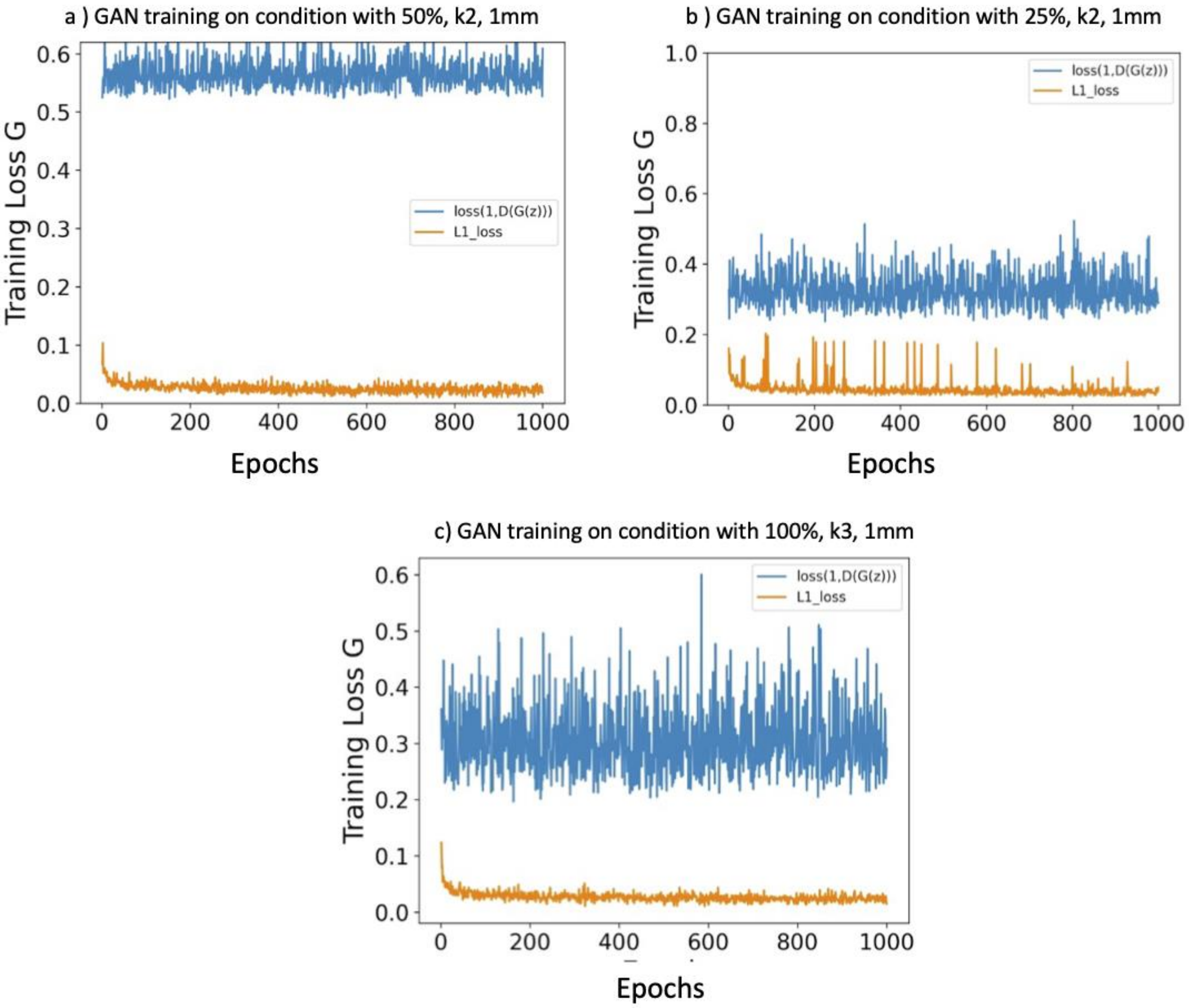


Figure B-7. Adversarial and L1 loss of the generator (G) during training for three non-reference conditions. loss(1,D(G(z)) stands for) Adversarial loss according to Eq. 1 in Chapter 6.

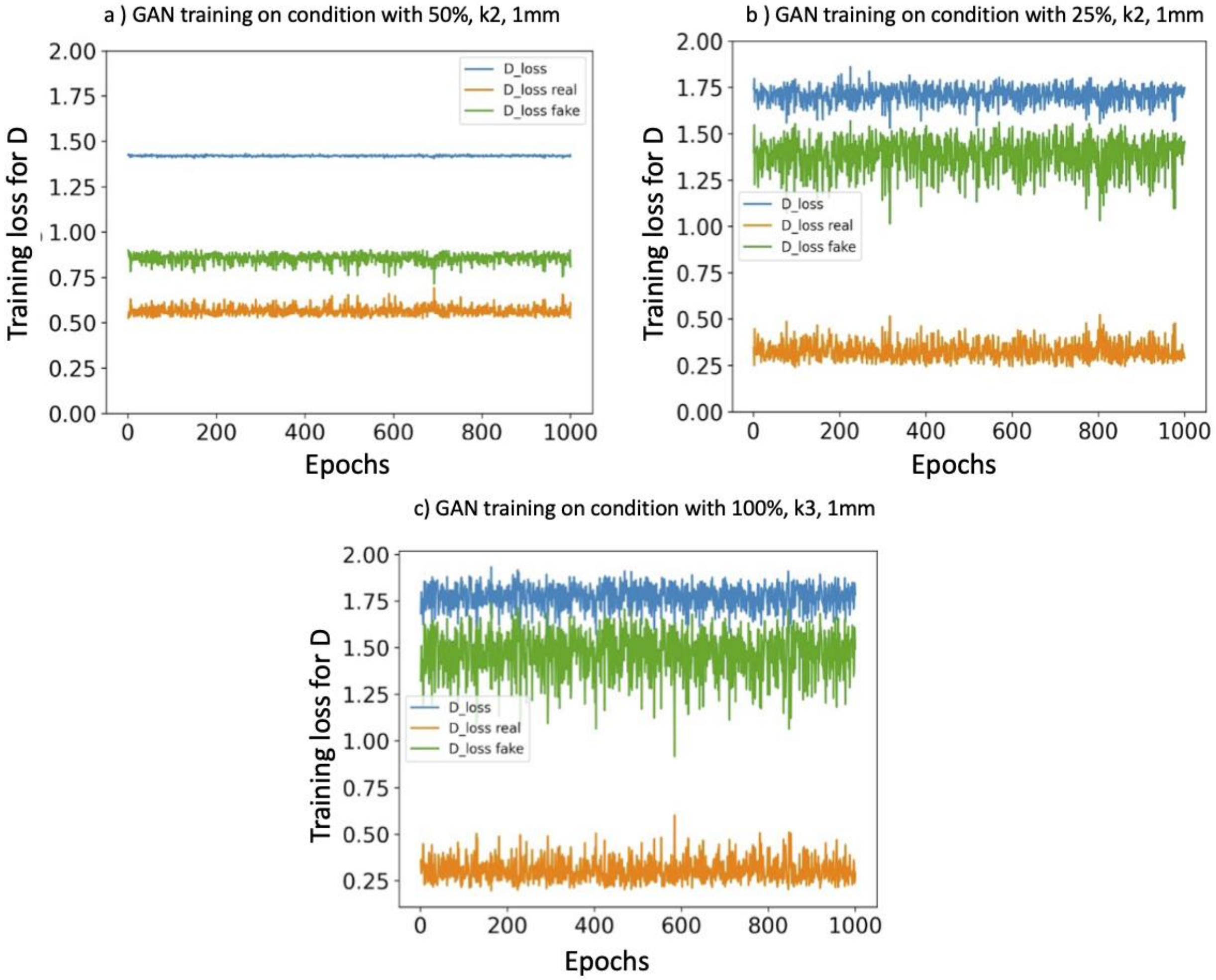


Figure B-8. Loss values for discriminator (D) during training of GAN for three non-reference condition. ‘D_loss’ is the sum of losses for real and fake images.

Appendix C Additional Results for Chapter 7

C.1 Additional Results

In harmonizing radiomic features in the condition with 50% dose, harmonization did not substantially affect the radiomic features. Figure C-1(a) also indicates the similarity of CCC_1 and CCC_2 values for this condition. According to Figure C-1 (b) versus (c), both of the conditions with 25% and 10% dose have more radiomic features with $CCC_2 \geq 0.9$ after the harmonization. Although the condition with 10% dose level has few reproducible features after harmonization, we see a considerable improvement in its CCC values (according to Figure C-1 (c)).

Figure C-2 visualizes CCC_1 and CCC_2 values (vertical axis) before and after harmonization of kernel variation effect for 93 radiomic features (horizontal axis). The data points that shift to higher values in the vertical axis indicate that the harmonization has improved the agreement of radiomic features with the reference condition.

Figure C-3 visualizes CCC values before and after harmonizing variation of slice thickness. A couple of radiomic features have increased CCC values.

