Title
Measurement Variation and Robustness in Quantitative Thoracic Computed Tomography

Author
Chong, Daniel Y.

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Measurement Variation and Robustness in Quantitative Thoracic Computed Tomography

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biomedical Physics

by

Daniel Yun Chong

2015

ABSTRACT OF THE DISSERTATION

Measurement Variation and Robustness in Quantitative Thoracic Computed Tomography

by

Daniel Yun Chong

Doctor of Philosophy in Biomedical Physics

University of California, Los Angeles, 2015

Professor Matthew Sherman Brown, Chair

Quantitative computed tomography (CT) plays a major role in large-scale, longitudinal, multi-center clinical trials. Minimizing measurement variation by identifying robust CT imaging biomarkers and developing robust techniques for quantitative CT has implications for clinical trial management and for patient care. We investigated robustness with respect to two sources of measurement variation in quantitative CT: repeat-scan variation (reproducibility) and variation due to changing CT technical parameters.

In this dissertation, we conducted two separate but related studies in the area of quantitative CT robustness. In the first, we characterized and compared the reproducibilities of several widely-accepted measures of emphysema by examining repeat CT images from a multi-center clinical trial taken one week apart. We investigated the influence of breathhold on reproducibility of

emphysema measures. We also investigated variations in reproducibility characteristics across sites. Our results have implications for multi-center clinical trials that rely on accurate and reproducible measures of emphysema.

In the second study, we investigated feature and classifier robustness with respect to slice thickness, reconstruction kernel, and tube current in the setting of classification of fibrotic interstitial lung disease (FILD). We developed a quantitative Robustness Index measure by examining the stability of imaging features across multiple systematic reconstructions of CT raw sinogram data. We proposed a novel Robustness-Driven Feature Selection (RDFS) method for identifying a subset of robust features, then used these features to develop a robust support vector classifier for lung structure and parenchymal abnormalities in FILD. We demonstrated the superior robustness of this classifier compared to a similar classifier that did not utilize RDFS. Our results have implications for improving the robustness of classifier-based CT CAD systems, which is of importance in multi-center clinical trials that rely on imaging biomarkers that can be generalized across many sites and timepoints.

The dissertation of Daniel Yun Chong is approved.

David A. Lynch

Hyun J. Kim

Michael F. McNitt-Gray

Jonathan G. Goldin

Matthew Sherman Brown, Committee Chair

University of California, Los Angeles

2015

To God, who created a fascinating universe and endowed us with the capacity to understand it, if only a little;

My friends from Yukai Daiko, who were my life outside of lab;

My uncle Chris, who showed me the way;

My grandparents, that my life may be worthy of their sacrifices;

Jessica, for her steadfast love and support;

And my parents Peter and Joyce, who above all taught me never to be satisfied.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor Professor Matthew Brown, whose patience, support, and guidance carried me through the long journey of my PhD career. I would also like to express my gratitude for my committee members, Professor Jonathan Goldin, Professor Michael McNitt-Gray, Professor Hyun J. Kim, and Professor David Lynch. This dissertation is founded upon the intersection of a diverse set of disciplines, and I am privileged to benefit from the insight, experience, and wisdom of expert authorities in each of these fields.

Chapter two of this dissertation is adapted from the final submitted manuscript of a paper that was published as Chong D, Brown MS, Kim HJ, van Rikxoort EM, Guzman L, McNitt-Gray MF, Khatonabadi M, Galperin-Aizenberg M, Coy H, Yang K, Jung Y, Goldin JG. Reproducibility of volume and densitometric measures of emphysema on repeat computed tomography with an interval of 1 week. Eur Radiol 2012; 22(2):287-94.

Chapter five of this dissertation is adapted from the manuscript "Robustness-driven feature selection in classification of fibrotic interstitial lung disease patterns in computed tomography using 3D texture features" by D Chong, HJ Kim, P Lo, S Young, MF McNitt-Gray, F Abtin, JG Goldin, MS Brown, which has been submitted to IEEE Transactions on Medical Imaging for consideration for publication.

I am grateful to each of my coauthors, without whom this research would not have been possible. Pechin Lo constantly supported me with advice, suggestions, and feedback for developing methodology, interpreting results, and writing manuscripts. Stefano Young assisted with collection of CT raw data and generously shared his tools for simulated dose reduction. Eva van Rikxoort provided numerous suggestions for reproducibility methods. Peiyun Lu worked closely with me to perform statistical analyses and help interpret results. Maryam Khatonabadi

Bostani helped ensure quality control of CT imaging devices across sites and assisted with collection of CT raw data. Mazyar Aryanfar, Heidi Coy, Yongha Jung, Maya Khatonabadi, David Oria, and Katherine Yang spent countless hours contouring and editing ROIs; truly, lab technologists are the backbone of medical imaging research. Laura Guzman served as project manager and coordinated data collection and quality control efforts on a massive scale. Maya Galperin-Aizenberg provided valuable medical insight for pulmonary emphysema. Fereidoun Abtin generously donated his time to provide thousands of annotations for ground truth. Hyun J. Kim served as a mentor throughout my PhD career, providing valuable guidance with study design, patiently answering my questions about statistical methods, and offering much-appreciated career and life advice. Michael McNitt-Gray taught me everything I know about CT physics, and his insight into the role of physics in quantitative imaging is second to none. Jonathan Goldin generously donated his time to provide thousands of annotations for ground truth, and his tireless commitment to basic science was a constant inspiration. Matthew Brown worked closely with me, advising me on all aspects of research, reading and rereading every manuscript I've ever written, and inspiring me to never stop asking questions and seeking answers.

I would also like to personally thank the following individuals: my fellow computer team members Pechin Lo, Greg Chu, Bharath Ramakrishna, Pang Yu Teng, Mahesh Nagarajan, for their encouragement; Jing Huo, who got there first; Erin Angel Sisto, who was an early inspiration; and last but not least, everyone who helped shape my life during my time at UCLA. There is not room enough on these pages to adequately express the depth of my gratitude and appreciation for each of you.

VITA

2000	Graduated valedictorian Wichita High School East, Wichita, KS
2000-2005	Camras Scholar Illinois Institute of Technology, Chicago, IL
2003-2005	Academic Tutor, Mathematics Academic Resource Center Illinois Institute of Technology, Chicago, IL
2003-2005	Undergraduate Student Researcher Department of Computer Science Illinois Institute of Technology, Chicago, IL
2004	Research in Industrial Projects for Students Institute for Pure and Applied Mathematics University of California Los Angeles, Los Angeles, CA
2005	Clinton E. Stryker Distinguished Service Award Illinois Institute of Technology, Chicago, IL
2005	John L. Way Award for Excellence in Tutoring Academic Resource Center Illinois Institute of Technology, Chicago, IL
2005	B.S. Applied Mathematics B.S. Computer Science Minor, Physics Illinois Institute of Technology, Chicago, IL
2005-Present	Graduate Student Researcher Center for Computer Vision and Imaging Biomarkers Department of Radiology David Geffen School of Medicine at UCLA, Los Angeles, CA

PUBLICATIONS AND PRESENTATIONS

Kim HJ, Brown MS, Chong D, Gjertson DW, Lu P, Kim HJ, Coy H, Goldin JG. Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months. Acad Radiol 2015. 22(1):70-80.

Chong D, Kim HJ, Goldin JG, Abtin F, Brown MS. Computer-aided classification of interstitial lung diseases in high-resolution computed tomography using 3D multiscale texture features. Poster presentation at the American Thoracic Society International Conference 2013.

Chong D, Lu P, Kim HJ, Lo P, Da Costa I, Yang K, Aryanfar M, Oria D, Jung Y, Khatonabadi M, Brown MS, Goldin JG. Reproducibility of breathhold and densitometric measures of emphysema in computed tomography scans with a one-week interval. Poster presentation at the American Thoracic Society International Conference 2013.

Chong D, Lu P, Kim HJ, Da Costa I, Yang K, Aryanfar M, Oria D, Jung Y, Khatonabadi M, Brown MS, Goldin JG. Quantitative reproducibility of densitometric measures of emphysema in a multicenter clinical trial: The influence of breathhold, site, and scanner manufacturer. Poster presentation at the American Thoracic Society International Conference 2013.

Chong D, Lu P, Kim HJ, Brown MS, Goldin JG. Reproducibility of in vivo measurements of attenuation values in computed tomography scans with a one-week interval. Poster presentation at the American Thoracic Society International Conference 2013.

Brown MS, Kim HJ, Abtin FG, Strange C, Galperin-Aizenberg M, Pais R, Da Costa IG, Ordookhani A, Chong D, Ni C, McNitt-Gray MF, Tashkin DP, Goldin JG. Emphysema lung lobe volume reduction: effects on the ipsilateral and contralateral lobes. Eur Radiol. 2012 22(7):1547-55. Epub 2012 Apr 1.

Chong D, Brown MS, Kim HJ, van Rikxoort EM, Guzman L, McNitt-Gray MF, Khatonabadi M, Galperin-Aizenberg M, Coy H, Yang K, Jung Y, Goldin JG. Reproducibility of volume and densitometric measures of emphysema on repeat computed tomography with an interval of 1 week. Eur Radiol. 2012 22(2):287-94.

Kim HJ, Brown MS, Elashoff R, Li G, Gjertson DW, Lynch DA, Strollo DC, Kleerup E, Chong D, Shah SK, Ahmad S, Abtin F, Tashkin DP, Goldin JG. Quantitative texture-based assessment of one-year changes in fibrotic reticular patterns on HRCT in scleroderma lung disease treated with oral cyclophosphamide. Eur Radiol. 2011 21(12):2455-65.

Chong D, Kim HJ, van Rikxoort EM, Galperin M, Yang K, Jung Y, McNitt-Gray MF, Goldin JG, Brown MS. Reproducibility of densitometric measures of emphysema in computed tomography scans one week apart: the effect of breathhold and scanner calibration. Poster presentation at the American Thoracic Society International Conference 2011.

Chong D, van Rikxoort EM, Kim HJ, Goldin JG, Brown MS. Scan-rescan reproducibility of CT densitometric measures of emphysema. Poster presentation at SPIE Medical Imaging 2011.

Chong D, Brown MS, Ordookhani A, Kim HJ, Ochs R, Angel E, McNitt-Gray MF, Goldin JG. Investigation of influence of technical factors on quantitative CT emphysema scoring. Poster presentation at the American Thoracic Society International Conference 2009.

Chong D, Brown MS, Ochs R, Abtin F, Brown M, Ordookhani A, Shaw G, Kim HJ, Gjertson D, Goldin JG. The effect of CT technical factors on quantification of lung fissure integrity. Presented at the SPIE Medical Imaging 2009.

Chong D, Angel E, Kim HJ, Cole GB, Boyadzhyan L, Panknin C, Gomez AM, Goldin JG, Brown MS, McNitt-Gray MF. Separation of bone from iodine- and gadolinium-based contrast agents using dual energy CT. Poster presentation at SPIE Medical Imaging 2008.

1. Introduction

This dissertation is an investigation into sources of variation in quantitative computed tomography (CT) and methods to improve the robustness of computer-aided techniques in quantitative CT. We focus on robustness with respect to two particular sources of variation: variation due to repeated image acquisition (reproducibility) and variation due to changing CT technical parameters, namely slice thickness, reconstruction kernel, and tube current. The purpose of this dissertation is twofold: first, to quantitatively characterize and compare reproducibility in commonly-used CT densitometric measures of pulmonary emphysema; and second, to develop a methodology for robust feature selection that produces a classifier model which is robust to changes in CT slice thickness, reconstruction kernel, and tube current in the setting of classification for fibrotic interstitial lung disease.

Since its inception, computed tomography has played an increasing role in the diagnosis and treatment of a variety of diseases, including lung diseases such as pulmonary emphysema and fibrotic interstitial lung disease. Today, quantitative CT (QCT) is firmly established as a valuable diagnostic and prognostic tool, and CT-derived imaging biomarkers play a major role in longitudinal multi-center clinical trials as a primary or secondary endpoint. It is therefore critical to understand the measurement variations associated with CT imaging biomarkers, particularly their reproducibility and their stability with regards to CT technical parameters, which can be extremely challenging to standardize across a large-scale study. Minimizing measurement variation by identifying robust CT imaging biomarkers and developing robust techniques for quantitative CT has implications for clinical trial management and for patient care.

1.1 Introduction to Computed Tomography

Computed tomography (CT) is a medical imaging modality with high contrast and high spatial resolution. Initially popularized as a noninvasive method for radiologists to visualize patient anatomy and pathology, CT has been increasingly recognized as a powerful quantitative tool as well. Advances in technology have steadily improved both scanning speed and image quality, improving the diagnostic utility of CT and paving the way for new applications.

The basic principles of tomography were outlined by Johann Radon, who in 1917 developed the mathematical techniques for recreating an image from a series of projections through the image [1]. Given the projection data of an image (also called the image sinogram), the original image can be obtained via the inverse Radon transform, which is commonly implemented in medical CT via filtered backprojection. This process of recovering an image from its sinogram is known as reconstruction.

As its name implies, filtered backprojection involves the use of a filter (variously referred to as "convolution kernel", "reconstruction kernel", or "reconstruction algorithm"). The purpose of this filter is to correct image artifacts introduced by the backprojection method. The choice of kernel greatly affects the reconstructed image, influencing image quality characteristics such as spatial resolution, contrast resolution, and noise power spectrum. In practice, CT device manufacturers have implemented numerous proprietary reconstruction kernels for various anatomies and imaging tasks. However, these kernels are optimized for visual assessment, and the influence of kernel selection on quantitative image analysis is not well characterized. An illustration of several kernels for thoracic CT is shown in Fig. 1.1.

In x-ray computed tomography, projections of a patient's anatomy are taken via x-ray radiography. The intensity at any given image voxel therefore reflects the x-ray attenuation

properties at the corresponding anatomical location. Image intensity values in CT are measured on a graylevel scale known as the Hounsfield Unit (HU) scale, named for Sir Godfrey Hounsfield, the inventor of CT [1]. The Hounsfield Unit is a relative scale which is defined as follows:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water}} \tag{1.1}$$

where μ represents the x-ray linear attenuation coefficient for the voxel in question and μ_{water} represents the linear attenuation coefficient of water. Therefore, water corresponds to a CT value of 0 HU, while air corresponds to -1000 HU (since air is assumed to be radiotransparent, i.e. $\mu_{air} = 0$). By convention, low HU values are rendered as dark while high HU values are bright; actual intensity on a computer screen depends on the choice of window and level, which compresses the wide dynamic range of CT to a range that can be displayed on a digital screen by truncating HU values above and below the chosen limits.

Signal strength in CT is influenced by the number of x-ray photons captured by the detector relative to the number of photons given off by the x-ray source. This is essentially a counting problem for a random variable which is governed by the Poisson distribution. It is known that signal-to-noise (SNR) ratio for this distribution is proportional to the square root of N, the total number of photons emitted by the x-ray source. Therefore, as the number of x-ray photons increases, the CT signal strength increases, resulting in a corresponding decrease in image noise, which can be estimated by computing the standard deviation of HU values in a relatively homogeneous region of interest. The number of x-ray photons is controlled by the x-ray tube current-time product, which is commonly measured in units of milliampere-seconds (mAs). To summarize all these concepts, if the tube current is halved, then the image noise will increase roughly by a factor of $\sqrt{2} \approx 1.41$, resulting in an noisier image. This result is illustrated in Fig. 1.1.

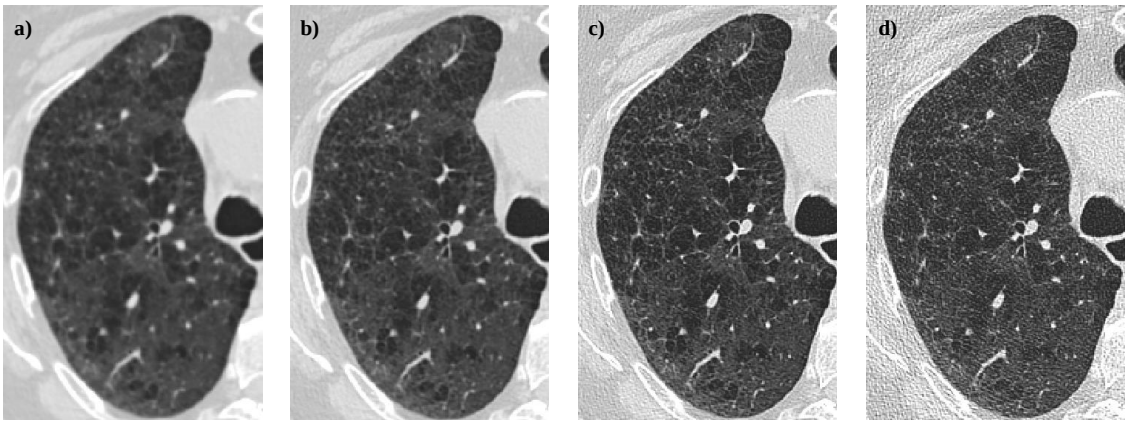

Figure 1.1. Image of lung reconstructed under various CT technical factors. (a) Smooth kernel, high tube current; (b) Medium-sharp kernel, high tube current; (c) Sharp kernel, high tube current; (d) Sharp kernel, simulated low tube current.

Additionally, since tube current is directly proportional to patient radiation dose, the CT community has adopted the ALARA principle to keep tube currents "as low as reasonably achievable" while maintaining the image quality necessary for a particular diagnostic task [2].

Computed tomography produces an image volume which is conventionally presented as a stack of 2D slabs or "slices", each of which represents an axial cross-section of the patient's body. The slices are of a finite thickness which may be specified at the time of image reconstruction, subject to restrictions relating to image acquisition geometry. The slices may be contiguous, overlapping, or separated. In the past, CT devices typically produced so-called "thick-section" slices of 5-10 mm, significantly larger than in-plane image resolution which is commonly below 1mm. Advances in technology have steadily improved CT devices to the point where isotropic or near-isotropic images, which are often desirable for quantitation, are more commonly achievable. However, thinner slices also result in increased image noise since fewer photons are collected for each slice.

1.2 Introduction to CT Image Classification

Classification refers to the problem of assigning a category (or class label) to a new observation based on patterns learned from a training set of previous observations with known categories. When the observations take the form of images, this problem falls under the field of computer vision, a discipline that concerns itself with analyzing, understanding, and perceiving digital images. CT image classification typically involves diagnostic categories such as healthy versus abnormal lung tissue or benign versus cancerous lesions.

Image classification can be broken down into a series of smaller tasks which make up an image processing pipeline. A pipeline for a typical classification problem might be comprised of the following steps: image acquisition, image preprocessing, feature extraction, feature selection, model training, and classification. We will briefly discuss each of these in turn in the context of CT image classification.

Image acquisition refers to the process of creating the image that is to be analyzed. In CT, this step encompasses both CT acquisition (the act of acquiring projection data for a patient) and reconstruction (backprojecting to obtain the CT image volume). These steps have been discussed in the previous section. We reiterate that the choice of technical parameters such as reconstruction kernel, tube current, and slice thickness will influence the resulting image, which will in turn impact quantitation.

Image preprocessing refers to any steps that must be taken in order to prepare the image for further analysis. These steps may include applying noise-reduction or contrast-enhancement filters, resampling to achieve a desired spatial resolution, and segmentation of relevant anatomy. Some of these steps are nontrivial research topics in their own right, but a detailed discussion is beyond the scope of this dissertation.

Feature extraction refers to the task of computing imaging features, which are quantitative measures that represent image content for the purposes of classification. Some examples are first-order descriptors such as mean, standard deviation, or median, which reflect the graylevel intensity distribution of the image; and second-order descriptors such as texture features, which characterize the spatial relationships between voxels in the image. Many different types of imaging features exist, and the task of selecting a set of features for a particular classification problem can be both science and art. Depending on the nature of the classification task, feature extraction may be performed globally for the image as a whole, or locally in regions of interest within the image volume.

Feature selection refers to the task of pruning the original feature space to obtain a subset of appropriate features. Feature extraction typically produces a large initial set of features, many of which may be redundant, uninformative, or susceptible to noise. It is often difficult to identify *a priori* which features will be the most useful for a given classification problem; feature selection techniques therefore examine the values of features in the training set to make an *a posteriori* decision. The set of selected feature values for a particular observation is referred to as a feature vector.

Model training refers to the task of tuning a machine learning algorithm based on training data. There are numerous classification techniques such as Bayesian classifiers, support vector machines, decision trees, random forests, artificial neural networks, and many, many more. Each of these is represented by a mathematical model which must be fit to the training data. Additionally, many of these techniques have parameters that influence the behavior of the classifier; these are frequently tuned according to the training data as well.

Finally, once a machine learning algorithm has been trained, the resulting model can be used to classify new observations. Classification of a new observation is typically done by subjecting it to the same preprocessing and feature extraction steps, constructing a feature vector consisting of the previously-identified subset of features, then feeding this feature vector into the classifier model to yield a category label. By classifying a testing set of observations with known categories, the performance of a classifier model can be measured.

1.3 Literature Review

1.3.1 *CT densitometry of emphysema*

Pulmonary emphysema, which is characterized by destruction of lung tissue and the subsequent replacement of lung tissue with air, manifests in computed tomography as areas of low attenuation [3]. In 1988, Mueller, *et al*, proposed the relative area of voxels below a certain image intensity threshold as a measure of emphysema extent [4]. Around the same time, Gould, *et al*, proposed a percentile density score for emphysema that reported the graylevel value associated with a certain percentage on the cumulative image intensity histogram [5]. These CT-derived measures have been shown to correlate well with both spirometric and histopathological assessment of emphysema [6-9].

As a quantitative tool, densitometric measures such as relative area and percentile density are subject to sources of variation. In particular, CT attenuation in the lungs is known to be sensitive to the patient's breathhold at the time of image acquisition, a finding that was reported as early as 1979 [10]. Since then, breathhold has been firmly established as one of the major sources of variation in lung CT densitometry [8, 11].

As a result of the dependence between lung densitometric measures and breathhold, several researchers have adopted volume correction of relative area and percentile density scores in order to account for differences in levels of inspiration between different patients or between multiple examinations of the same patient [11-15]. Many of these approaches essentially build a statistical model of the raw density score as a function of lung volume, then normalize scores by adjusting to a reference breathhold. Volume correction methods have been shown to improve the robustness of density-based emphysema scores, but other researchers have raised concerns that volume correction may result in diminished signal of physiological changes due to emphysema disease progression or treatment effect in longitudinal studies [16].

1.3.2 Quantitative CT approaches for fibrotic interstitial lung disease

Fibrotic interstitial lung diseases (FILD) such as idiopathic pulmonary fibrosis (IPF) and lung disease associated with systemic scleroderma (SSc) are a class of diseases characterized by scarring of lung tissue. In computed tomography, lung abnormalities due to FILD manifest in several classic visual patterns such as fibrosis, ground-glass opacity, and honeycombing (Fig. 1.2). Assessment of disease extent has been shown to play an important role in prognosis [17, 18] and to correlate with independent indicators of disease such as spirometry [19, 20] and histology [21].

Semiquantitative visual assessment of disease has been shown to be a strong independent predictor of patient survival in longitudinal studies [22]. However, visual assessment of disease is a difficult and subjective task, and inter- and intra-reader agreement have been identified as a concern [23]. Honeycombing in particular has proven to be challenging; in a multicenter longitudinal study for IPF, Lynch, *et al*, reported Cohen's kappa scores of 0.21 and 0.31 for reader

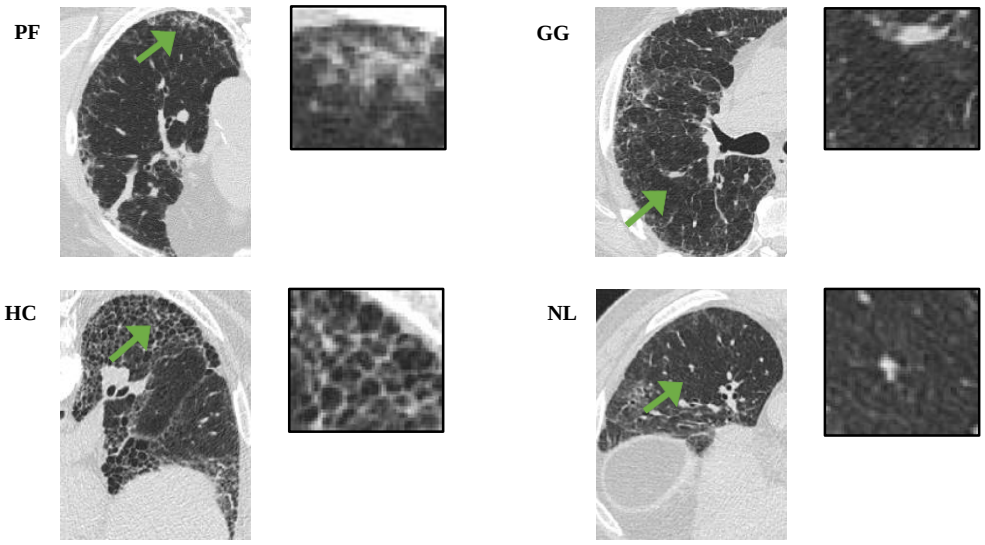

Figure 1.2. Illustration of visual disease patterns for fibrotic interstitial lung disease. (PF) pulmonary fibrosis; (GG) ground-glass opacity; (HC) honeycombing; (NL) normal lung parenchyma.

agreement on the binary presence of honeycombing [17]; a visual assessment of honeycombing extent is likely to be even more challenging.

In light of these concerns, there has been a growing interest in objective, repeatable QCT methods for assessing FILD. In 1994, Hartley, *et al*, reported that mean and median lung attenuation were independently associated with physiological measures of IPF, laying the foundation for density- and histogram-based analysis of FILD [24]. This work was extended by Best, *et al*, who further examined histogram index measures and demonstrated strong associations between skewness and especially kurtosis with spirometric measures of IPF [25].

In contrast to the above methods, a number of studies have examined classification-based methods to characterize FILD disease patterns. In 1999, Uppaluri, *et al*, proposed a method based on histogram features, graylevel co-occurrence matrix (GLCM) and run-length matrix (RLM) texture features, and fractal features [26]. Similar approaches have been proposed by numerous authors [27-29]. Of particular note are a three-dimensional extension of GLCM and RLM texture features [30], and a variation of the approach that improves performance by denoising images prior

to feature extraction [31]. Many of these approaches have been shown to perform well in both region-of-interest and voxelwise classification tasks; some have also demonstrated value in longitudinal assessment of disease progression and treatment efficacy in clinical trials [29, 32]. Furthermore, Kim, *et al*, demonstrated the superiority of a texture-classification-based approach over histogram index measures such as kurtosis in detecting longitudinal change [33].

1.3.3 Robustness of quantitative methods in thoracic CT

There are a number of technical factors chosen during CT acquisition and reconstruction, each of which influences the appearance and quality of the resulting images. An exhaustive examination of all of these is beyond the scope of this dissertation, so for the present discussion we will be focusing our attention on slice thickness, reconstruction kernel, and tube current. Additionally, we will also be discussing reproducibility, which we define to be robustness with respect to repeat-scan variation[1].

Reproducibility of CT lung densitometry in emphysema has been examined by several researchers. In 2001, Gierada, *et al*, published a same-day repeat-scan study consisting of 58 candidates for lung volume reduction surgery, concluding that QCT measurements of emphysema are highly repeatable with no clinically important improvement from spirometric gating [34]. This study is significant for being one of the first emphysema repeat-scan experiments in the literature and for the extremely short interval between baseline and repeat scans; however, it suffers from a relatively small sample size, as only 29 of the subjects underwent repeat scans with identical, non-spirometrically-gated protocols. Furthermore, the study was conducted using thick-section (8mm)

[1] According to guidelines put forth by the Quantitative Imaging Biomarkers Alliance [53], repeat-scan variation more appropriately falls under the definition of repeatability, rather than reproducibility. However, for the sake of internal consistency, we will be using the term "reproducibility" throughout this dissertation.

images, which are less relevant today due to the increasing prevalence of thin-section high-resolution CT, which is particularly common in clinical trials.

Since then, numerous reproducibility studies have been reported for CT lung densitometry with a wide range both in the number of subjects (from 10 to 100+) and in the time interval between repeat scans (from 2 weeks to over a year) [11, 35-39]. Perhaps in part due to concerns over radiation exposure, studies with short time intervals tend to involve fewer subjects as well. Larger sample sizes are preferred because they afford increased statistical power and generalizability of results; however, with longer time intervals it becomes difficult to distinguish repeat-scan variability from variability due to disease progression and/or treatment effect.

A recent study by Balagurunathan, *et al*, examined the reproducibility of CT imaging features in non-small cell lung cancer [40]. Using a 15-minute repeat-scan dataset, the authors evaluated a large number of 3D size, shape, and texture features according to three criteria: reproducibility, dynamic range, and redundancy. A final feature subset was produced by filtering the initial set on these criteria, and each individual feature was evaluated based on its ability to discriminate between two prognostic groups of tumors. This study is significant for its use of reproducibility as a feature selection criterion. However, it stops short of evaluating the selected features in a classification-based framework beyond single-feature discrimination. Moreover, there was no comparison between reproducible and non-reproducible features, and so the study fails to demonstrate the importance of selecting based on feature reproducibility.