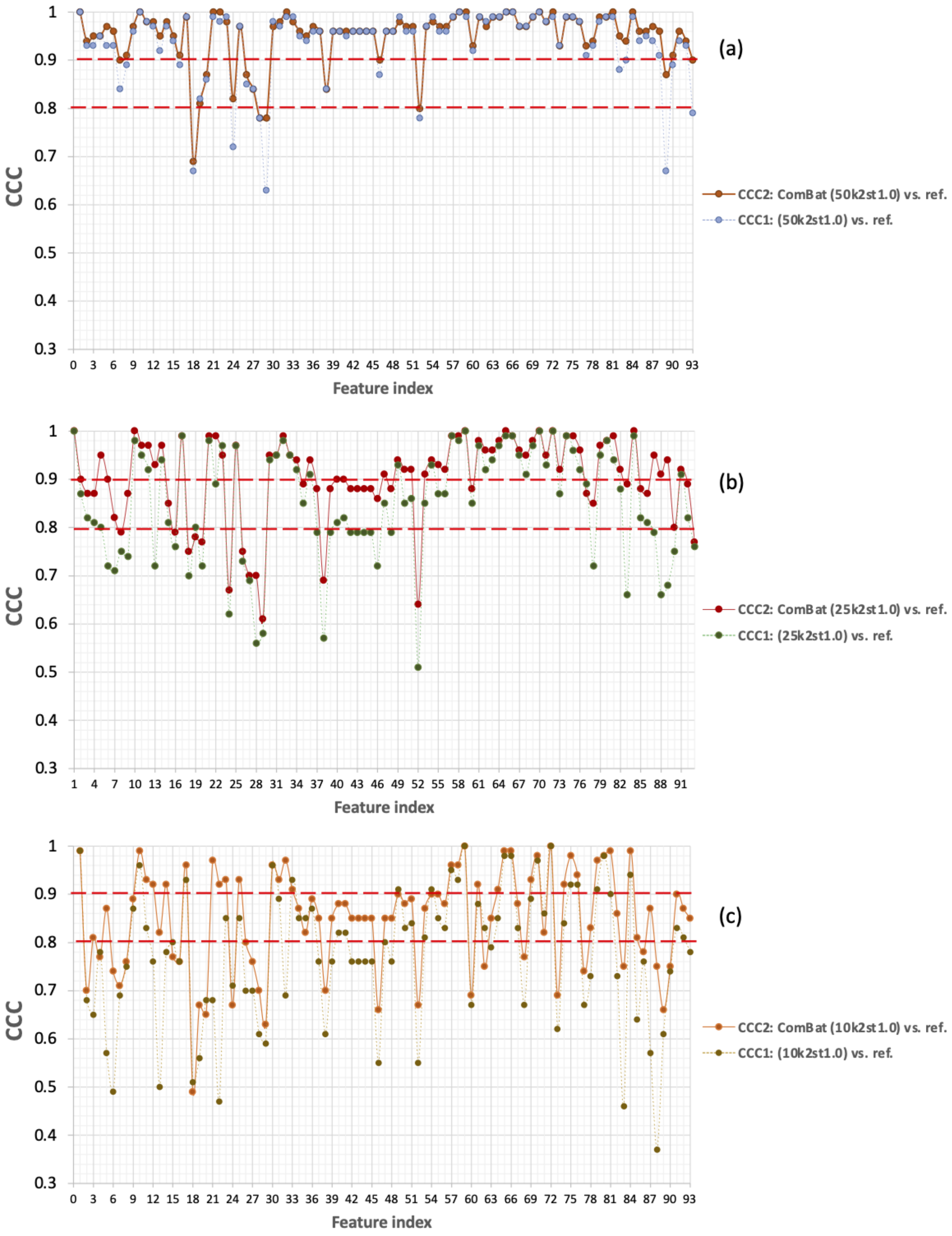


Figure C-1. CCC_1 for 15 uncorrelated features and the corresponding CCC_2 values for non-reference conditions with different dose level than reference. Harmonization of (a) 50% dose, (b) 25% dose, (c) 10% dose. The red dashed lines show the CCC cut-off values for moderate and strong agreements. Name of features are indicated in Table 1.

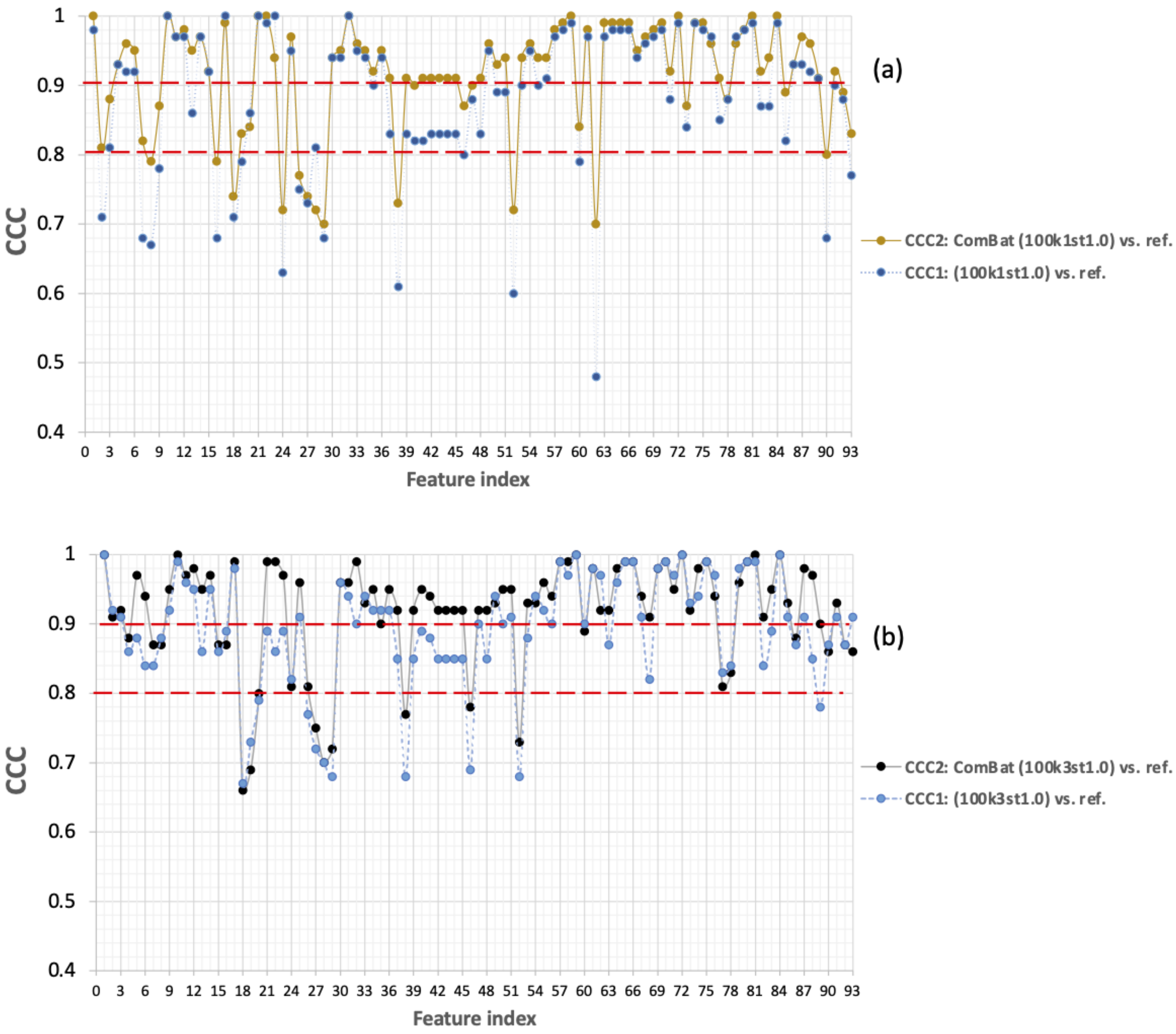


Figure C-2. CCC_1 and CCC_2 values for conditions with different kernel than reference. Harmonization of (a) smooth kernel, (b) sharp kernel. The red dashed lines show the CCC cut-off values for moderate and strong agreements. Name of features are indicated in Table 1.

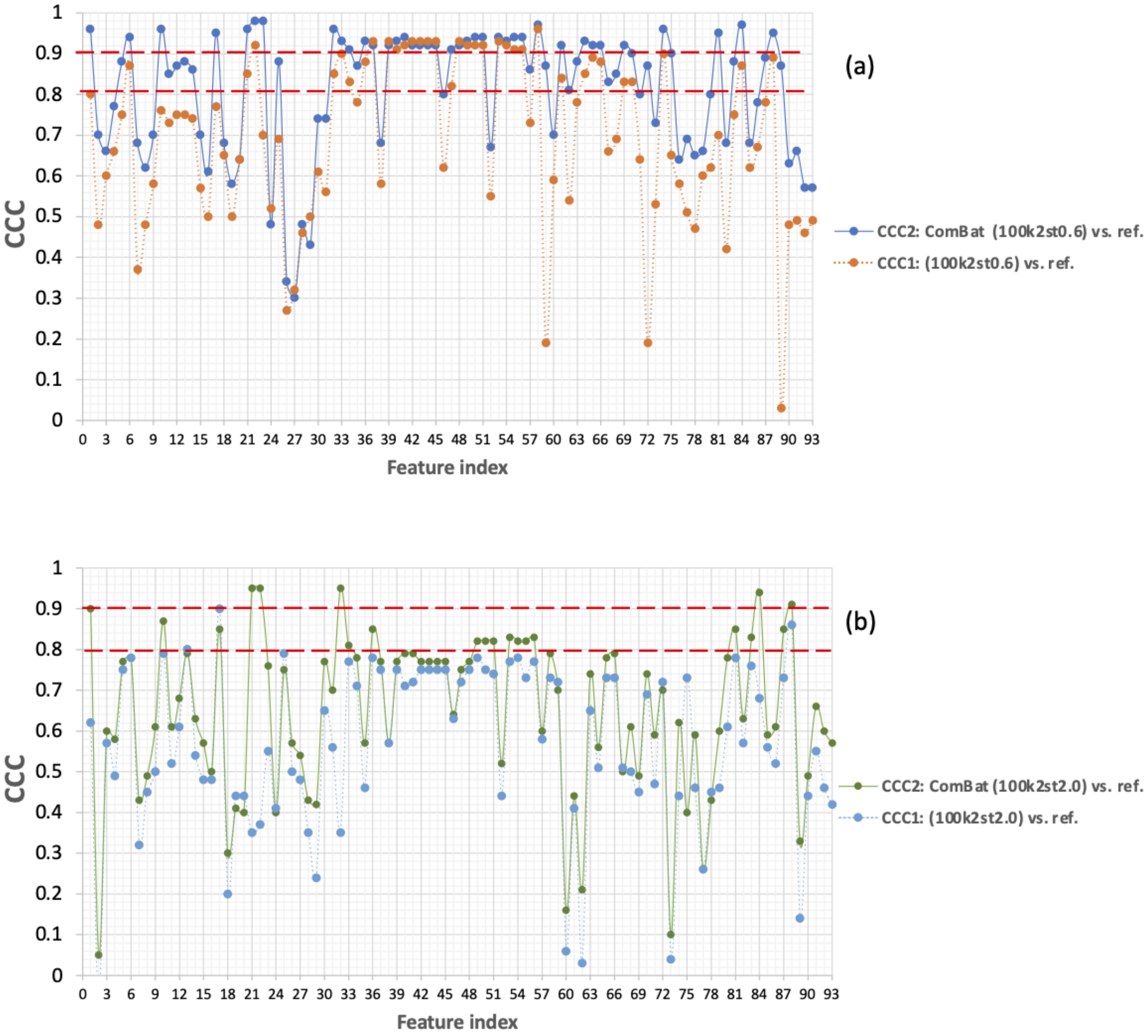


Figure C-3. CCC_1 and CCC_2 values for conditions with different slice thickness than reference. Harmonization of (a) 0.6mm thickness, (b) 2mm thickness. The red dashed lines show the CCC cut-off values for moderate and strong agreements.

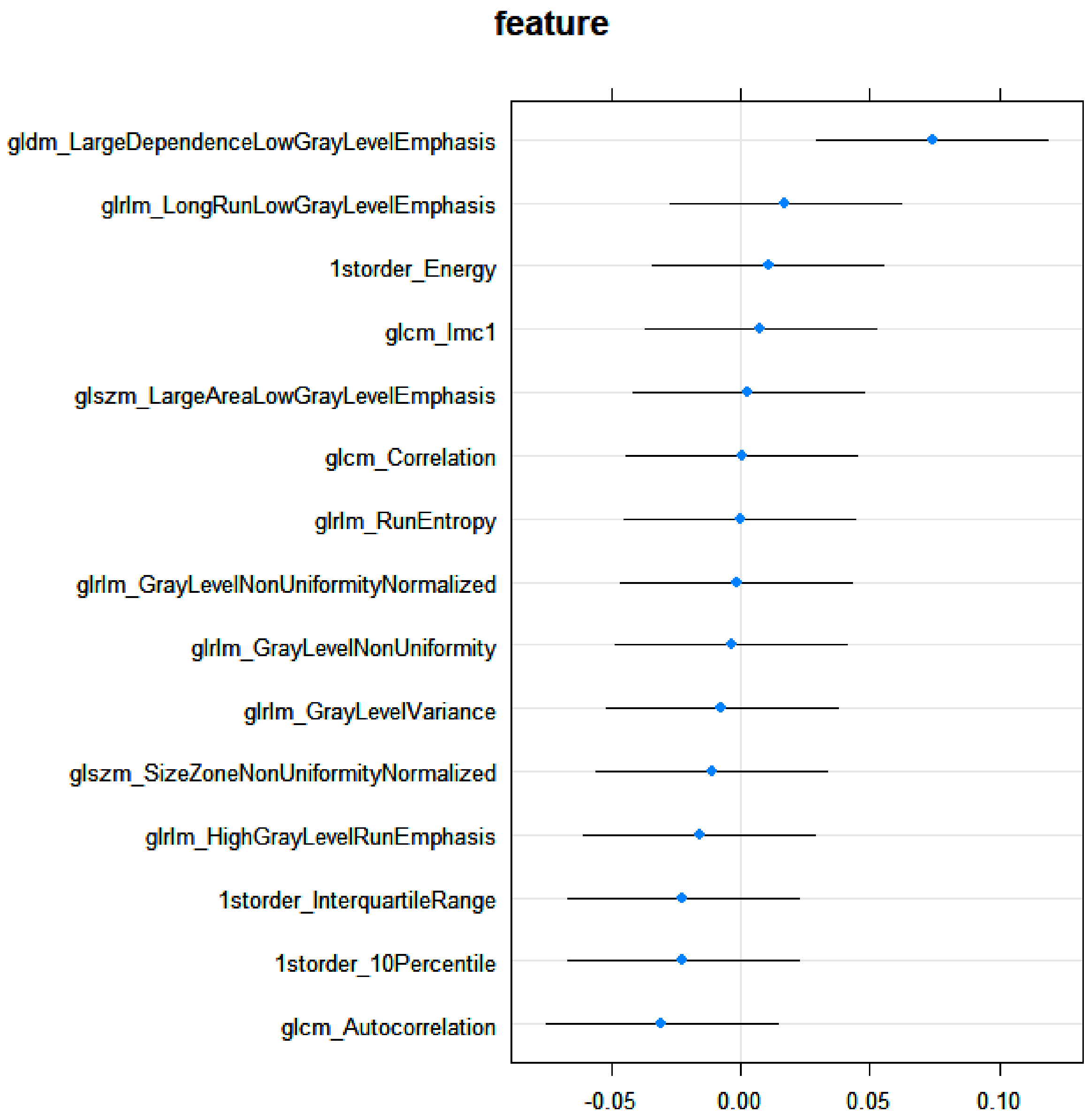


Figure C-4. Confidence Intervals for the random effects in the mixed-effect model after ComBat harmonization

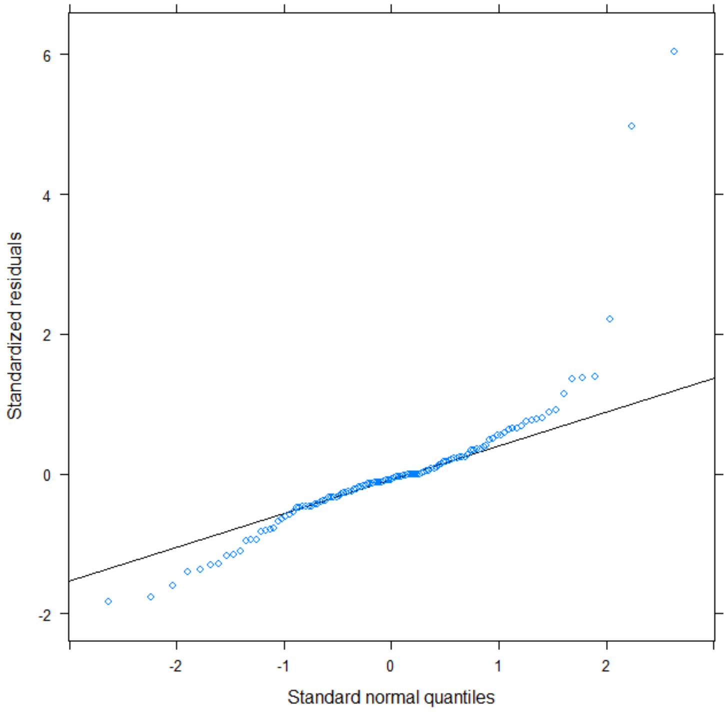


Figure C-5. Quantile-quantile plot to check assumption of normality of residuals in the linear mixed-effect model

C.2 Experimenting with ComBat performance for different sample sizes

This experiment was performed to understand how the performance of ComBat harmonization changes across different sizes. We gradually decreased the size of data from 134 to 20 and for each sample size, we performed a bootstrap sampling with replacement (for 100 times). ComBat was applied for each bootstrap sample to estimate and adjust the batch effects. After each ComBat application CCC values between radiomic features of harmonized non-reference and reference condition was calculated. At each sample size, a mean CCC and a confidence interval (CI) was calculated. Figure C-6Figure C-8 show the mean and CI for harmonization of seven non-reference conditions with different CT parameters of dose, kernel and slice thickness.