Slice thickness impacts image quality in two distinct ways. First, for a given tube current, thinner slices result in noisier images, since the number of x-ray photons captured in each slice is directly proportional to the thickness of the slice. Second, thicker slices result in decreased spatial

resolution due to reduced sampling in the longitudinal direction combined with increased partial volume effect.

The reconstruction kernel is an essential component of the filtered backprojection method which is used to produce CT images. The choice of kernel has a profound impact on image quality by influencing spatial resolution and image noise. Broadly speaking, kernels can be described by their sharpness or smoothness. Sharp kernels (also referred to as enhancing or bone kernels) typically provide enhancement of edges at the cost of increased image noise. By contrast, smooth kernels (sometimes referred to as soft tissue kernels) provide increased blurring, resulting in decreased image noise but reduced spatial resolution. Reconstruction kernels are often chosen on the basis of the anatomy of interest or the imaging task at hand.

Studies investigating the influence of slice thickness or reconstruction kernel often make use of CT raw sinogram data reconstructions to produce multiple images from a single acquisition. This approach has the dual advantage of allowing a researcher to isolate variation due to these technical factors (since there is no source of repeat-scan variation) while simultaneously saving subjects the additional radiation exposure that a repeat-scan study would involve. Following this study design, numerous researchers have demonstrated that densitometric measures of emphysema are sensitive to the choice of slice thickness and kernel, although lung volume measurements have been shown to be quite robust [41-45].

The effect of tube current on CT image quality is well known. Tube current, which is measured in units of milliampere-seconds (mAs), is directly proportional to the number of x-ray photons which are used to generate the image. Decreasing tube current reduces image quality (by increasing the Poisson noise), yielding images that appear grainier and more speckled. This has

been shown to have a noticeable impact on emphysema QCT, particularly at low levels of tube current (20 mAs or lower) [46].

Compared to slice thickness and reconstruction kernel, studies investigating tube current are more difficult because tube current is specified at the time of image acquisition rather than reconstruction. However, the development of CT noise simulation algorithms, such as those described in [47-49], have made it possible to create simulated reduced tube current images from a single acquisition. Zaporozhan, *et al*, examined simulated low-dose CT images, reporting that dose reduction down to 30 simulated mAs was possible without introducing a clinically relevant degree of variation in a density-based emphysema index score, while a separate morphological analysis remained stable down to 50 simulated mAs [50].

Additionally, several studies have investigated the robustness of texture features in various CT applications. Al-Kadi, *et al*, analyzed a variety of texture features including model, statistical, and wavelet features in lung tumors under conditions of simulated noise reduction and enhancement, concluding that graylevel co-occurrence matrix (GLCM) and run-length matrix (RLM) features were among the least robust to image noise [51]. Although this study examined multiple features, the feature robustness within each category was analyzed collectively, and data on individual features was not reported. In another study, Guggenbuhl, *et al*, investigated the robustness of GLCM and RLM features with respect to slice thickness by imaging bovine bone samples at multiple slice thicknesses (1, 3, 5, 8 mm), reporting significant changes in many of the features [52]. These and other studies illustrate the potential susceptibility of GLCM and RLM texture features to CT technical factors.

Two observations are apparent from the preceding review. The first observation is the importance of distinguishing between statistical significance and clinical relevance. For example,

Behrendt, *et al*, reported statistically significant differences in total lung volume across multiple reconstruction kernels; however, these differences were quite small (within 0.15% of the reference kernel), leading the authors to conclude that the finding was not clinically important [42]. Similarly, Zaporozhan, *et al*, reported statistically significant differences in emphysema measures across a wide range of simulated reduced tube currents, but only tube currents below 30 mAs resulted in a clinically relevant difference (defined to be 2% variation from the original clinical dose) [50]. However, we note that in the context of CT classification, it is difficult to define clinical relevance for imaging features because the value of an individual imaging feature is often not easily relatable to clinical outcome.

The second observation is that not all QCT measures are equally sensitive to changes in technical parameters, such as total lung volume versus densitometric measures of emphysema extent. Even among emphysema measures, the choice of HU threshold for relative area has a substantial impact on the measure's sensitivity to slice thickness and kernel, with higher thresholds being more robust than lower ones despite still being correlated with histology [44]. In a similar vein, Al-Kadi has shown that some categories of texture features are more robust than others [51]. These findings suggest that there is value in examining robustness as a criterion for evaluating imaging features.

1.4 Summary of Key Contributions

This dissertation consists of two separate but related studies in the area of robustness of quantitative CT. In the first study, we examined repeat CT images taken one week apart from a multi-center clinical trial for chronic obstructive pulmonary disease (COPD). We characterized and compared several widely-accepted QCT measures of emphysema disease extent, demonstrating the superiority of non-volume-adjusted relative area below -950 HU from a signal-

to-noise perspective. We also demonstrated that densitometric reproducibility is significantly impacted by breathhold reproduction. Lastly, we investigated reproducibility characteristics across multiple sites and proposed a framework for assessing each site's performance in achieving reproducible breathholds. This study contributes to the field of emphysema QCT by performing these valuable analyses on one of the largest known short-term repeat-scan datasets for COPD.

In the second study, we investigated quantitative feature and classifier robustness with respect to slice thickness, reconstruction kernel, and tube current in the setting of CT classification of fibrotic interstitial lung disease (FILD). We developed a novel approach for assessing the robustness of imaging features by examining their stability across multiple systematic reconstructions of CT raw sinogram data. Leveraging this approach, we proposed a novel method called Robustness-Driven Feature Selection (RDFS) for identifying a subset of robust features, then used these features to develop a robust support vector classifier for lung structure and parenchymal abnormalities in FILD. We demonstrated the effectiveness of our proposed methodology by comparing our robust classifier to a similar classifier that did not utilize RDFS. To our knowledge, we were the first to propose a framework for assessing feature and classifier robustness in the area of FILD by systematically applying CT raw sinogram reconstructions. Moreover, we were the first to examine the connection between feature robustness and classifier robustness.

The rest of this dissertation is organized as follows. Our study on emphysema reproducibility is described in chapters 2 and 3. In chapter 2, we present a paper titled "Reproducibility of volume and densitometric measures of emphysema on repeated computed tomography with an interval of 1 week", which we previously published in *European Radiology*.

In chapter 3, we present an extension of this study with a much larger sample size and expanded analysis to examine reproducibility characteristics across multiple sites.

Our study on quantitative feature and classifier robustness is described in chapters 4 through 6. In chapter 4, we introduce our classification approach for FILD and investigate several previously-published techniques for addressing the problem of imbalanced data learning. In chapter 5, we develop our Robustness-Driven Feature Selection method and demonstrate its effectiveness at improving classifier robustness. In chapter 6, we extend our work on robustness to examine slice thicknesses of 3.0 mm. Lastly, in Appendix A, we examine reader agreement between the expert readers who provided the ground truth for this study.

1.5 References

[1] Bushberg JT, Seibert JA, Leidholt EM Jr, Boone JM. The Essential Physics of Medical Imaging, 2nd ed. Lippincott Williams & Wilkins 2002.

[2] McCollough CH, Primak AN, Braun N, Kofler J, Yu L, Christner J. Strategies for reducing radiation dose in CT. Radiol Clin North Am 2009. 47(1):27-40.

[3] Hayhurst MD, MacNee W, Flenley DC, Wright D, McLean A, Lamb D, Wightman AJ, Best J. Diagnosis of pulmonary emphysema by computerized tomography. Lancet 1984. 2(8398):320-2.

[4] Muller NL, Staples CA, Miller RR, Abboud RT. "Density mask". An objective method to quantitate emphysema using computed tomography. Chest 1988. 94(4):782-7.

[5] Gould GA, MacNee W, McLean A, Warren PM, Redpath A, Best JJ, Lamb D, Flenley DC. CT measurements of lung density in life can quantitate distal airspace enlargement – an essential defining feature of human emphysema. Am Rev Respir Dis 1988. 137(2):380-92.

[6] Spouge D, Mayo JR, Cardoso W, Muller NL. Panacinar emphysema: CT and pathologic findings. J Comput Assist Tomogr 1993. 17(5):710-3.

[7] Gevenois PA, De Vuyst P, de Maertelaer V, Zanen J, Jacobovitz D, Cosio MG, Yernault JC. Comparison of computed density and microscopic morphometry in pulmonary emphysema. Am J Respir Crit Care Med 1996. 154(1):187-92.

[8] Goldin JG. Quantitative CT of emphysema and the airways. J Thorac Imaging 2004. 19(4):235-40.

[9] Pauls S, Gulkin D, Feuerlein S, Muche R, Kruger S, Schmidt SA, Dharaiya E, Brambs HJ, Hetzel M. Assessment of COPD severity by computed tomography: correlation with lung functional testing. Clin Imaging 2010. 34(3):172-8.

[10] Robinson PJ, Kreel L. Pulmonary tissue attenuation with computed tomography: comparison of inspiration and expiration scans. J Comput Assist Tomogr 1979. 3(6):740-8.

[11] Shaker SB, Dirksen A, Laursen LC, Skovgaard LT, Holstein-Rathlou NH. Volume adjustment of lung density by computed tomography scans in patients with emphysema. Acta Radiol 2004. 45(4):417-23.

[12] Dirksen A, Friis M, Olesen KP, Skovgaard LT, Sorensen K. Progress of emphysema in severe alpha 1-antitrypsin deficiency as assessed by annual CT. Acta Radiol 1997. 38(5):826-32.

[13] Parr DG, Stoel BC, Stolk J, Stockley RA. Validation of computed tomographic lung densitometry for monitoring emphysema in alpha1-antitrypsin deficiency. Thorax 2006. 61(6):485-90.

[14] Stoel BC, Putter H, Bakker ME, Dirksen A, Stockley RA, Piitulainen E, Russi EW, Parr D, Shaker SB, Reiber JH, Stolk J. Volume correction in computed tomography densitometry for follow-up studies on pulmonary emphysema. Proc Am Thorac Soc 2008. 5(9):919-24.

[15] Lynch DA, Al-Qaisi MA. Quantitative computed tomography in chronic obstructive pulmonary disease. J Thorac Imaging 2013. 28(5):284-90.

[16] Cazzola M, MacNee W, Martinez FJ, Rabe KF, Franciosi LG, Barnes PJ, Brusasco V, Burge PS, Calverley PM, Celli BR, Jones PW, Mahler DA, Make B, Miravitlles M, Page CP, Palange P, Parr D, Pistolesi M, Rennard SI, Rutten-van Mölken MP, Stockley R, Sullivan SD, Wedzicha JA, Wouters EF. Outcomes for COPD pharmacological trials: from lung function to biomarkers. Eur Respir J 2008. 31(2):416-69.

[17] Lynch DA, Godwin JD, Safrin S, Starko KM, Hormel P, Brown KK, Raghu G, King TE Jr, Bradford WZ, Schwartz DA, Richard Webb W. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. Am J Respir Crit Care Med 2005. 172(4):488-93.

[18] Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, Arriola E, Silver R, Strange C, Bolster M, Seibold JR, Riley DJ, Hsu VM, Varga J, Schraufnagel DE, Theodore A, Simms R, Wise R, Wigley F, White B, Steen V, Read C, Mayes M, Parsley E, Mubarak K, Connolly MK, Golden J, Olman M, Fessler B, Rothfield N, Metersky M. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006. 354(25):2655-66.

[19] Staples CA, Muller NL, Vedal S, Abboud R, Ostrow D, Miller RR. Usual interstitial pneumonia: correlation of CT with clinical, functional, and radiologic findings. Radiology 1987. 162(2):377-81.

[20] Xaubet A, Agusti C, Luburich P, Roca J, Monton C, Ayuso MC, Barbera JA, Rodriguez-Roisin R. Pulmonary function tests and CT scan in the management of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 1998. 158(2):431-6.

[21] Kazerooni EA, Martinez FJ, Flint A, Jamadar DA, Gross BH, Spizarny DL, Cascade PN, Whyte RI, Lynch JP 3rd, Toews G. Thin-section CT obtained at 10-mm increments versus limited three-level thin-section CT for idiopathic pulmonary fibrosis: correlation with pathologic scoring. AJR Am J Roentgenol 1997. 169(4):977-83.

[22] Best AC, Meng J, Lynch AM, Bozic CM, Miller D, Grunwald GK, Lynch DA. Idiopathic pulmonary fibrosis: physiologic tests, quantitative CT indexes, and CT visual scores as predictors of mortality. Radiology 2008. 246(3):935-40.

[23] Lynch DA. Quantitative CT of fibrotic interstitial lung disease. Chest 2007. 131(3):643-4.

[24] Hartley PG, Galvin JR, Hunninghake GW, Merchant JA, Yagla SJ, Speakman SB, Schwartz DA. High-resolution CT-derived measures of lung density are valid indexes of interstitial lung disease. J Appl Physiol 1994. 76(1):271-7.

[25] Best AC, Lynch AM, Bozic CM, Miller D, Grunwald GK, Lynch DA. Quantitative CT indexes in idiopathic pulmonary fibrosis: relationship with physiologic impairment. Radiology 2003. 228(2):407-14.

[26] Uppaluri R, Hoffman EA, Sonka M, Hartley PG, Hunninghake GW, McLennan G. Computer recognition of regional lung disease patterns. Am J Respir Crit Care Med 1999. 160(2):648-54.

[27] Uchiyama Y, Katsuragawa S, Abe H, Shiraishi J, Li F, Li Q, Zhang CT, Suzuki K, Doi K. Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography. Med PHys 2003. 30(9):2440-54.

[28] Wang J, Li F, Doi K, Li Q. Computerized detection of diffuse lung disease in MDCT: the usefulness of statistical texture features. Phys Med Biol 2009. 54(22):6881-99.

[29] Maldonado F, Moua T, Rajagopalan S, Karwoski RA, Raghunath S, Decker PA, Hartman TE, Bartholmai BJ, Robb RA, Ryu JH. Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. Eur Respir J 2014. 43(1):204-12.

[30] Xu Y, van Beek EJ, Hwanjo Y, Guo J, McLennan G, Hoffman EA. Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). Acad Radiol 2006. 13(8):969-78.

[31] Kim HJ, Li G, Gjertson D, Elashoff R, Shah SK, Ochs R, Vasunilashorn F, Abtin F, Brown MS, Goldin JG. Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. Acad Radiol 2008. 15(8):1004-16.