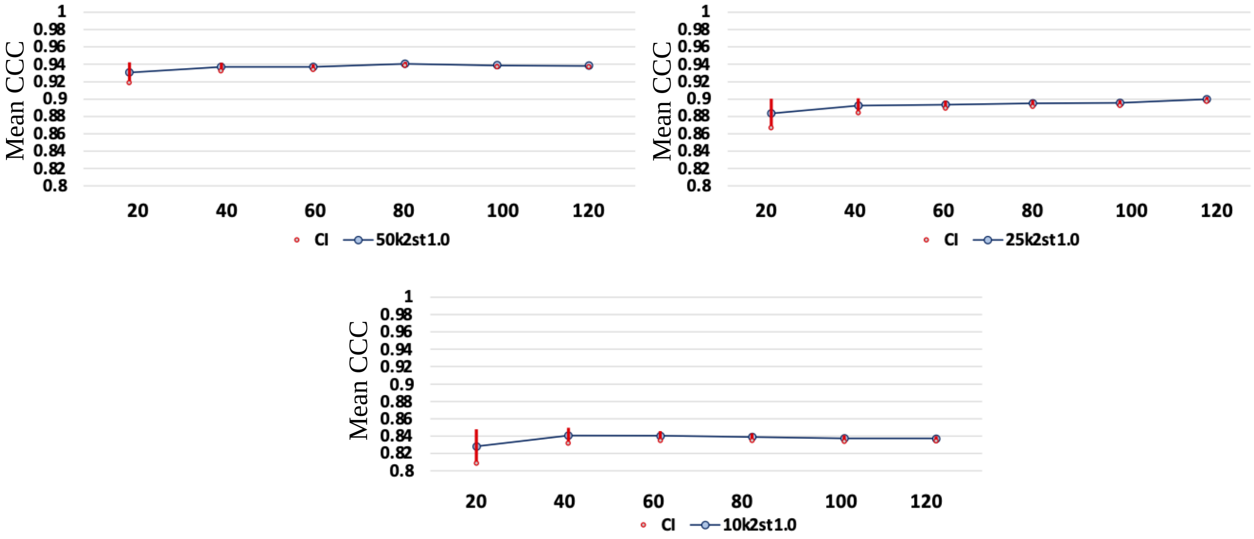


Figure C-6. Mean and CI for CCC values of radiomic features after ComBat application across 100 bootstrapped samples with sizes determined on the horizontal axis. Blue dots show the mean CCC and red lines show the CI.

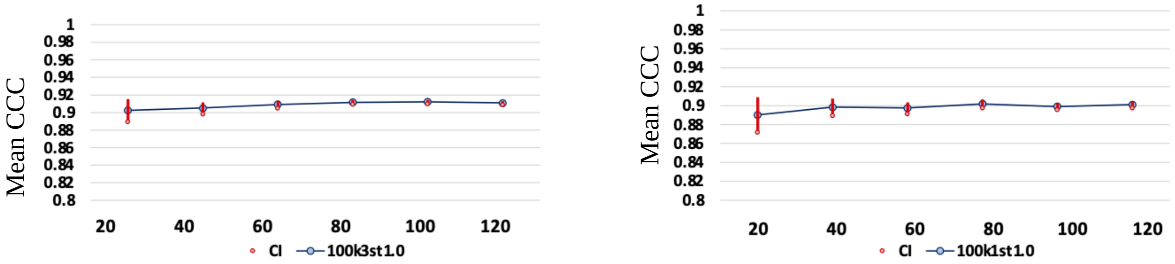


Figure C-7. Mean and CI for CCC values of radiomic features after ComBat application across 100 bootstrapped samples with sizes determined on the horizontal axis. Blue dots show the mean CCC and red lines show the CI.

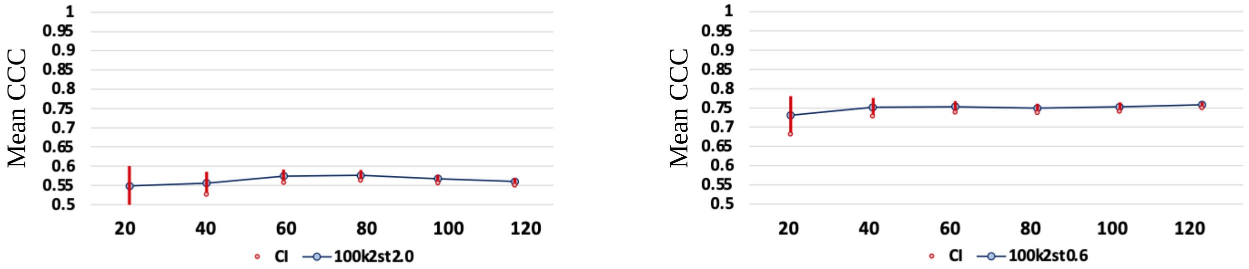


Figure C-8. Mean and CI for CCC values of radiomic features after ComBat application across 100 bootstrapped samples with sizes determined on the horizontal axis. Blue dots show the mean CCC and red lines show the CI.

Appendix D Additional Results for Chapter 8

Table D-1. Results from harmonization of 93 radiomic features using either GAN or ComBat method: No. of radiomic features with improved CCC values, the percentage of improvement of CCC, and No. of radiomic features that meet the cut-off value of 0.9 before and after harmonization

		100%, st1		k2, st1			100%, k2	
		k1	k3	50%	25%	10%	st0.6	st2
Unharmonized	$\boldsymbol{CCC_U \geq 0.9}$	51	50	75	40	20	20	1
GAN Harmonization	$\boldsymbol{CCC_{GAN} > CCC_U}$	85	83	31	68	85	43	64
	Average $\boldsymbol{CCC}$ difference	10.6%	8.2%	-0.1%	3.8%	19.3%	5.9%	60%
	$\boldsymbol{CCC_2 \geq 0.9}$	82	79	74	53	55	15	13
ComBat Harmonization	$\boldsymbol{CCC_{ComBat} > CCC_U}$	63	75	42	70	86	75	79
	Average $\boldsymbol{CCC}$ difference	2.1%	3.6%	1.02%	4.8%	10.2%	11.1%	12.6%
	$\boldsymbol{CCC_{ComBat} \geq 0.9}$	57	75	77	51	33	30	7

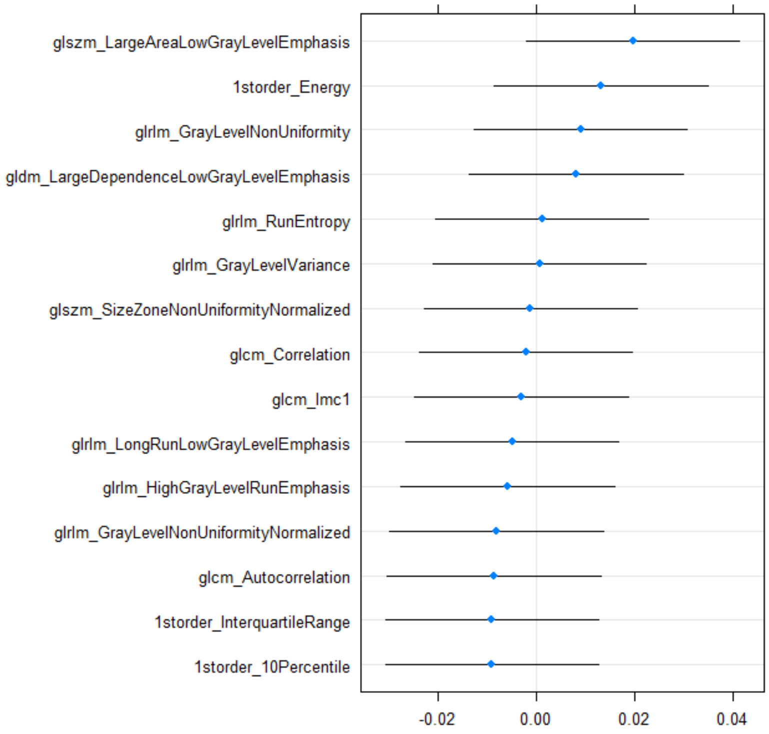


Figure D-1. The confidence interval for the random effect included for 15 radiomic features in the mixed effect model fit to the harmonized data by ComBat. Since the intervals lie around zero, there is not a large randomness in the model.