[32] Kim HJ, Brown Ms, Elashoff R, Li G, Gjertson DW, Lynch DA, Strollo DC, Kleerup E, Chong D, Shah SK, Ahmad S, Abtin F, Tashkin DP, Goldin JG. Quantitative texture-based assessment of one-year changes in fibrotic reticular patterns on HRCT in scleroderma lung disease treated with oral cyclophosphamide. Eur Radiol 2011. 21(12):2455-65.

[33] Kim HJ, Brown MS, Chong D, Djertson DW, Lu P, Kim HJ, Coy H, Goldin JG. Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months. Acad Radiol 2015. 22(1):70-80.

[34] Gierada DS, Yusen RD, Pilgram TK, Crouch L, Slone RM, Bae KT, Lefrak SS, Cooper JD. Repeatability of quantitative CT indexes of emphysema in patients evaluated for lung volume reduction surgery. Radiology 2001. 220(2):448-54.

[35] Stolk J, Dirksen A, van der Lugt AA, Hutsebaut J, Mathieu J, de Ree J, Reiber JH, Stoel BC. Repeatability of lung density measurements with low-dose computed tomography in subjects with alpha-1-antitrypsin deficiency-associated emphysema. Invest Radiol 2001. 36(11):648-51.

[36] McGregor A, Roberts HC, Dong Z, Menezes R, Kauczor HU, Weinheimer O, Heussel CP. Repeated low-dose computed tomography in current and former smokers for quantification of emphysema. J Comput Assist Tomogr 2010. 34(6):933-8.

[37] Keller BM, Reeves AP, Yankelevitz DF, Henschke CI. Automated quantification of pulmonary emphysema from computed tomography scans: comparison of variation and correlation of common measures in a large cohort. In Proc SPIE Medical Imaging 2010. Vol 7624.

[38] Mets OM, Igsum I, Mol CP, Gietama HA, Zanen P, Prokop M, de Jong PA. Variation in quantitative CT air trapping in heavy smokers on repeat CT examinations. Eur Radiol 2012. 22(12):2710-7.

[39] Park SJ, Lee Ch, Goo JM, Heo CY, Kim JH. Inter-scan repeatability of CT-based lung densitometry in the surveillance of emphysema in a lung cancer screening setting. Eur J Radiol 2012. 81(4):554-60.

[40] Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, Goldgof DB, Hall Lo, Kkorn R, Zhao B, Schwartz LH, Basu S, Eschrich S, Gatenby RA, Gilles RJ. Test-retest reproducibility analysis of lung CT image features. J Digit Imaging 2014. 27(6):805-23.

[41] Boedeker KL, McNitt-Gray MF, Rogers SR, Truong DA, Brown MS, Gjertson DW, Goldin JG. Emphysema: effect of reconstruction algorithm on CT imaging measures. Radiology 2004. 232(1):295-301.

[42] Behrendt FF, Das M, Mahnken AH, Kraus T, Bakai A, Stanzel S, Gunther RW, Wildberger JE. Computer-aided measurements of pulmonary emphysema in chest multidetector-row spiral computed tomography: effect of image reconstruction parameters. J Comput Assist Tomogr 2008. 32(6):899-904.

[43] Ley-Zaporozhan J, Ley S, Weinheimer O, Iliyushenko S, Erdugan S, Eberhardt R, Fuxa A, Mews J, Kauczor HU. Quantitative analysis of emphysema in 3D using MDCT: influence of different reconstruction algorithms. Eur J Radiol 2008. 65(2):228-34.

[44] Gierada DS, Bierhals AJ, Choong CK, Bartel ST, Ritter JH, Das NA, Hong C, Pilgram TK, Bae KT, Whiting BR, Woods JC, Hogg JC, Lutey BA, Battafarano RJ, Cooper JD, Meyers BF, Patterson GA. Effects of CT section thickness and reconstruction kernel on emphysema quantification: relationship to the magnitude of the CT emphysema index. Acad Radiol 2010. 17(2):146-56.

[45] Bartel ST, Bierhals AJ, Pilgram TK, Hong C, Schechtman KB, Conradi SH, Gierada DS. Equating quantitative emphysema measurements on different CT image reconstructions. Med Phys 2011. 38(8):4898-902.

[46] Madani A, De Maertelaer V, Zanen J, Gevenois PA. Pulmonary emphysema: radiation dose and section thickness at multidetector CT quantification – comparison with macroscopic and microscopic morphometry. Radiology 2007. 243(1):250-7.

[47] Mayo JR, Whittall KP, Leung AN, Hartman TE, Park CS, Primack SL, Chambers GK, Limkeman MK, Toth TL, Fox SH. Simulated dose reduction in conventional chest CT: validation study. Radiology 1997. 202(2):453-7.

[48] Massoumzadeh P, Don S, Hildebolt F, Bae KT, Whiting BR. Validation of CT dose-reduction simulation. Med Phys 2009. 36(1):174-89.

[49] Zabic S, Wang Q, Morton T, Brown KM. A low dose simulation tool for CT systems with energy integrating detectors. Med Phys 2013. 40(3):031102.

[50] Zaporozhan J, Ley S, Weinheimer O, Eberhardt R, Tsakiris I, Noshi Y, Herth F, Kauczor HU. Multi-detector CT of the chest: influence of dose onto quantitative evaluation of severe emphysema: a simulation study. J Comput Assist Tomogr 2006. 30(3):460-8.

[51] Al-Kadi OS. Assessment of texture measures susceptibility to noise in conventional and contrast enhanced computed tomography lung tumour images. Comput Med imaging Graph 2010. 34(6):494-503.

[52] Guggenbuhl P, Chappard D, Garreau M, Bansard JY, Chales G, Rolland Y. Reproducibility of CT-based bone texture parameters of cancellous calf bone samples: influence of slice thickness. Eur J Radiol 2008. 67(3):514-20.

[53] Obuchowski NA, Reeves AP, Huang EP, Wang XF, Buckler AJ, Kim HJ, Barnhart HX, Jackson EF, Giger ML, Pennello G, Toledano AY, Kalpathy-Cramer J, Apanasovich TV, Kinahan PE, Myers KJ, Koldgof DB, Barboriak DP, Gillies RJ, Schwartz LH, Sullivan DC. Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons. Stat Methods Med Res 2015. 24(1):68-106.

2. Reproducibility of volume and densitometric measures of emphysema on repeat computed tomography with an interval of 1 week

The following chapter is adapted from the final submitted manuscript of the paper "Reproducibility of volume and densitometric measures of emphysema on repeat computed tomography with an interval of 1 week", by D Chong, MS Brown, HJ Kim, EM van Rikxoort, L Guzman, MF McNitt-Gray, M Khatonabadi, M Galperin-Aizenberg, H Coy, K Yang, Y Jung, JG Goldin in *European Radiology* 2012, 22(2):287-94.

The final publication is available at springerlink.com. An electronic version of the paper may be retrieved at http://link.springer.com/article/10.1007/s00330-011-2277-1.

2.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a significant cause of morbidity and mortality and represents a large economic and social burden worldwide, particularly in developed countries [1, 2]. COPD is typically diagnosed and assessed clinically via pulmonary function tests (PFT) such as forced expiratory volume in 1 s (FEV_1) and diffusion of carbon monoxide (D_LCO). Emphysema, a subset of COPD which is characterized by lung tissue destruction, is known to manifest as areas of low attenuation in computed tomography (CT) [3]. Densitometric measures such as density mask and percentile density were therefore proposed as quantitative means of assessing extent and progression of emphysema [4, 5]. These CT-derived measures have been shown to correlate well with emphysema assessment via PFT and histopathology [4, 6-9].

More recently, CT densitometric measures have been proposed as an efficacy endpoint in emphysema treatment trials. The dependence between CT attenuation and level of inspiration was reported as early as 1979 by Robinson and Kreel [10], and differences in breath-hold became well-

established as one of the major sources of variation in densitometric measures [6, 11]. The question of density measure reproducibility is therefore intimately tied to breath-hold reproducibility. When tracking the progression of emphysema over the course of months and years, the reproducibility of these measures is an important concern. In a clinical trial setting, this has both economic and scientific impacts – more reproducible measures lead to smaller subject populations and shorter study durations.

The purpose of this study is to assess the reproducibility of both breath-hold (as measured by CT lung volume) and commonly used CT densitometric measures of emphysema by evaluating patients in a scan-rescan setting within a strictly controlled clinical trial over a period of 1 week. Additionally, the relationship between the reproducibilities of density measures and breath-hold is investigated.

2.2 Materials and Methods

2.2.1 Patients

The study population consisted of patients enrolled in multicenter clinical trials whose CT images were made available in an anonymized research database. Each of these patients gave signed consent with the approval of their local institutional review board to sending their anonymized data to a central imaging core. All patients were current or former smokers with a clinical diagnosis of COPD. Patients were selected for this study based on the availability of two baseline scans, and 44 patients (17 male, 27 female, median age 56 years) were found to match this criterion. All use of anonymized image data was in keeping with the Health Insurance Portability and Accountability Act.

2.2.2 CT image acquisition

All CTs were performed under the auspices of a central imaging core that provided training and standardized guidelines for image acquisition. Subjects underwent CT at full inspiration (total lung capacity, or TLC) on two occasions, with a median interval of 7 days between the two visits (minimum 5 days, maximum 17 days). Patients were instructed in proper breath-holding technique and were given the opportunity to practice before imaging. A variety of CT systems were used from General Electric (GE Healthcare, Little Chalfont, Buckinghamshire, UK), Toshiba (Toshiba America Medical Systems, Tustin, CA, USA), or Siemens (Siemens Healthcare, Erlangen, Germany). Devices were calibrated for air and water, as per site protocols. CT was performed using an X-ray tube voltage of 120 kVp, with tube currents ranging from 200 to 300 mA in the GE devices, 300 mA in the Toshiba ones, and 80 to 150 effective mAs for Siemens systems. Contiguous, thin-section imaging was performed with slice thicknesses ranging from 0.6 mm to 1.25 mm. Images were reconstructed using a sharp algorithm of Bone (GE), FC51 (Toshiba), or B45f (Siemens). For each patient, the same equipment, imaging protocol, and reconstruction parameters were used for both the first and second visit CT with the exception of one patient, whose tube current differed between the two visits (300 mA and 240 mA). Additionally, a water phantom was imaged (with a field of view large enough to include air) using the same imaging protocol and system within 24 h of each patient's CT.

2.2.3 Quantitative CT Image Analysis

The CT images were analyzed using a software package developed in-house [12, 13]. Lungs were segmented by an automated algorithm using a threshold of -500 HU followed by manual editing to correct segmentation errors. A guided semi-automatic system was used to identify the airway

tree, which was then used to exclude the gross airway structure from the lung segmentation, resulting in a region of interest (ROI) representing the lungs. The volume of the lungs was computed based on the size of this ROI.

Next, a histogram representing the distribution of Hounsfield Unit (HU) values was generated for the lung ROI. This histogram was used to derive two different densitometric measures: the relative area below -950 HU (RA950-raw) and the 15^{th} percentile density (PD15-raw). These densitometric measures were chosen because they have been shown to correlate well with histopathology [4, 14] and because the thresholds of -950 HU and the 15^{th} percentile have been widely adopted in the assessment of emphysema in thin-section CT [9, 15-17].

Next, relative area and percentile density values were adjusted via volume correction. Volume correction was performed due to the endorsement of this procedure by numerous authors [5, 11, 15, 17] and involved fitting the data to the mixed-effects regression model described in [5] to produce RA950-adj and PD15-adj, respectively.

2.2.4 Statistical Analysis

For each patient, the reproducibilities between the first and second scans of the various measures (volume, RA950-raw, RA950-adj, PD15-raw, PD15-adj) were assessed via the following methods. First, a paired t-test was used to test for a difference in means between the first and second CTs. Next, the concordance correlation coefficient (CCC) was used to assess the agreement between the measured values in the first and second CTs [18]. Lastly, Bland–Altman analysis was used to compute the repeatability coefficient (RC) for each measure, giving an indication of the variability of each measure [19].

The reproducibility of RA950-raw was compared against that of RA950-adj by using a robust difference-of-variances test (Brown-Forsythe's test) to compare their RCs. The RCs of PD15-raw and PD15-adj were compared in a similar fashion.

Scatter plots of density versus volume were used to show the relationship between densitometric and breath-hold reproducibilities. A scatter plot was created by plotting the paired difference in RA950-raw against the paired difference in volume. In a similar fashion, paired differences in RA950-adj, PD15-raw, and PD15-adj were all plotted against the paired difference in volume, yielding a total of four scatter plots. Linear regression analysis was performed on each of these plots to evaluate the relationship between the difference in volume and each of the four densitometric measures.

Using an *a priori* chosen threshold of 0.25 L to represent superior breath-hold reproduction, the patients were divided into two subgroups based on the paired difference in CT lung volumes. The statistics of CCC and RC in RA950-raw and PD15-raw were computed for the two subgroups. Brown–Forsythe's test was used to compare the RC of RA950-raw between the <0.25-L and ≥0.25-L subgroups. Similarly, the RC of PD15-raw of the <0.25-L and ≥0.25-L subgroups were compared.

All data were analyzed using a combination of Microsoft Excel 2003 (Microsoft; Redmond, WA) and Stata 8.0 (StataCorp; College Station, TX). For statistical tests, a p value smaller than 0.05 was considered to be significant.

Table 2.1. Subject baseline characteristics

Measure	Overall[c] (n = 44) mean ± SD	<0.25 L[d] (n = 33) mean ± SD	≥0.25 L[d] (n = 11) mean ± SD
Volume[a] (L)	5.77 ± 1.54	5.93 ± 1.61	5.29 ± 1.20
RA950-raw[a] (%)	16.0 ± 11.7	18.1 ± 12.3	9.4 ± 6.2
RA950-adj[b] (%)	15.1 ± 9.9	---	---
PD15-raw[a] (HU)	-946.3 ± 28.3	-951.3 ± 29.3	-931.2 ± 19.1
PD15-adj[b] (HU)	-950.4 ± 21.0	---	---

[a] Calculated directly from image histograms. [b] Obtained via volume correction techniques [5]. [c] Overall population of subjects. [d] Subgroup analysis using breath-hold reproduction threshold of 0.25 L.

Table 2.2. Quantitative measure reproducibility characteristics

Measure	n	CCC[a]	Difference mean ± SD	p value[b]	RC[c]
V (L)	44	0.976	-0.056 ± 0.336	0.279	0.672
RA950-raw (%)					
Overall	44	0.995	0.01% ± 1.17%	0.959	2.34%
<0.25 L	33	0.998	0.15% ± 0.81%	0.286	1.62%
≥0.25 L	11	0.954	-0.42% ± 1.87%	0.472	3.75%
RA950-adj (%)	44	0.996	0.22% ± 0.83%	0.083	1.65%
PD15-raw (HU)					
Overall	44	0.982	0.52 ± 5.29	0.516	10.59
<0.25 L	33	0.998	-0.67 ± 1.85	0.046*	3.70
≥0.25 L	11	0.862	4.09 ± 9.53	0.185	19.07
PD15-adj (HU)	44	0.996	-0.42 ± 1.82	0.135	3.63

[a] Concordance correlation coefficient. Indicates the agreement between two sets of measurements, with perfect agreement yielding +1. [b] Paired t-test comparing difference mean to zero. [c] Bland–Altman repeatability coefficient, defined as RC = 2 × Difference SD. * $p < 0.05$ different from zero

2.3 Results

2.3.1 Volume analysis and reproducibility

The baseline patient characteristics for the subjects in this study are summarized in the first column of Table 2.1. The lung volumes demonstrated very good agreement between the first and second CTs (CCC=0.976; Table 2.2). The volumes had a mean difference of -0.056 L, which was not significantly different from zero ($p = 0.279$). The distribution of volumes is illustrated in Fig. 2.1.

Aside from one outlier (ΔV = -1.49 L), all of the patients reproduced their breath-holds to within 1 L, and the overall RC was approximately two-thirds of a liter.

Of the 44 subjects, 33 achieved superior breath-hold reproduction ($|\Delta V| < 0.25$ L), leaving 11 subjects in the ≥0.25-L subgroup. Lung volumes were similar between the two subgroups, but the <0.25-L subgroup exhibited more severe emphysema as measured by both RA950-raw and PD15-raw (Table 2.1).

2.3.2 Densitometric analysis and reproducibility

The densitometric measures all exhibited very good agreement between the first and second CTs (CCC > 0.95; Table 2.2) with the exception of PD15-raw in the ≥0.25-L subgroup (CCC = 0.862). The mean differences were not significantly different from zero with the exception of PD15-raw in the <0.25-L subgroup ($p = 0.046$). The distributions of the densitometric measures are illustrated in Fig. 2.2. The outlier observed in the plots of RA950-raw (Fig. 2.2a, ΔRA950-raw = -4.94%) and PD15-raw (Fig. 2.2c, ΔPD15-raw = 26 HU) corresponds to the outlier in Fig. 2.1b (ΔV = -1.49).

For both RA950 and PD15, the volume corrected measure exhibited a smaller RC than the raw measure (Table 2.3, rows 1–2). However, this difference was only significant for PD15 ($p = 0.012$). For both RA950-raw and PD15-raw, the RC was significantly smaller in the <0.25-L subgroup than in the ≥0.25-L subgroup ($p = 0.010$, $p < 0.001$ respectively; rows 3–4).

Compared with RA950-raw in the overall population, the coefficient of determination R^2 decreased in the <0.25-L subgroup as well as in the volume corrected measure RA950-adj (Table 2.4). A similar trend was observed in PD15, but the effect in PD15-adj was much more pronounced.

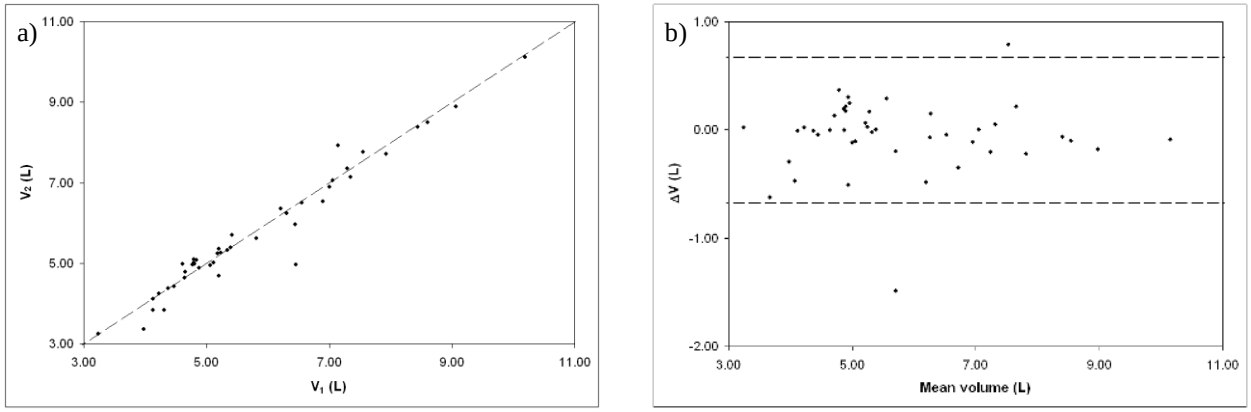

Figure 2.1. (a) Scatter plot illustrating distribution of lung volumes between first and second CTs. Dashed line indicates the identity line. (b) Bland-Altman plot illustrating distribution of lung volumes between first and second CTs. Dashed lines indicate ± repeatability coefficient (RC = 0.67 L).

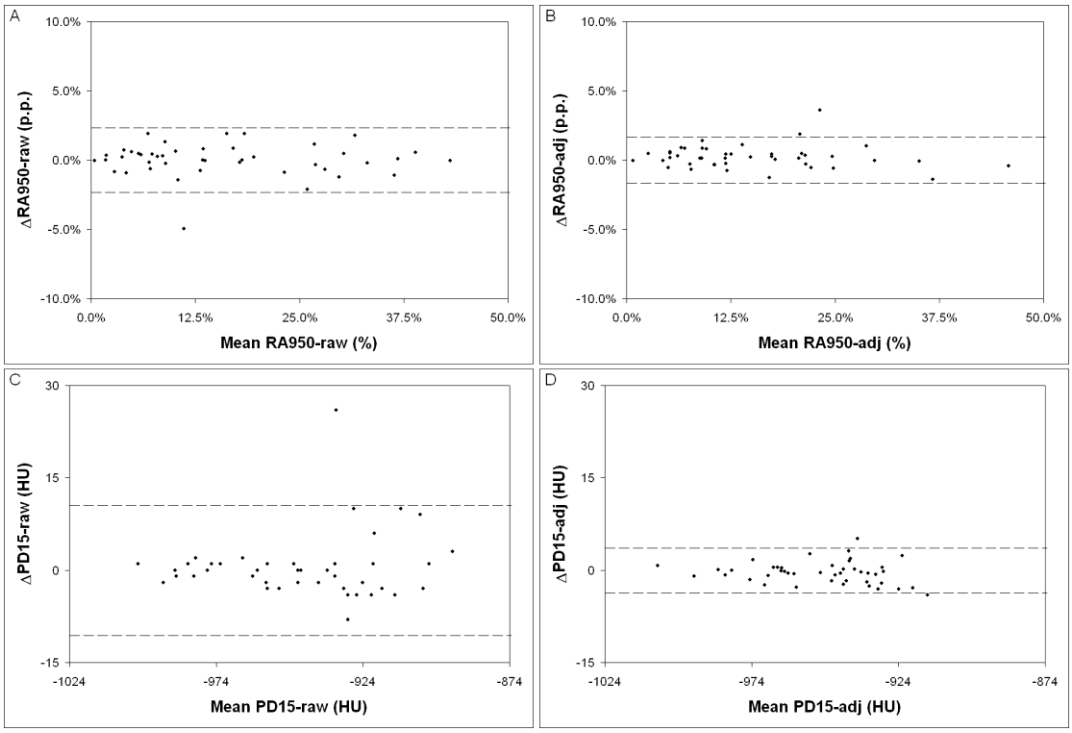

Figure 2.2. Bland-Altman plots illustrating distribution of densitometric measures. Dashed lines indicate ±repeatability coefficient (RC). (a) RA950-raw (RC = 2.31%); (b) RA950-adj (RC = 1.62%); (c) PD15-raw (RC = 10.47 HU); (d) PD15-adj (RC = 3.76 HU).

Table 2.3. Comparisons of repeatability coefficients

Comparison	RC_1	RC_2	P value[c]
RA950-raw vs. RA950-adj (Overall group)[a]	2.34%	1.65%	0.193
PD15-raw vs. PD15-adj (Overall group)[a]	10.59 HU	3.63 HU	0.012*
≥0.25-L vs. <0.25-L (RA950-raw)[b]	3.75%	1.62%	0.010*
≥0.25-L vs. <0.25-L (PD15-raw)[b]	19.07 HU	3.70 HU	<0.001*

[a] Comparison between raw and volume-corrected measures in the overall population (n = 44). [b] Comparison of raw measures between the ≥0.25-L (n = 12) and the <0.25-L (n = 33) subgroups. [c] Brown–Forsythe's robust difference-of-variances test.

* $p < 0.05$ significant difference between RC_1 and RC_2

Table 2.4. Coefficients of determination of differences in densitometric measures against differences in volume

Measure	n	R^2 against ΔV
RA950-raw (%)		
Overall	44	0.7113
<0.25 L	33	0.4225
≥0.25 L	11	0.8843
RA950-adj (%)	44	0.1023
PD15-raw (HU)		
Overall	44	0.7526
<0.25 L	33	0.3052
≥0.25 L	11	0.7996
PD15-adj (HU)	44	<0.0001

2.4 Discussion

This study offers insight into the measurement variability inherent in the patient–CT imaging chain by investigating CT lung volume as well as CT densitometric measures of emphysema. The values of CT lung volume, the relative area below -950 HU, and the volume-adjusted 15th percentile density are reproducible over a period of 1 week. This short time interval between CTs is a unique aspect that is crucial to the study as it minimizes the potential impact of changes in a patient's anatomy or histopathology on the quantitative measures. Unlike a coffee break (same-day repeat imaging) study, an interval of 1 week more realistically assesses the variation due to the interaction between the patient and technologist rather than just CT variability.

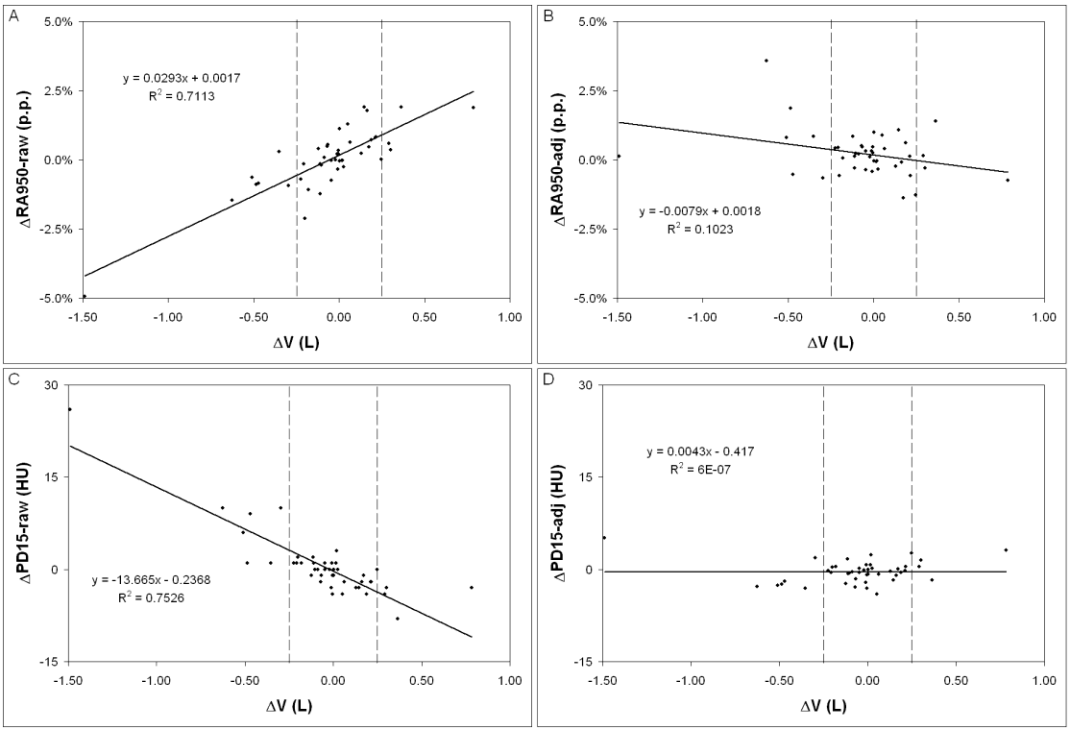

Figure 2.3. Scatter plot illustrating the relationship between density reproducibility and volume reproducibility. Solid lines represent lines of best fit, while dotted lines represent ±0.25 L, the threshold for superior breathhold reproduction. (a) ΔRA950-raw vs ΔV; (b) ΔRA950-adj vs ΔV; (c) ΔPD15-raw vs ΔV; (d) ΔPD15-adj vs ΔV

Our dataset revealed good breath-hold reproduction between the first and second CTs. Out of the 44 subjects in this study, 39 successfully reproduced their breath-holds to within 10%, and 33 were within an even stricter standard of 0.25 L. This compares favorably with ATS and ERS guidelines for acceptability of TLC measurements obtained on conventional pulmonary function tests [20]. It seems that with a carefully controlled imaging protocol that includes detailed breath-holding instructions, it is reasonable to expect well-reproduced TLC breath-holds from the same patient across multiple visits, a result that has implications for longitudinal studies that rely on full-inspiration CT images [21, 22].

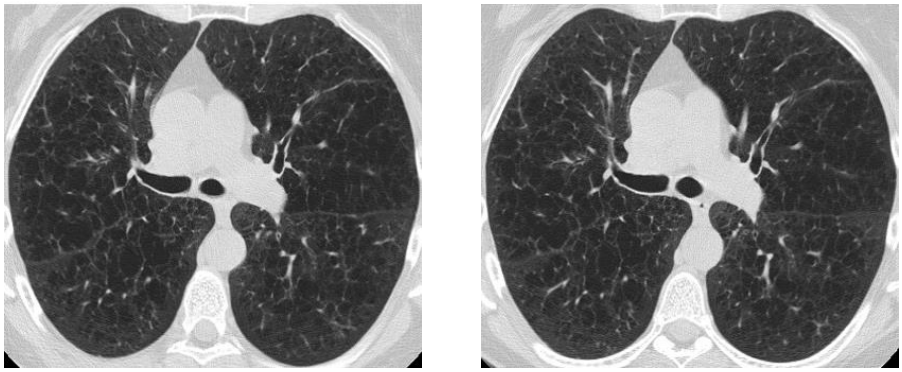

Figure 2.4. Representative paired images from a subject with well-reproduced breath-holds. Subject demonstrates ΔV= -0.02 L, ΔRA950-raw = 0.01 p.p., ΔPD15-raw = 0 HU, ΔRA950-adj = 0.12 p.p., ΔPD15-adj = -0.15 HU.