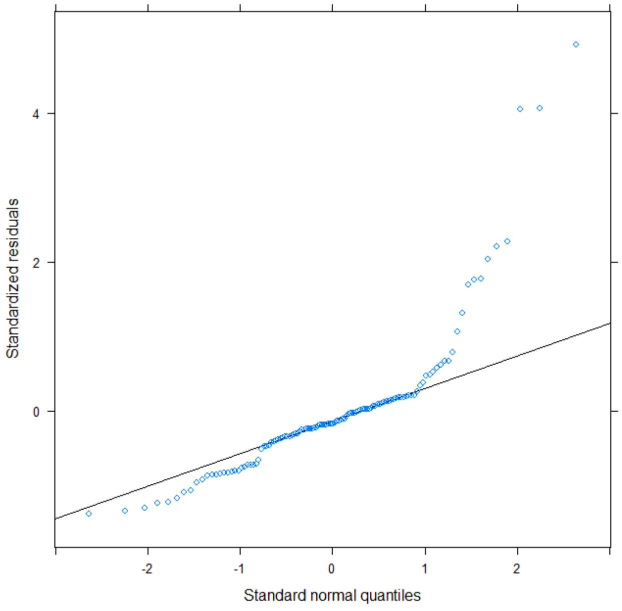


Figure D-2. QQPlot for the mixed-effect model fitted to the harmonized data after ComBat application on 55 cases.

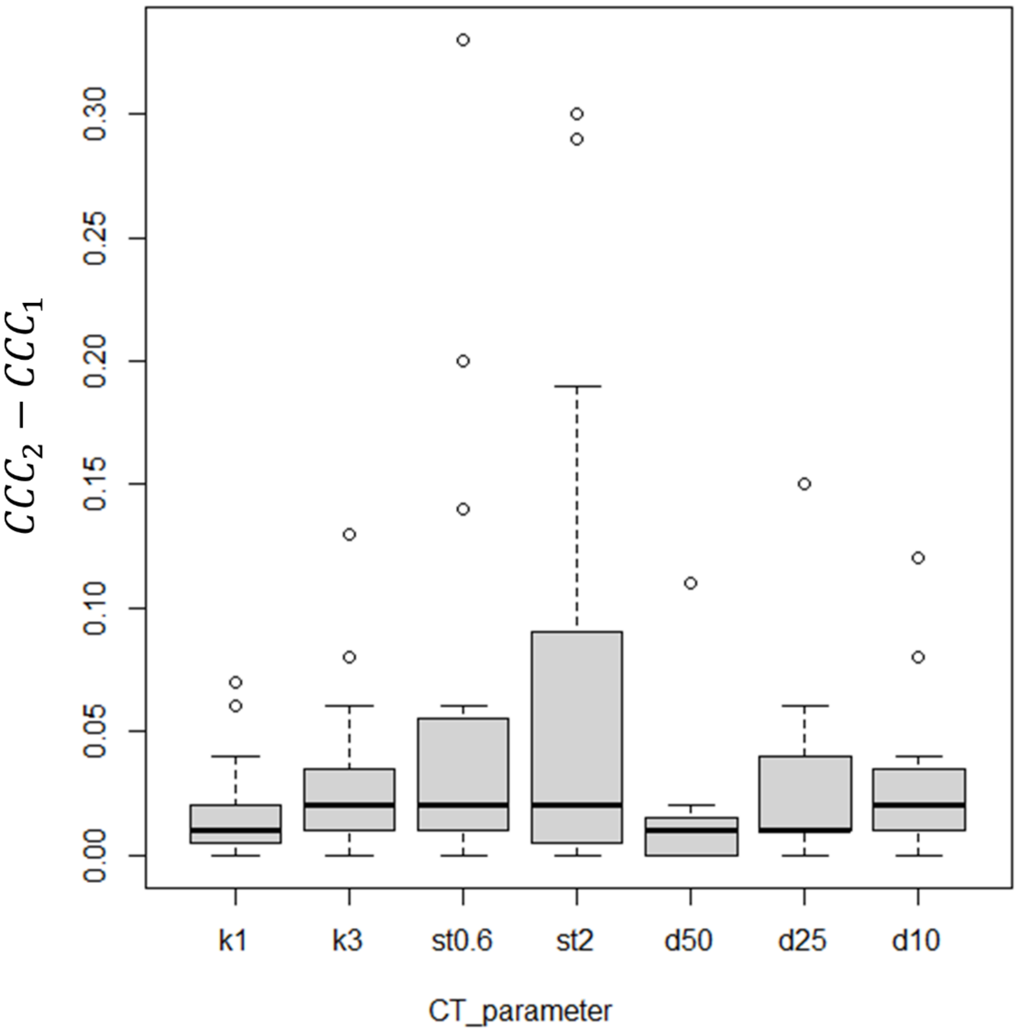


Figure D-3. Boxplots showing the range of differences in CCC before and after harmonization of 15 radiomic features using ComBat technique. Harmonization was done for seven non-reference condition. Conditions are identified by the level of the varying CT parameter. e.g., d50: (50%, k2, st1), d25: (25%, k2, st1), d10

Bibliography

1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036

2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. Published online 2016. doi:10.1148/radiol.2015151169

3. Lambin P, Leijenaar RT, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Published online 2017. doi:10.1038/nrclinonc.2017.141

4. Ravanelli M, Farina D, Morassi M, et al. Texture analysis of advanced non-small cell lung cancer (NSCLC) on contrast-enhanced computed tomography: prediction of the response to the first-line chemotherapy. doi:10.1007/s00330-013-2965-0

5. Fave X, Zhang L, Yang J, et al. Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer. doi:10.1038/s41598-017-00665-z

6. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. http://dx.doi.org/10.1038/ncomms5006

7. Mao L, Chen H, Liang M, et al. Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT screening. *Quant Imaging Med Surg*. 2019;9(2):263-272. doi:10.21037/qims.2019.02.02

8. Cherezov D, Hawkins SH, Goldgof DB, et al. Delta radiomic features improve prediction for lung cancer incidence: A nested case–control analysis of the National Lung Screening Trial. *Cancer Med*. 2018;7(12):6340-6356. doi:10.1002/cam4.1852

9. Hawkins S, Wang H, Liu Y, et al. Predicting Malignant Nodules from Screening CT Scans.

Published online 2016. doi:10.1016/j.jtho.2016.07.002

10. Schabath MB, Gillies RJ. *Noninvasive Quantitative Imaging - Based Biomarkers and Lung Cancer Screening*.; 2015. Accessed June 15, 2020. www.atsjournals.org.

11. Ganeshan B, Goh V, Mandeville HC, et al. non–small cell lung cancer: Histopathologic Correlates for Texture Parameters at CT. *Radiol n Radiol*. 2013;266(1—January). doi:10.1148/radiol.12112428/-/DC1

12. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: The promise of radiomics. *Phys Medica*. 2017;38:122-139. doi:10.1016/j.ejmp.2017.05.071

13. B O JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. Published online 2017. doi:10.1038/nrclinonc.2016.162

14. Chalkidou A, O ’doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. doi:10.1371/journal.pone.0124165

15. Fournier L, Costaridou L, Bidaut L, et al. Incorporating radiomics into clinical trials: expert consensus on considerations for data-driven compared to biologically driven quantitative biomarkers. *Eur Radiol*. Published online January 25, 2021:1-12. doi:10.1007/s00330-020-07598-8

16. *QIBA CT Volumetry Technical Committee. Lung Nodule Assessment in CT Screening Profile*.; 2017. Accessed August 24, 2020. https://qibawiki.rsna.org/images/f/fb/QIBA_CT_Vol_LungNoduleAssessmentInCTScreening_2017.07.rev15.pdf

17. Zwanenburg A, Leger S, Vallières M, Löck S, Initiative for the IBS. Image biomarker standardisation initiative. Published online 2016. doi:10.17195/candat.2016.08.1

18. Emaminejad N, Kim GH, Brown MS, McNitt-Gray MF, Hoffman J, Wahi-Anwar M. The

effects of variations in parameters and algorithm choices on calculated radiomics feature values: initial investigations and comparisons to feature variability across CT image acquisition conditions. *Med Imaging 2018 Comput Diagnosis*. 2018;(March):140. doi:10.1117/12.2293864

19. McNitt-Gray M, Napel S, Jaggi A, et al. studies. *Tomography*. 2020;6(2):118-128. doi:10.18383/j.tom.2019.00031

20. Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. doi:10.1186/s41747-018-0068-z

21. Haralick RM, Dinstein I, Shanmugam K. Textural Features for Image Classification. *IEEE Trans Syst Man Cybern*. 1973;SMC-3(6):610-621. doi:10.1109/TSMC.1973.4309314

22. Wahi-Anwar M, Emaminejad N, Hoffman J, Kim GH, Brown MS, McNitt-Gray MF. Towards quantitative imaging: stability of fully automated nodule segmentation across varied dose levels and reconstruction parameters in a low-dose CT screening patient cohort. In: Mori K, Petrick N, eds. *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol 10575. SPIE; 2018:56. doi:10.1117/12.2293269