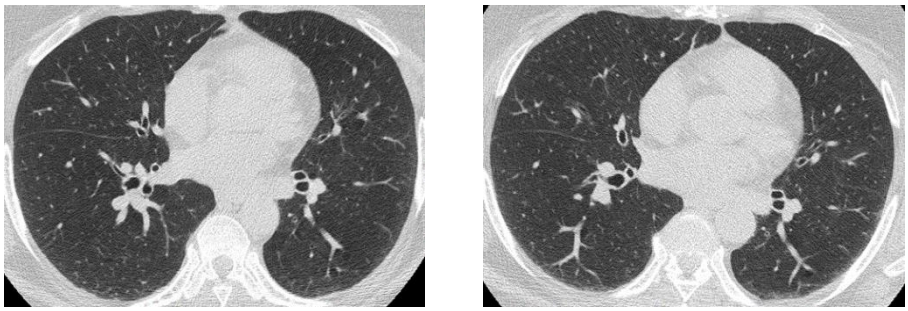

Figure 2.5. Representative paired images from a subject with relatively poorly-reproduced breath-holds. Subject demonstrates ΔV = -0.63 L, ΔRA950-raw = -1.45 p.p., ΔPD15-raw = 10 HU, ΔRA950-corr = 3.49 p.p., ΔPD15-corr = 2.69 HU

The PD15 measure is inherently more sensitive to changes in breath-hold than RA950. This is evidenced by the relatively poorer CCC demonstrated by PD15-raw in the ≥0.25-L subgroup (Table 2.2). Furthermore, when comparing between the ≥0.25-L and <0.25-L subgroups, PD15-raw demonstrated a much larger relative improvement in RC than RA950-raw (Table 2.3), indicating that poorer breath-hold reproduction has a much greater negative impact on the reproducibility of PD15-raw. Lastly, volume correction resulted in a significant improvement in the RC for PD15 but not for RA950 (Table 2.3). These observations indicate that the reproducibility of PD15 is more sensitive to changes in breath-hold.

The linear regression analysis (Fig. 2.3, Table 2.4) reveals a strong relationship between breath-hold reproducibility and density reproducibility, echoing the findings of numerous studies [10, 11]. In particular, the linearity evident in Fig. 2.3a and c is consistent with the intuitive idea that larger differences in breath-holds between two CTs should lead to larger differences in densitometric measures as well, as illustrated in Figs. 2.4 and 2.5. Volume correction results in a flattening of regression lines and a reduction in the coefficients of regression R^2 (Fig. 2.3b, d), indicating that the volume correction model successfully accounts for variation due to differences in volume. However, the effect was more pronounced for PD15 than RA950, a result that is consistent with the findings of Shaker et al., who reported [11] that PD is more robust following volume correction than RA. An alternate interpretation (as discussed in the previous paragraph) is that PD15 is more strongly correlated with breath-hold to begin with and is therefore more responsive to volume correction.

It is essential to evaluate the reproducibility of these densitometric measures in a clinical context as well as a statistical one. In one longitudinal study, Parr et al., followed 71 patients with emphysema associated with α_1-antitrypsin deficiency over a period of 2 years, reporting mean annual progressions of +1.34 p.p., +0.97 p.p., -3.53 HU, and -1.79 HU for RA950-raw, RA950-adj, PD15-raw, and PD15-adj, respectively [23]. These annual rates of change serve as a useful benchmark against which to evaluate the measures considered in our study. The statistical strength of a measure can be represented by its effect size, defined as the ratio of the change to detect (signal) to the variability associated with the measure (noise), with larger effect sizes representing greater statistical strength. Taking the annual rates of change from [23] as the signal and the difference SDs from our study (Table 2.2) as the noise yields an effect size of 1.15 for RA950-raw and 0.67 for PD15-raw, indicating that in the absence of volume correction, RA950 does have better

reproducibility than PD15 from a signal-to-noise standpoint. In contrast, the volume-corrected effect sizes are 1.17 for RA950-adj and 0.98 for PD15-adj. Although volume correction results in an increase in the effect size for both RA950 and PD15, this improvement is much smaller than what is indicated by the RCs alone, a result which illustrates the importance of considering signal-to-noise in variability analysis. By comparison, if the difference SD can be decreased without sacrificing signal (by achieving breath-hold reproduction of <0.25 L, as in our subgroup analysis), then the corresponding effect sizes increase to 1.65 for RA950 and 1.91 for PD15. The preceding discussion relies upon a number of assumptions and is not intended to replace a formal power analysis for a specific study. In practice, the variabilities of these measures would be much larger simply because there are additional sources of variation in a longitudinal study beyond the two CT examinations.

It has often been suggested that volume correction is essential for proper interpretation of densitometric measures [5, 11, 15]. This is motivated by the well-established finding that breath-hold is one of the biggest sources of variation in these measures. Our data show that while breath-hold is indeed a contributor to densitometric measure variation, volume correction may not be the only answer. In particular, volume correction did not lead to a statistically significant improvement in RC for RA950. Furthermore, for both RA950 and PD15, the volume corrected RCs are nearly indistinguishable from the RCs obtained in the <0.25-L subgroup, which may carry the additional advantage of larger effect sizes due to higher signal. Two things seem clear from these observations. First, when it comes to reproducibility of RA950 and PD15, volume correction is no better than ensuring that patient breath-holds are well reproduced to begin with. Second, RA950 is more robust to small variations in breath-hold (up to 1 L) than PD15, which seems to require breath-hold reproduction to within 0.25 L.

There are three important limitations of this study. First, because breath-holds were reproduced so well in this study, our data provide no insight into the behavior of RA950 and PD15 over a wider range of breath-hold differences. Second, the subjects in this study exhibited predominantly mild to moderate emphysema, limiting the generalizability of our findings when considering patients suffering from more severe disease. Lastly, in a longitudinal setting, both breath-hold and densitometric reproducibility depend upon a number of factors that are not addressed in this study. In particular, lung volumes can vary widely due to disease progression, potential treatment effect, and inconsistent patient effort. However, we have previously reported that long-term reproducibility of lung volumes is feasible in multi-center trials [24].

2.5 Conclusion

Our study shows that it is possible to collect patient images across multiple visits and sites with good breath-hold reproducibility that compares favorably with current guidelines for conventional measures of lung volumes. Under these conditions, RA950 showed a high reproducibility of ±2.34 percentage points over the entire set of images. PD15 was relatively less reproducible at ±10.59 HU, but it improved considerably with statistical volume correction or by selecting a subgroup with the most consistent lung volumes. However, volume correction also reduces the magnitude of density change between images, and careful consideration of both signal and noise is necessary during study design.

2.6 References

[1] Gold. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. Executive summary, updated 2009. http://www.goldcopd.org

[2] Hurd S (1984) The impact of COPD on lung health worldwide: epidemiology and incidence. Chest 117(2 Suppl):1S-4S

[3] Hayhurst MD, Flenley DC, McLean A, Wightman AJA, MacNec W, Wright D, et al (1984) Diagnosis of pulmonary emphysema by computerised tomography. Lancet 2:320-322

[4] Muller NL, Staples CA, Miller RR, Abboud RT (1988) Density mask. An objective method to quantitate emphysema using computed tomography. Chest 94:782-787

[5] Dirksen A, Friis M, Olesen KP, Skovgaard LT, Sorensen K (1997) Progress of emphysema in severe α_1-antitrypsin deficiency as assessed by annual CT. Acta Radiol 38:826-832

[6] Goldin JG (2004) Quantitative CT of emphysema and the airways. J Thorac Imaging 19(4):235-240

[7] Spouge D, Mayo JR, Cardoso W, Muller NL (1993) Panacinar emphysema: CT and pathologic findings. J Comput Assist Tomogr 17(5):710-713

[8] Gevenois PA, De Vuyst P, de Maertelaer V, Zanen J, Jacobovitz D, Cosio MG, Yernault JC (1996) Comparison of computed density and microscopic morphometry in pulmonary emphysema. Am J Respir Crit Care Med 154(1):187-192

[9] Pauls S, Gulkin D, Feuerlein S, Muche R, Kruger S, Schmidt SA, Dharaiya E, Brambs HJ, Hetzel M (2010) Assessment of COPD severity by computed tomography: correlation with lung functional testing. Clin Imaging 34:172-178

[10] Robinson PJ, Kreel L (1979) Pulmonary tissue attenuation with computed tomography: comparison of inspiration and expiration scans. J Comput Assist Tomogr 3(6):740-748

[11] Shaker SB, Dirksen A, Laursen LC, Skovgaard LT, Holstein-Rathlou NH (2004) Volume adjustment of lung density by computed tomography scans in patients with emphysema. Acta Radiol 45:417-423

[12] Brown MS, McNitt-Gray MF, Pais R, Shah SK, Qing D, Da Costa I, Aberle DR, Goldin JG (2007) CAD in Clinical Trials: Current role and architectural requirements. J Comput Med Imaging Graph 31:332-337

[13] Brown MS, McNitt-Gray MF, Mankovich NJ, Hiller J, Wilson LS, Goldin JG, Aberle DR (1997) Method for segmenting chest CT image data using an anatomical model: preliminary results. IEEE Trans Med Imaging 16(6):828-839

[14] Gould GA, MacNee W, McLean A, Warren PM, Redpath A, Best JJ, et al (1988) CT measurements of lung density in life can quantitate distal airspace enlargement: an essential defining feature of human emphysema. Am Rev Respir Dis 137:380-392

[15] Stoel BC, Putter H, Bakker ME, Dirksen A, Stockley RA, Piitulainen E, Russi EW, Parr D, Shaker SB, Reiber JH, Stolk J (2008) Volume correction in computed tomography densitometry for follow-up studies on pulmonary emphysema. Proc Am Thorac Soc 5(9):919-924

[16] Dirksen A, Dijkman JH, Madsen F, Stoel B, Hutchison DCS, Ulrik CS, Skovgaard LT, Kok-Jensen A, Rudolphus A, Seersholm N, Vrooman HA, Reiber JH, Hansen NC, Heckscher T, Viskum K, Stolk J (1999) A randomized clinical trial of alpha 1-antitrypsin augmentation therapy. Am J Respir Crit Care Med 160:1468-1472

[17] Parr DG, Stoel BC, Stolk J, Stockley RA (2006) Validation of computed tomographic lung densitometry for monitoring emphysema in alpha 1-antitrypsin deficiency. Thorax 61(6):485-490

[18] Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1):255-268

[19] Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476):307-310

[20] Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright P, van der Grinten CPM, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navajas D, Pedersen OF, Pellegrino R, Viegi G, Wanger J (2005) Standardization of spirometry. Eur Respir J 26(2):319-338

[21] Sciurba FC, Ernst A, Herth FJ, Strange C, Criner GJ, Marquette CH, Kovitz KL, Chiacchierini RP, Goldin J, McLennan G, for the VENT Study Group (2010)A Randomized study of endobronchial valves for advanced emphysema. N Engl J Med 363:1233-1244

[22] Coxson HO, Nasute Fauerbach PV, Storness-Bliss C, Muller NL, Cogswell S, Dillard DH, Finger CL, Springmeyer SC (2008) Computed tomography assessment of lung volume changes after bronchial valve treatment. Eur Respir J 32(6):1443-1450

[23] Parr DG, Sevenoaks M, Deng CQ, Stoel BC, Stockley RA (2008) Detection of emphysema progression in alpha 1-antytrpsin deficiency using CT densitometry; Methodological advances. Respir Res 9:21

[24] Brown MS, Kim HJ, Abtin F, Da Costa I, Pais R, Ahmad S, Angel E, Ni C, Kleerup EC, Gjertson DW, McNitt-Gray MF, Goldin JG (2010) Reproducibility of lung and lobar volume measurements using computed tomography. Acad Radiol 17(3):316-322

3. Reproducibility of breathhold and densitometric measures of emphysema in repeat thoracic computed tomography examinations in the setting of a multicenter clinical trial

3.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a significant cause of morbidity and mortality affecting millions of individuals worldwide. COPD is often associated with emphysema, which is characterized by destruction of lung parenchymal tissue which has long been known to manifest in computed tomography (CT) scans of the lungs as areas of low attenuation [1]. CT densitometric assessment of emphysema has been shown to correlate well with spirometry and pathology [2-6], and densitometric measures such as relative area and percentile density have been proposed as efficacy endpoints in emphysema clinical trials.

In the context of emphysema treatment trials, where investigators are interested in measuring disease progression and treatment efficacy over time, the reproducibility of CT densitometric measures is of paramount importance. One of the most important factors influencing the measurement of CT density is the lung volume at which it is measured; as such, breathhold variation due to inconsistent patient effort is known to be a major factor influencing density reproducibility [3, 7-8]. Additional factors that influence CT density measures include CT technical factors, other patient-related variability, and technologist competence. Standardization of these factors over multiple sites and timepoints is required if CT lung density is to be useful as an outcome measure in clinical trials or to assess change over time in epidemiologic studies.

The purpose of this study was to assess the short-term reproducibility of breathhold and commonly-used CT densitometric measures of emphysema in a scan-rescan setting within a

strictly controlled multicenter clinical trial. Additionally, the variation of reproducibility characteristics across sites was examined.

3.2 Materials and methods

3.2.1 Subjects

The study population consisted of adult subjects with an established clinical diagnosis of COPD who gave signed consent with the approval of a local institutional review board to send their data to a central imaging core. All imaging data was anonymized and used in keeping with the Health Insurance Portability and Accountability Act. A total of 93 subjects (35 male, 39 female, 19 unknown) were collected for the study. The subjects' ages ranged from 42 to 71 years, with a median age of 57 years. These subjects were imaged at eleven different sites, with a minimum of 1 and a maximum of 17 subjects per site. 43 of these subjects have been reported in a previous publication [9].

3.2.2 CT image acquisition

All CTs were performed under the auspices of a central imaging core that provided training and standardized guidelines for image acquisition. Subjects received CT examinations at two timepoints with a one week interval (median 7 days, S.D. 2.4 days). All 93 subjects were imaged at full inspiration (total lung capacity, or TLC) at both timepoints; corresponding full expiration (residual volume, or RV) scans were obtained for a subset of 85 subjects. Subjects were instructed on proper breathholding technique by CT imaging technologists at each site and were given the opportunity to practice prior to imaging.

A variety of CT systems were used from General Electric (GE Healthcare, Little Chalfont, Buckinghamshire, UK), Toshiba (Toshiba America Medical Systems, Tustin, CA, USA), and

Siemens (Siemens Healthcare, Erlangen, Germany). CT devices were calibrated for air and water, as per site protocols. The image acquisition protocol consisted of an X-ray tube voltage of 120 kVp, with tube currents ranging from 57 to 225 mAs. Volumetric high-resolution computed tomography (HRCT) was performed with slice thicknesses ranging from 0.6 to 1.25 mm, and images were reconstructed using a medium-sharp algorithm of Bone (GE), FC51 (Toshiba), or B45f (Siemens). For each subject, the same CT device, image acquisition protocol, and reconstruction parameters were used for both timepoints. Additionally, a water phantom was imaged (with a large field of view to include background air) within 24 hours of each subject's CT examination to monitor device calibration.

3.2.3 Quantitative CT image analysis

CT image data was analyzed using an in-house software package [10]. Lungs were segmented by an automated region-growing algorithm using an initial threshold of -500 Hounsfield Units (HU) followed by manual editing to refine the segmentation. A guided semi-automatic system was used to identify the airway tree, which was then used to exclude gross airway structure from the lung segmentation, resulting in a region of interest (ROI) representing the lungs. The volume of the lungs was computed based on the size of the lung ROI. This CT volume was used as a proxy to represent the TLC and RV breathholds (TLC-V, RV-V).

Next, an image intensity histogram representing the distribution of HU values was extracted from the lung ROI. For TLC images, this histogram was used to compute the relative area below -950 HU (RA950-raw) and the 15^{th} percentile density (PD15-raw). For RV images, the relative area below -860 HU (RA860-raw) was computed in a similar fashion. Next, for TLC only, RA and PD values were adjusted via a volume correction scheme (RA950-adj, PD15-adj). Briefly,

this method involved fitting raw density measures to a mixed-effects regression model, taking breathhold and timepoint as covariates [11].

3.2.4 Statistical analysis

The reproducibility of each measure (TLC-V, RA950-raw, RA950-adj, PD15-raw, PD15-adj for TLC; RV-V, RA860-raw for RV) was assessed according to the following procedures. First, the concordance correlation coefficient (CCC) was computed to assess the agreement in measured values between the first and second timepoints [12]. Next, a paired t-test was performed to test for a statistically significant difference in means between timepoints. Lastly, Bland-Altman analysis was performed to compute the Limits of Agreement (LoA = $\bar{d} \pm 1.96s$) and Reproducibility Coefficient (RC = $\sqrt{2}\times 1.96s$) for each measure, where $\bar{d}$ and s represent the mean and standard deviation of the differences between timepoints, respectively [13-15]. For both LoA and RC, smaller numerical values indicate a narrower range of differences between timepoints, which in turn is associated with better reproducibility.

The measures RA950-adj and RA950-raw were compared using Brown-Forsythe's robust difference-of-variances test to compare their RC. The RC of PD15-adj and PD15-raw were compared in a similar fashion.

The subjects were divided into two subgroups based on the quality of their TLC breathhold reproducibility to assess the impact of breathholds on the reproducibility of density-based measurements. An *a priori* threshold of 0.25L was chosen to represent superior breathhold reproduction. The first subgroup (superior breathholds) was defined to include those subjects with $|\Delta$TLC-V$| < 0.25$L, while the second subgroup (normal breathholds) was defined as those subjects with $|\Delta$TLC-V$| \geq 0.25$L. The reproducibility analysis described above (CCC, LoA, RC) was performed for the two subgroups, and Brown-Forsythe's test was used to compare the RC of

RA950-raw between the <0.25L and ≥0.25L subgroups. Similarly, the RC of PD15-raw was compared between the <0.25L and ≥0.25L subgroups.

Next, the subjects were stratified according to the site where their image data was collected. Sites with fewer than four subjects were excluded from this portion of the analysis. Analysis of variance (ANOVA) was performed on the remaining sites to test for differences in baseline and reproducibility characteristics between the sites. Next, the median and interquartile range (IQR) were computed for ΔTLC-V, ΔRA950-raw, and ΔPD15-raw. Sites were divided into two subgroups based on their ability to achieve good TLC breathhold reproduction. The first subgroup (excellent breathhold performance) was defined according to the following two criteria: the median of ΔTLC-V not significantly different from zero (according to the Wilcoxon signed-rank test) and the IQR of ΔTLC-V smaller than 0.5 L. The second subgroup (normal breathhold performance) was defined as any site failing either of these criteria. The site threshold of 0.5 L was chosen to correspond to the subject threshold of ±0.25L. The reproducibility characteristics of RA950-raw and PD15-raw were compared between these two site subgroups via a two-sample t-test.

Linear regression analysis was used to build two different regression models for ΔRA950-raw. In the first model, ΔRA950-raw was regressed against ΔTLC-V only. In the second model, ΔTLC-V was nested in site. Two different regression models were evaluated for ΔPD15-raw in a similar fashion.

In order to put our results into a clinically relevant context, it is important to consider our measurements from the perspective of both signal and noise. Guyatt, et al, formalized this concept by defining the responsiveness index of a measurement instrument to be the ratio of clinically important difference to the variability in stable subjects [16]. For this purpose, we defined

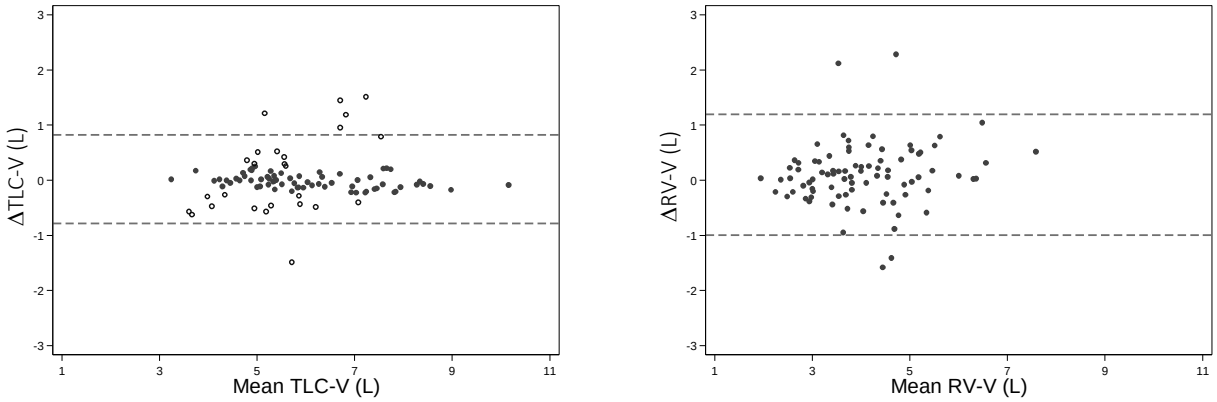

Figure 3.1. Bland-Altman plots illustrating distribution of CT lung volumes between timepoints for TLC (left) and RV (right). For TLC, solid circles indicate superior breathhold reproduction (|ΔTLC-V| < 0.25 L), while open circles represent normal breathhold reproduction (|ΔTLC-V| ≥ 0.25 L). Dashed lines indicate limits of agreement for overall population (TLC *n*=93; RV *n*=85).

variability to be equal to the RC for each of our measures. To establish clinically important difference, we examined a study that followed emphysema patients with α_1-antitrypsin deficiency for two-years, taking their reported mean annual progression rates to be our clinically important difference; these rates were +1.34 p.p., +0.97 p.p, -3.53 HU, and -1.79 HU for RA950-raw, RA950-adj, PD15-raw, and PD15-adj, respectively [17].

All data were analyzed using Microsoft Excel (Microsoft; Redmond, WA) and Stata (StataCorp; College Station, TX). A P value of 0.05 was considered to be significant for all statistical tests.

3.3 Results

3.3.1 CT volume reproducibility analysis

The baseline characteristics of all of the subjects are summarized in Table 3.1, and reproducibility characteristics are summarized in Table 3.2. TLC lung volumes demonstrated very good agreement between the two timepoints, with CCC=0.954 and an RC of 1.14 L. RV lung volume

Table 3.1. Subject baseline characteristics

Measure	Breathhold	Overall (*n*=93 TLC; *n*=85 RV) mean ± SD	<0.25L (*n* = 67) mean ± SD	≥0.25L (*n* = 26) mean ± SD
TLC-V (L)	TLC	5.92 ± 1.35	6.11 ± 1.44	5.44 ± 0.95
RA950-raw (%)	TLC	17.5 ± 10.8	19.0 ± 11.2	13.9 ± 8.7
RA950-adj (%)	TLC	17.7 ± 11.3	---	---
PD15-raw (HU)	TLC	-951.7 ± 26.3	-954.7 ± 26.8	-944.1 ± 24.0
PD15-adj (HU)	TLC	-954.1 ± 22.7	---	---
RV-V (L)	RV	4.00 ± 1.12	---	---
RA860-raw (%)	RV	36.8 ± 18.6	---	---

Table 3.2. Quantitative measure reproducibility characteristics

Measure	*N*	CCC[a]	LoA[b]	RC[c]	P value[d]
TLC-V (L)	93	0.954	(-0.78 0.82)	1.14	0.714
RA950-raw (%)					
Overall	93	0.985	(-3.57, 3.80)	5.21	0.591
<0.25L	67	0.996	(-1.90, 1.94)	2.72	0.900
≥0.25L	26	0.931	(-5.97, 6.65)	8.93	0.598
RA950-adj (%)	93	0.993	(-2.56, 2.78)	3.77	0.441
PD15-raw (HU)					
Overall	93	0.975	(-11.67, 11.23)	16.19	0.724
<0.25L	67	0.995	(-5.46, 4.96)	7.37	0.438
≥0.25L	26	0.911	(-20.39, 20.15)	28.67	0.955
PD15-adj (HU)	93	0.992	(-5.75, 5.35)	7.84	0.495
RV-V (L)	85	0.877	(-1.00, 1.20)	1.55	0.107
RA860-raw (HU)	85	0.922	(-13.48, 15.64)	20.60	0.185

[a] Concordance correlation coefficient. [b] Bland-Altman limits of agreement. [c] Reproducibility coefficient. [d] Paired t-test for difference of means between timepoints.

Table 3.3. Comparisons of Bland-Altman Reproducibility Coefficients for TLC density

	Comparison of Bland-Altman RC[c]			P value[d]
All subjects[a]	RA950-raw 5.21%	vs	RA950-adj 3.77%	0.077
All subjects[a]	PD15-raw 16.19 HU	vs	PD15-adj 7.84 HU	<0.001*
RA950-raw[b]	≥0.25L 8.93%	vs	<0.25L 2.72%	<0.001*
PD15-raw[b]	≥0.25L 28.67 HU	vs	<0.25L 7.37 HU	<0.001*

[a] Comparison between raw and volume-corrected measures on all subjects (*n*=93). [b] Comparison of raw measures between normal (|ΔTLC-V| ≥ 0.25 L; *n* = 26) and superior (|ΔTLC-V| < 0.25 L; *n* = 67) breathhold subgroups. [c] Reproducibility coefficient. [d] Brown-Forsythe's robust difference-of-variances test. *Statistically significant at 0.05.

reproducibility was somewhat poorer than TLC, with CCC=0.877 and an RC of 1.55 L. There was no significant difference in TLC-V or RV-V between the two timepoints (P=0.714 for TLC; P=0.107 for RV). The distributions of the CT lung volumes for TLC and RV are illustrated in Figure 3.1.

Of the overall population of 93 subjects, 67 subjects (72%) achieved superior TLC breathhold reproduction ($|\Delta$TLC-V$| < 0.25$ L). Compared to the normal breathhold subgroup ($|\Delta$TLC-V$| \geq 0.25$ L; $n = 26$), the superior breathhold subgroup demonstrated larger TLC breathholds and more severe emphysema according to RA950-raw and PD15-raw (Table 3.1).

3.3.2 CT densitometric reproducibility analysis

The densitometric measures exhibited good statistical reproducibility, with all CCCs exceeding 0.90 (Table 3.2). For all densitometric measures, there was no significant difference between timepoints ($P > 0.400$ for TLC; $P = 0.185$ for RV). The distributions of the densitometric measures are illustrated in Figure 3.2.

For both RA950 and PD15, volume correction resulted in a smaller value of Reproducibility Coefficient (RC) than the raw measure, indicating improved reproducibility (Table 3.3). This comparison was significant for PD15 ($P<0.001$) but not for RA950 (P=0.077). Next, the superior breathhold subgroup (<0.25L) demonstrated a statistically smaller RC than the normal breathhold subgroup (≥0.25L) for both RA950-raw and PD15-raw ($P<0.001$).

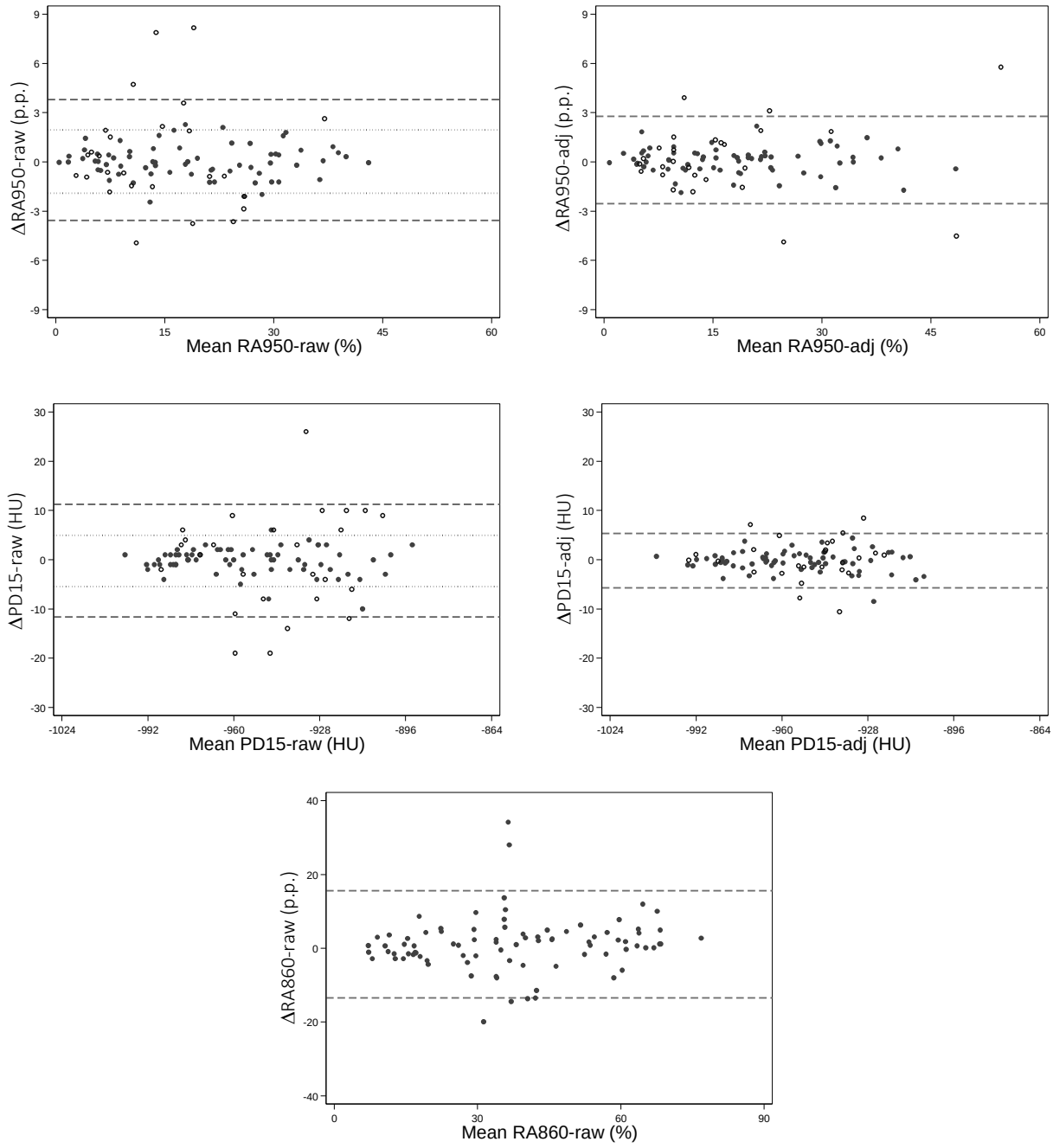

Figure 3.2. Bland-Altman plots illustrating distribution of densitometric measures between timepoints for TLC (top and middle) and RV (bottom). For TLC, solid circles indicate superior breathhold reproduction ($|\Delta$TLC-V$| < 0.25$ L), while open circles represent normal breathhold reproduction ($|\Delta$TLC-V$| \geq 0.25$ L). Horizontal lines indicate limits of agreement for overall population (dashed lines) and superior breathhold subgroup (dotted lines).