23. Young S, Lo P, Kim G, et al. The effect of radiation dose reduction on computer-aided detection (CAD) performance in a low-dose lung cancer screening population. *Med Phys*. 2017;44(4):1337-1346. doi:10.1002/mp.12128

24. Mackin D, Fave X, Zhang L, et al. Measuring CT scanner variability of radiomics features HHS Public Access. *Invest Radiol*. 2015;50(11):757-765. doi:10.1097/RLI.0000000000000180

25. Yasaka K, Akai H, Mackin D, et al. Precision of quantitative computed tomography texture analysis using image filtering A phantom study for scanner variability.

doi:10.1097/MD.0000000000006993

26. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*. 2018;288(2):407-415. doi:10.1148/radiol.2018172361

27. Shafiq-ul-hassan M, Zhang GG, Hunt DC, et al. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *J Med Imaging*. 2017;5(1):1. doi:10.1117/1.JMI.5.1

28. Kim H, Park CM, Lee M, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: Analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One*. 2016;11(10). doi:10.1371/journal.pone.0164924

29. Gavrielides MA, Kinnard LM, Myers KJ, et al. A resource for the assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom. *Opt Express*. 2010;18(14):15244. doi:10.1364/oe.18.015244

30. Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. *Transl Oncol*. 2014;7:88-93. doi:10.1593/tlo.13865

31. Kim YJ, Lee H-J, Kim KG, Lee SH. The Effect of CT Scan Parameters on the Measurement of CT Radiomic Features: A Lung Nodule Phantom Study. Published online 2019. doi:10.1155/2019/8790694

32. Lo P, Young S, Kim HJ, Brown MS, McNitt-Gray MF. Variability in CT lung-nodule quantification: Effects of dose reduction and reconstruction methods on density and texture based features. *Med Phys*. 2016;43(8):4854. doi:10.1118/1.4954845

33. Hoye J, Richards TW, Solomon JB, Samei E. A method to assess the performance and the relevance of segmentation in radiomic characterization. In: Bosmans H, Chen G-H, eds. *Medical Imaging 2020: Physics of Medical Imaging*. Vol 11312. SPIE-Intl Soc Optical Eng; 2020:95. doi:10.1117/12.2549097

34. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images 1,2. *Transl Oncol*. 2014;7:72-87. doi:10.1593/tlo.13844

35. Foy JJ, Robinson KR, Li H, Giger ML, Al-Hallaq H, Armato SG. Variation in algorithm implementation across radiomics software. *J Med Imag*. 2018;5(4):44505. doi:10.1117/1.JMI.5.4.044505

36. Balagurunathan Y, Kumar V, Gu Y, et al. Test-Retest Reproducibility Analysis of Lung CT Image Features. *J Digit Imaging*. 2014;27(6):805-823. doi:10.1007/s10278-014-9716-x

37. Hunter LA, Krafft S, Stingo F, et al. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. *Med Phys*. 2013;40(12). doi:10.1118/1.4829514

38. Midya A, Chakraborty J, Gönen M, Do RKG, Simpson AL. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. *J Med Imaging*. 2018;5(01):1. doi:10.1117/1.jmi.5.1.011020

39. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT radiomic features within the same patient: Influence of radiation dose and CT reconstruction settings. *Radiology*. 2019;293(3):583-591. doi:10.1148/radiol.2019190928

40. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:1-7. doi:10.1038/srep23428

41. Fave X, Cook M, Frederick A, et al. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Comput Med Imaging Graph*. 2015;44:54-61. doi:10.1016/j.compmedimag.2015.04.006

42. Hoffman J, Emaminejad N, Wahi-Anwar M, et al. Design and Implementation of a High Throughput Pipeline for Reconstruction and Quantitative Analysis of CT Image Data. *Med Phys*. Published online 2019. doi:10.1002/mp.13401

43. Brown MS, Lo P, Goldin JG, et al. Toward clinically usable CAD for lung cancer screening with computed tomography. *Eur Radiol*. Published online 2014:2719-2728. doi:10.1007/s00330-014-3329-0

44. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer*. 45:228-247. doi:10.1016/j.ejca.2008.10.026

45. Hu Q, Whitney HM, Giger ML. Radiomics methodology for breast cancer diagnosis using multiparametric magnetic resonance imaging. *J Med Imaging*. 2020;7(04):1-15. doi:10.1117/1.jmi.7.4.044502

46. Young S, Kim HJG, Ko MM, Ko WW, Flores C, Mcnitt-Gray MF. Variability in CT lung-nodule volumetry: Effects of dose reduction and reconstruction methods. *Med Phys*. 2015;42(5):2679-4095. doi:10.1118/1.4918919

47. Young S, Lo P, Kim G, et al. The Effect of Radiation Dose Reduction on Computer-Aided Detection (CAD) Performance in a Low-Dose Lung Cancer Screening Population. *Med Phys*. Published online January 25, 2017. doi:10.1002/mp.12128

48. Žabić S, Wang Q, Morton T, Brown KM, Žabi S. A low dose simulation tool for CT systems with energy integrating detectors A low dose simulation tool for CT systems with energy

integrating detectors. 2013;031102. doi:10.1118/1.4789628

49. Hoffman J, Young S, Noo F, Mcnitt-Gray M. Technical Note: FreeCT_wFBP: A robust, efficient, open-source implementation of weighted filtered backprojection for helical, fan-beam CT. doi:10.1118/1.4941953

50. Stierstorfer K, Rauscher A, Boese J, Bruder H, Schaller S, Flohr T. Weighted FBP - A simple approximated 3D FBP algorithm for multislice spiral CT with good dose usage for arbitrary pitch. *Phys Med Biol*. 2004;49(11):2209-2218. doi:10.1088/0031-9155/49/11/007

51. Hoffman J, UCLA CT Physics and Reconstruction Group. Free_CT Homepage. http://cvib.ucla.edu/freect

52. Boedeker KL, Cooper VN, McNitt-Gray MF. Application of the noise power spectrum in modern diagnostic MDCT: Part I. Measurement of noise power spectra and noise equivalent quanta. *Phys Med Biol*. 2007;52(14):4027-4046. doi:10.1088/0031-9155/52/14/002

53. Hoffman J, Young S, Noo F, McNitt-Gray M. Technical Note: FreeCT_wFBP: A robust, efficient, open-source implementation of weighted filtered backprojection for helical, fan-beam CT. *Med Phys*. 2016;43(3):1411-1420. doi:10.1118/1.4941953

54. American College of Radiology. *Lung- RADS*.; 2019. Accessed May 18, 2021. https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/LungRADSAssessmentCategoriesv1-1.pdf

55. HTCondor. https://research.cs.wisc.edu/htcondor

56. Wu Z, Lan T, Wang J, Ding Y, Qin Z. Medical Image Registration Using B-Spline Transform. doi:10.5013/IJSSST.a.17.48.01

57. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107.

doi:10.1158/0008-5472.CAN-17-0339

58. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*. 2002;58(4):1020-1027. doi:10.1111/j.0006-341X.2002.01020.x

59. Lin LI-K. *A Concordance Correlation Coefficient to Evaluate Reproducibility*. Vol 45.; 1989.

60. McBride G. A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. *NIWA Client Rep*. 2005;45(1):307-310. doi:10.2307/2532051

61. Yang J, Zhang L, Fave XJ, et al. Uncertainty analysis of quantitative imaging features extracted from contrast-enhanced CT in lung tumors. *Comput Med Imaging Graph*. 2016;48:1-8. doi:10.1016/j.compmedimag.2015.12.001

62. Lecler A, Duron L, Balvay D, et al. Combining Multiple Magnetic Resonance Imaging Sequences Provides Independent Reproducible Radiomics Features. *Sci Rep*. 2019;9(1):1-8. doi:10.1038/s41598-018-37984-8

63. Hoffman JM. *Dissertation: Characterizing and Minimizing the Impacts of Diagnostic Computed Tomography Acquisition and Reconstruction Parameter Selection on Quantitative Emphysema Scoring*.; 2018.

64. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*. Published online 2018:172361. doi:10.1148/radiol.2018172361

65. Kim H, Park CM, Lee M, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: Analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One*. 2016;11(10). doi:10.1371/journal.pone.0164924

66. MacKin D, Ger R, Dodge C, et al. Effect of tube current on computed tomography radiomic features. *Sci Rep*. 2018;8(1):1-10. doi:10.1038/s41598-018-20713-6

67. D.R. Aberle, A.M. Adams, C.D. Berg, W.C. Black, J.D. Clapp et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med*. 2011;365:395-409. doi:10.1056/NEJMoa1414264

68. Ahn SY, Park CM, Park SJ, et al. Prognostic Value of Computed Tomography Texture Features in Non–Small Cell Lung Cancers Treated With Definitive Concomitant Chemoradiotherapy. *Invest Radiol*. 2015;50(10):719-725. doi:10.1097/RLI.0000000000000174

69. Weiss GJ, Ganeshan B, Miles KA, et al. Noninvasive image texture analysis differentiates K-ras mutation from pan-wildtype NSCLC and is prognostic. *PLoS One*. 2014;9(7). doi:10.1371/journal.pone.0100244

70. Whitney HM, Li H, Ji Y, Liu P, Giger ML. Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. *J Med Imaging*. 2020;7(01):1. doi:10.1117/1.jmi.7.1.012707

71. Gang G, Stayman J. Modeling and Recovering Gray-Level Co-Occurrence-Based Radiomics in the Presence of Blur and Noise. In: *AAPM*. ; 2020. https://w3.aapm.org/meetings/2020AM/programInfo/programAbs.php?t=specific&shid[]=1575&sid=8801&aid=53374

72. Foy J, Mitta P, Nowosatka LR, et al. Variations in algorithm implementation among quantitative texture analysis software packages. In: Mori K, Petrick N, eds. *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol 10575. SPIE; 2018:55. doi:10.1117/12.2292573

73. Li Y, Lu L, Xiao M, et al. CT Slice Thickness and Convolution Kernel Affect Performance of a Radiomic Model for Predicting EGFR Status in Non-Small Cell Lung Cancer: A Preliminary Study. *Sci Rep*. 2018;8(1):1-10. doi:10.1038/s41598-018-36421-0

74. Kim H, Park CM, Gwak J, et al. Effect of CT Reconstruction Algorithm on the Diagnostic Performance of Radiomics Models: A Task-Based Approach for Pulmonary Subsolid Nodules. *Am J Roentgenol*. 2019;212(3):505-512. doi:10.2214/AJR.18.20018

75. Fletcher JG, Levin DL, Sykes A-MG, et al. Observer Performance for Detection of Pulmonary Nodules at Chest CT over a Large Range of Radiation Dose Levels. *Radiology*. 2020;(13):200969. doi:10.1148/radiol.2020200969

76. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS One*. 2017;12(9). doi:10.1371/journal.pone.0178524

77. Kim H, Park CM, Lee M, et al. Impact of Reconstruction Algorithms on CT Radiomic Features of Pulmonary Tumors: Analysis of Intra- and Inter-Reader Variability and Inter-Reconstruction Algorithm Variability. Published online 2016. doi:10.1371/journal.pone.0164924

78. Emaminejad N, Wahi- Anwar MW, Kim GHJ, Hsu W, Brown M, McNitt- Gray M. Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters. *Med Phys*. Published online April 13, 2021. doi:10.1002/mp.14830

79. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050-1062. doi:10.1002/mp.12123

80. Caramella C, Allorant A, Orlhac F, et al. Can we trust the calculation of texture indices of

CT images? A phantom study. *Med Phys*. 2018;45(4):1529-1536. doi:10.1002/mp.12809

81. Foy JJ, Al-Hallaq HA, Grekoski V, et al. Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: Assessment in a cadaveric liver. *Phys Med Biol*. 2020;65(20):205008. doi:10.1088/1361-6560/abb172

82. Shafiq M, Zhang G, Latifi K, Ullah G, Gillies R, Moros EG. Computed Tomography Texture Phantom Dataset for Evaluating the Impact of CT Imaging Parameters on Radiomic Features . Abstract : :1-9.

83. Salon E, Abidin Sulaiman Z, Ahmad A, et al. *Determination of Effective Atomic Number of Rubber*. Vol 6.; 1983.

84. Chang K-P, Hung S-H, Chie Y-H, Shiau A-C, Huang R-J. A Comparison of physical and dosimetric properties of lung substitute materials. *Med Phys*. 2012;39(4):2013-2020. doi:10.1118/1.3694097

85. Larue RTHM, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol (Madr)*. 2017;56(11):1544-1553. doi:10.1080/0284186X.2017.1351624

86. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. Published online 2016. doi:10.1038/srep34921

87. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings. Published online 2016. doi:10.1371/journal.pone.0166550

88. Petrou M, Quint LE, Nan B, Baker LH. Pulmonary Nodule Volumetric Measurement Variability as a Function of CT Slice Thickness and Nodule Morphology I. Published online 2007. doi:10.2214/AJR.05.1063

89. Zhao B, Schwartz LH, Moskowitz CS, et al. Pulmonary Metastases: Effect of CT Section Thickness on Measurement-Initial Experience 1. Published online 2005. doi:10.1148/radiol.2343040020

90. Winer-Muram HT, Jennings SG, Meyer CA, et al. From the Departments of Radiology Effect of Varying CT Section Width on Volumetric Measurement of Lung Tumors and Application of Compensatory Equations 1 Thoracic Imaging. *Radiology*. 2003;229:184-194. doi:10.1148/radiol.2291020859

91. Way TW, Chan H-P, Sahiner B, et al. Effect of CT scanning parameters on volumetric measurements of pulmonary nodules by 3D active contour segmentation: A phantom study. doi:10.1088/0031-9155/53/5/009

92. Robins M, Solomon J, Hoye J, Abadi E, Marin D, Samei E. Systematic analysis of bias and variability of texture measurements in computed tomography. *J Med Imaging*. 2019;6(03):1. doi:10.1117/1.jmi.6.3.033503

93. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30:1234-1248. doi:10.1016/j.mri.2012.06.010

94. Ger RB, Zhou S, Chi P-CM, et al. Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. *Sci RepoRTS* |. 2018;8:13047. doi:10.1038/s41598-018-31509-z

95. Ger RB, Craft DF, Mackin DS, et al. Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis. *Comput Med Imaging*

Graph. 2018;69:134-139. doi:10.1016/j.compmedimag.2018.09.002

96. Chatterjee A, Vallieres M, Dohan A, et al. Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization. *IEEE Trans Radiat Plasma Med Sci*. 2019;3(2):210-215. doi:10.1109/trpms.2019.2893860

97. Dissaux G, Visvikis D, Da-Ano R, et al. Pretreatment 18F-FDG PET/CT Radiomics Predict Local Recurrence in Patients Treated with Stereotactic Body Radiotherapy for Early-Stage Non-Small Cell Lung Cancer: A Multicentric Study. *J Nucl Med*. 2020;61(6):814-820. doi:10.2967/jnumed.119.228106

98. Whitney HM, Giger ML. Improvement of classification performance using harmonization across field strength of radiomic features extracted from DCE-MR images of the breast. 2020;(March 2020):33. doi:10.1117/12.2548129

99. Peeken JC, Shouman MA, Kroenke M, et al. A CT-based radiomics model to detect prostate cancer lymph node metastases in PSMA radioguided surgery patients. *Eur J Nucl Med Mol Imaging*. 2020;47(13):2968-2977. doi:10.1007/s00259-020-04864-1

100. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. *Generative Adversarial Nets*. Accessed March 17, 2021. http://www.github.com/goodfeli/adversarial

101. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037

102. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep*. 2018;8(1):1-9. doi:10.1038/s41598-018-28895-9

103. Hoffman J. Characterizing and Minimizing the Impacts of Diagnostic Computed

Tomography Acquisition and Reconstruction Parameter Selection on Quantitative Emphysema Scoring. Published online 2018:1-193.

104. Loizou CP, Pantziaris M, Seimenis I, Pattichis CS. Brain MR image normalization in texture analysis of multiple sclerosis. In: *Final Program and Abstract Book - 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009*. ; 2009. doi:10.1109/ITAB.2009.5394331

105. Gang GJ, Deshpande R, Stayman JW. Standardization of histogram- and gray-level co-occurrence matrices-based radiomics in the presence of blur and noise. *Phys Med Biol*. 2021;66:74004. doi:10.1088/1361-6560/abeea5

106. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: Bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol*. 2019;19:33-38. doi:10.1016/j.ctro.2019.07.003

107. Gammex 467 Tissue Characterization Phantom. http://cspmedical.com/content/102-%0A1492_tissue_phantom_user_guide.pdf

108. Pomponio R, Erus G, Habes M, et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan HHS Public Access. *Neuroimage*. 2020;208:116450. doi:10.1016/j.neuroimage.2019.116450

109. Orlhac F, Boughdad S, Philippe C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med*. 2018;59(8):1321-1328. doi:10.2967/jnumed.117.199935

110. Mahon RN, Ghita M, Hugo GD, Weiss E. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Phys Med Biol*. 2020;65(1):015010. doi:10.1088/1361-6560/ab6177

111. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative Adversarial Networks: An Overview. *IEEE Signal Process Mag*. 2018;35(1):53-65. doi:10.1109/MSP.2017.2765202

112. generative adversarial text to image synthesis. Accessed April 1, 2021. https://www.google.com/search?q=generative+adversarial+text+to+image+synthesis&oq=Generative+adversarial+text+to+image+synthesis&aqs=chrome.0.0l3j0i22i30l7.1220j0j4&sourceid=chrome&ie=UTF-8

113. A convolutional generative adversarial network for symbolic-domain music generation using 1d and 2d conditions. Accessed April 1, 2021. https://www.google.com/search?q=A+convolutional+generative+adversarial+network+for+symbolic-domain+music+generation+using+1d+and+2d+conditions&safe=active&sxsrf=ALeKk03pa_3WpaOTwF1XvyEyu3nvF5_Q9w%3A1617300745545&ei=CQ1mYK3UIPnF0PEPxr27-A0&oq=A+convolution

114. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context Encoders: Feature Learning by Inpainting. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 2016-December. IEEE Computer Society; 2016:2536-2544. doi:10.1109/CVPR.2016.278

115. Ledig C, Theis L, Huszár F, et al. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*.

116. Kingma DP, Welling M. *Auto-Encoding Variational Bayes*.

117. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks.

Sci Rep. 2019;9(1):1-9. doi:10.1038/s41598-019-52737-x

118. Han Z, Wei B, Mercado A, Leung S, Li S. Spine-GAN: Semantic segmentation of multiple spinal structures. *Med Image Anal*. 2018;50:23-35. doi:10.1016/j.media.2018.08.005

119. Dong X, Lei Y, Wang T, et al. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Med Phys*. 2019;46(5):2157-2168. doi:10.1002/mp.13458

120. Jin C-B, Kim H, Jung W, et al. *Deep CT to MR Synthesis Using Paired and Unpaired Data*.; 2018.

121. Ying X, Guo H, Ma K, et al. *X2CT-GAN: Reconstructing CT from Biplanar X-Rays with Generative Adversarial Networks.*

122. Qadir SA, Zarshenas A, Yang L, Liu J, Fajardo L, Suzuki K. Radiation dose reduction in digital breast tomosynthesis (DBT) by means of neural network convolution (NNC) deep learning. In: Krupinski EA, ed. *14th International Workshop on Breast Imaging (IWBI 2018)*. Vol 10718. SPIE; 2018:15. doi:10.1117/12.2317789

123. Chen H, Zhang Y, Kalra MK, et al. Low-Dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*. 2017;36(12):2524-2535. doi:10.1109/TMI.2017.2715284

124. Shan H, Zhang Y, Yang Q, et al. 3-D Convolutional Encoder-Decoder Network for Low-Dose CT via Transfer Learning From a 2-D Trained Network. *IEEE Trans Med Imaging*. 2018;37(6):1522-1534. doi:10.1109/TMI.2018.2832217

125. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. *IEEE Trans Med IMAGING,*. 2017;36(12):2536-2545. Accessed April 1, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7934380

126. CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE). Accessed April 1, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8736838

127. Zhu J-Y, Park T, Isola P, Efros AA, Research BA. *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks Monet Photos*.

128. Ouyang J, Chen KT, Gong E, Pauly J, Zaharchuk G. Ultra-low-dose PET reconstruction using generative adversarial network with feature matching and task-specific perceptual loss. *Med Phys*. 2019;46(8):3555-3564. doi:10.1002/mp.13626

129. Wei L, Lin Y, Hsu W. *USING A GENERATIVE ADVERSARIAL NETWORK FOR CT NORMALIZATION AND ITS IMPACT ON RADIOMIC FEATURES*.; 2020.

130. Liang G, Fouladvand S, Zhang J, Brooks MA, Jacobs N, Chen J. GANai: Standardizing CT Images using Generative Adversarial Network with Alternative Improvement. *2019 IEEE Int Conf Healthc Informatics, ICHI 2019*. Published online 2019:1-11. doi:10.1109/ICHI.2019.8904763

131. Armanious K, Jiang C, Fischer M, et al. *MedGAN: Medical Image Translation Using GANs*.

132. Isola P, Zhu J-Y, Zhou T, Efros AA, Research BA. *Image-to-Image Translation with Conditional Adversarial Networks*. Accessed March 31, 2020. https://github.com/phillipi/pix2pix.

133. Mirza M, Osindero S. *Conditional Generative Adversarial Nets*.

134. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Vol 15.; 2014.

135. Efros AA, Freeman WT. *Image Quilting for Texture Synthesis and Transfer*.

136. Efros AA, Leung TK. *Texture Synthesis by Non-Parametric Sampling*.; 1999.

137. Cirillo MD, Abramian D, Eklund A. *Vox2Vox: 3D-GAN for Brain Tumour Segmentation*.

138. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition*. Accessed March 18, 2021. http://image-net.org/challenges/LSVRC/2015/

139. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612. doi:10.1109/TIP.2003.819861

140. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR; 2015.

141. TensorFlow White Papers. Accessed March 19, 2021. https://www.tensorflow.org/about/bib

142. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. *Improved Techniques for Training GANs*. Accessed March 23, 2021. https://github.com/openai/

143. Arjovsky M, Chintala S, Bottou L. *Wasserstein Generative Adversarial Networks*. PMLR; 2017. Accessed March 24, 2021. http://proceedings.mlr.press/v70/arjovsky17a.html

144. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*.

145. Johnson J, Alahi A, Fei-Fei L. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*.

146. Gatys L, Ecker A, Bethge M. A Neural Algorithm of Artistic Style. *J Vis*. 2016;16(12):326. doi:10.1167/16.12.326

147. Schreiber S, Geldenhuys J, De Villiers H. Texture synthesis using convolutional neural networks with long-range consistency and spectral constraints. In: *2016 Pattern*

Recognition Association of South Africa and Robotics and Mechatronics International Conference, PRASA-RobMech 2016. Institute of Electrical and Electronics Engineers Inc.; 2017. doi:10.1109/RoboMech.2016.7813173

148. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Published online 2014:1-14. doi:10.1016/j.infsof.2008.09.005

149. Lin L, Hedayat AS, Sinha B, Yang M. *Statistical Methods in Assessing Agreement: Models, Issues, and Tools*.; 2002. Accessed March 27, 2020. http://www.uic.edu/

150. Tomtom AK, Wang O, Research A. *MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks*.

151. Miyato T, Kataoka T, Koyama M, Yoshida Y. *SPECTRAL NORMALIZATION FOR GENERATIVE ADVERSARIAL NETWORKS*. Accessed March 31, 2021. https://github.com/pfnet-research/sngan_

152. Hunter LA, Krafft S, Stingo F, et al. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. *Med Phys*. 2013;40(12). doi:10.1118/1.4829514

153. Bu- by, George Casella -m. *AN INTRODUCTION TO EMPIRICAL BAYES DATA ANALYSIS*.; 1982.

154. Goh WW Bin, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol*. 2017;35(6):498-507. doi:10.1016/j.tibtech.2017.02.012

155. Chen C, Grennan K, Badner J, et al. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. Kliebenstein D, ed. *PLoS One*. 2011;6(2):e17238. doi:10.1371/journal.pone.0017238

156. Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements

across scanners and sites. *Neuroimage*. 2018;167:104-120. doi:10.1016/j.neuroimage.2017.11.024

157. Fortin JP, Parker D, Tunç B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*. 2017;161:149-170. doi:10.1016/j.neuroimage.2017.08.047

158. Yu M, Linn KA, Cook PA, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi- site fMRI data. *Hum Brain Mapp*. 2018;39(11):4213-4227. doi:10.1002/hbm.24241

159. Orlhac F, Lecler A, Savatovski J, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol*. Published online September 25, 2020:1-9. doi:10.1007/s00330-020-07284-9

160. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology*. 2019;291(1). doi:10.1148/radiol.2019182023

161. Stein CK, Qu P, Epstein J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics*. 2015;16(1):63. doi:10.1186/s12859-015-0478-3

162. Bowman KO, Shenton LR. Estimator: Method of Moments. In: *Encyclopedia of Statistical Sciences*. Wiley; 1998:2092-2098.

163. Team R core. R: A Language and Environment for Statistical Computing. Published online 2017. https://redirect.viglink.com/?format=go&jsonp=vglnk_161768534420013&key=949efb41171ac6ec1bf7f206d57e90b8&libId=kn5h1ovv01021u9s000DLc05ydsrb&loc=https%253A%252F%252Fwww.r-bloggers.com%252F2018%252F06%252Fits-easy-to-cite-and-

reference-r%252F&v=1&type=U&out=https%253A%252F%25

164. Team Rs. RStudio: Integrated Development for R. Published online 2020.

165. Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. *Phys Med Biol*. 2020;65(24). doi:10.1088/1361-6560/aba798

166. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 2007;3(9):1724-1735. doi:10.1371/journal.pgen.0030161

167. Nie D, Trullo R, Petitjean C, Ruan S, Shen D. *Medical Image Synthesis with Context-Aware Generative Adversarial Networks*.

168. Ning Z, Luo J, Li Y, et al. Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolutional features. *IEEE J Biomed Heal Informatics*. 2019;23(3):1181-1191. doi:10.1109/JBHI.2018.2841992

169. Kalpathy-Cramer J, Mamomov A, Zhao B, et al. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomography*. 2016;2(4):430-437. doi:10.18383/j.tom.2016.00235