Table 3.4. Subject baseline characteristics stratified by site

Site	*n*	TLC-V (L) Mean ± SD	RA950-raw (%) Mean ± SD	PD15-raw (HU) Mean ± SD
A	17	5.49 ± 0.91	12.21 ± 10.46	-938.12 ± 24.97
B	10	6.08 ± 1.19	15.11 ± 11.30	-943.30 ± 23.20
C	14	6.54 ± 2.08	17.88 ± 14.32	-946.93 ± 30.10
D	7	6.49 ± 1.52	14.56 ± 13.67	-940.86 ± 30.60
E	9	5.76 ± 1.07	18.74 ± 11.15	-952.56 ± 23.68
F	8	5.73 ± 1.29	25.22 ± 5.28	-977.25 ± 12.68
G	13	5.91 ± 1.10	20.14 ± 7.99	-962.23 ± 20.90
H	8	6.16 ± 1.13	21.55 ± 7.48	-964.75 ± 17.50
I	5	4.79 ± 0.91	14.19 ± 6.86	-942.20 ± 25.40
Aggregate	91	5.92 ± 1.34	17.44 ± 10.85	-951.13 ± 26.08
P value[a]	---	0.246	0.166	0.007*

[a] ANOVA test for difference in means between sites. *Statistically significant at 0.05.

3.3.3 Site performance analysis

The total study population of 93 subjects was collected across 11 different sites. Of these, 2 sites had fewer than 4 subjects and were excluded from further analysis. The baseline characteristics of the 9 remaining sites are summarized in Table 3.4. There was no difference in baseline TLC-V or RA950-raw between sites, but PD15-raw was significantly different between sites (P=0.007).

The reproducibility characteristics of the 9 sites are summarized in Table 3.5. ΔTLC-V, ΔRA950-raw, and ΔPD15-raw all demonstrated statistically significant difference between sites. Six of the sites (A, B, C, D, G, H) satisfied the criteria for excellent breathhold performance, with Site E failing due to having a median different from 0 (P=0.038) and Sites F and I failing due to having an IQR exceeding 0.5 L. ΔRA950-raw and ΔPD15-raw were significantly different between these two groups of sites (P=0.044 and P=0.033, respectively). The volume and densitometric reproducibilities for each site are illustrated in Figure 3.3.

Table 3.5. Summary of reproducibility stratified by site

Site	*n*	Breathhold performance	ΔTLC-V (L) Median (IQR)	ΔRA950-raw (p.p.) Median (IQR)	ΔPD15-raw (HU) Median (IQR)
A	17	Excellent	0.00 (0.18)	0.02 (0.94)	0.0 (3.0)
B	10	Excellent	-0.07 (0.10)	0.13 (0.63)	0.5 (3.0)
C	14	Excellent	-0.04 (0.33)	-0.04 (0.92)	0.5 (3.0)
D	7	Excellent	0.21 (0.49)	0.81 (1.49)	-6.0 (9.0)
E	9	Normal[a]	-0.13 (0.44)	-0.65 (1.19)	1.0 (5.0)
F	8	Normal[b]	-0.11 (0.54)	-0.90 (4.06)	1.5 (6.0)
G	13	Excellent	0.12 (0.35)	0.21 (2.28)	-1.0 (4.0)
H	8	Excellent	-0.02 (0.21)	-0.11 (3.37)	0.0 (5.5)
I	5	Normal[b]	0.17 (0.79)	-0.90 (3.34)	3.0 (17.0)
Aggregate	91	---	-0.02 (0.30)	-0.02 (1.58)	0.0 (4.0)
P value[c]	---	---	0.028*	0.035*	0.049*

[a] Median ΔTLC-V different from zero. [b] IQR ΔTLC-V greater than 0.5 L. [c] ANOVA test for difference in means between sites.

*Statistically significant at 0.05.

Table 3.6. Regression models for densitometric reproducibility

Measure	Covariates	R^2	R^2_{adj}
ΔRA950-raw	ΔTLC-V	0.6839	0.6804
ΔRA950-raw	ΔTLC-V nested in site	0.7602	0.7335
ΔPD15-raw	ΔTLC-V	0.7329	0.7299
ΔPD15-raw	ΔTLC-V nested in site	0.7671	0.7412

Table 3.6 describes the linear regression models for ΔRA950-raw. Compared with using ΔTLC-V alone, nesting ΔTLC-V in site resulted in a stronger model, as reflected by the adjusted coefficient of determination R^2_{adj}. A similar effect was observed for ΔPD15-raw.

3.4 Discussion

In this study, we compared HRCT scans performed at two different timepoints using the same imaging platforms. There are numerous sources of variation that influence quantitative HRCT reproducibility, including (but not limited to) CT acquisition and reconstruction parameters, X-ray and electronic noise, patient effort with respect to breathhold and motion, as well as patient anatomy and physiology. The strength of our study was our ability to control many of these factors, allowing us to focus more purely on the measurement variation inherent in the patient-CT imaging chain.

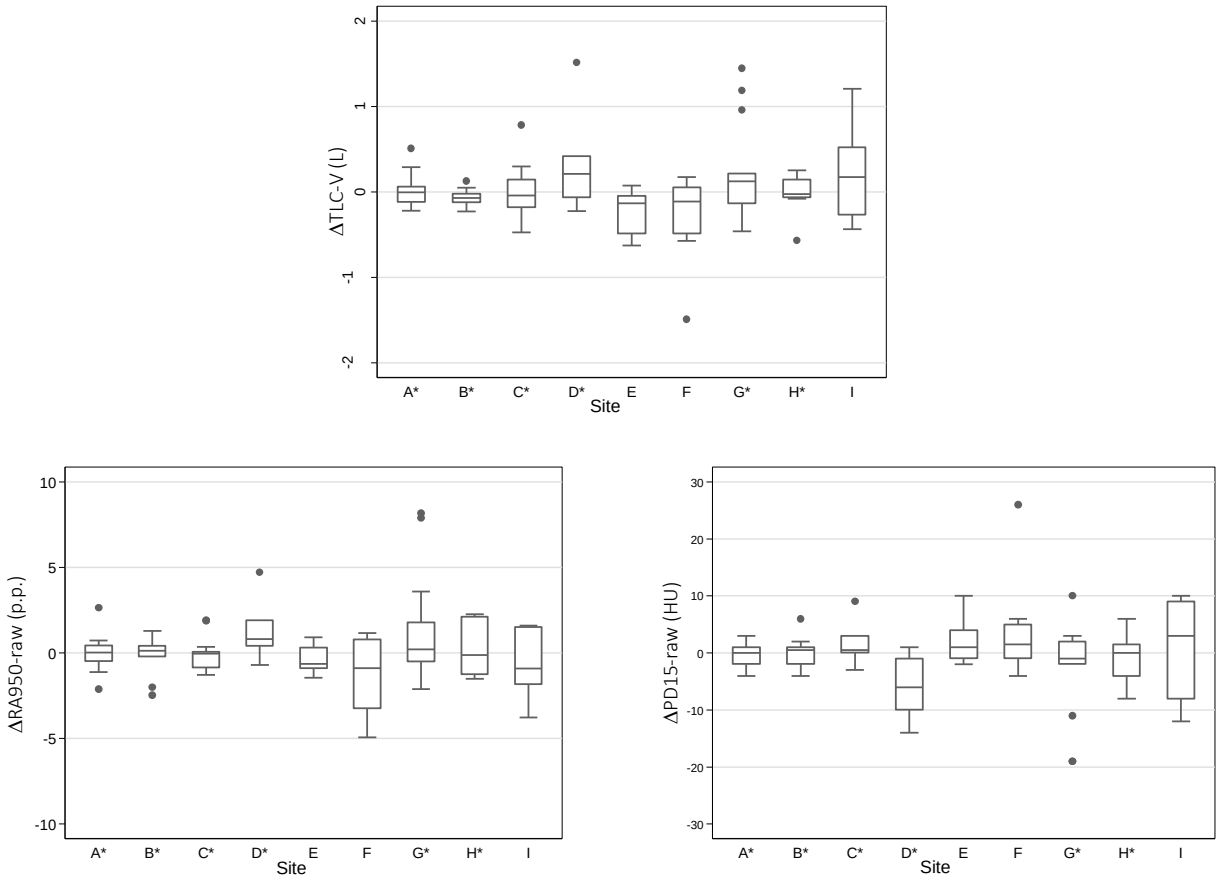

Figure 3.3. Box plots illustrating volume and density reproducibility, stratified by site. *Denotes sites with excellent breathhold performance (ΔV median not different from 0, IQR < 0.5L).

This study is an extension of a previously published study on breathhold and CT densitometric reproducibility in a multicenter emphysema treatment clinical trial [9]. Compared to our previous work, we have made two new key contributions. First, we more than doubled the size of our cohort (from 44 to 93 subjects), making this one of the largest short-term HRCT reproducibility datasets for emphysema in the literature. Second, we assessed the reproducibility of breathhold and CT density at the level of individual sites, providing insight as to how reproducibility characteristics can vary from site to site in the context of a multicenter clinical trial.

Our data reveals good breathhold reproducibility between the two timepoints. TLC breathholds in particular demonstrated excellent statistical and clinical reproducibility. RV

breathholds were somewhat less reproducible, as is the case for conventional measures of RV lung function. It appears that with careful coaching of subjects, good breathhold reproducibility is achievable in CT even without the use of spirometric gating, echoing the findings of Gierada et al [18].

The dependence of CT lung density on breathhold has been described by numerous authors [3, 7-8]. We have shown that RA950 and PD15 reproducibility can be dramatically improved either with well-reproduced breathholds (<0.25L) or through statistical volume-correction methods, a finding that is consistent with our previously reported results [9]. It should be noted that while volume correction can successfully improve scan-rescan reproducibility, this approach may diminish true signal and reduce effect size in a longitudinal setting because differences in breathhold may reflect physiological changes in the lungs rather than variation due to inconsistent patient effort [9].

Based on the mean annual progression reported by Parr et al [17] and the RC from our study (Table 3.2), we were able to evaluate the responsiveness of the densitometric measures. In the overall population, RA950 demonstrated a responsiveness index of 0.26 for both RA950-raw and RA950-adj. Similarly, PD15 demonstrated responsiveness indices of 0.22 for PD15-raw and 0.23 for PD15-adj. These results indicate that while volume correction does nominally improve the reproducibility as measured by RC, it does little to improve their responsiveness. By contrast, in the superior breathhold reproduction subgroup (<0.25L), the responsiveness indices of RA950-raw and PD15-raw rose dramatically to 0.49 and 0.45, respectively.

It is clear that reproducibility characteristics can vary widely between sites in a multicenter clinical study. The interquartile range of ΔTLC-V varied from a minimum of 0.10 L to a maximum of 0.79 L, and one out of the 9 sites demonstrated a statistically significant bias in TLC-V between

timepoints (Table 3.5). These findings could easily be missed if the data were considered in aggregate form only. Nevertheless, 6 out of the 9 sites (67%) achieved excellent breathhold performance, suggesting a role for targeted retraining of specific sites that fail performance criteria. Not surprisingly, breathhold performance at each site is associated with density reproduction as well; in the linear regression models, breathhold alone (irrespective of site) accounted for 68% and 73% of the variation in RA950-raw and PD15-raw, respectively, while including sites as covariates improved the R^2_{adj} to 73% and 74%, respectively (Table 3.6).

There are a number of factors that may contribute to the high variability in reproducibility characteristics across the sites. Numerous authors have emphasized the importance of good coaching to encourage patients to achieve desired breathholds [18-19], and in a multicenter study, it is possible that the quality of breathhold coaching differs from site to site. Additionally, we found statistically significant differences in baseline PD15-raw across sites (Table 3.4), and this difference in disease severity may be a contributing factor as well. Further investigation is needed with a larger number of sites and detailed site performance and compliance metrics.

There are two primary limitations to this study. First, the cohort consists mostly of mild to moderate emphysema, making it difficult to generalize our results to patients with more severe disease. Second, because we deliberately restricted the scope of this investigation to short term reproducibility with tightly controlled parameters, our results are likely to be overly optimistic in a longitudinal setting. Nevertheless, knowledge of a "best-case" scenario for reproducibility can provide valuable insight for understanding and designing multicenter clinical trials. Furthermore, we have previously reported the feasibility of long-term reproducibility of breathholds in multicenter trials [19].

3.5 Conclusion

We have demonstrated that it is possible to image a large number of patients at multiple timepoints across many sites while maintaining good breathhold reproducibility. Under these conditions, our data shows that RA950-raw is highly reproducible at with a reproducibility coefficient of 5.21 percentage points, while PD15-raw is somewhat less reproducible with a reproducibility coefficient of 16.19 HU. The reproducibility of both of these measures improves considerably in a subset of subjects with superior breathhold reproduction, or when statistical volume correction is applied, although volume correction carries the drawback of potentially diminishing true signal and responsiveness. Furthermore, we have demonstrated that reproducibility can vary widely between sites, and we have proposed a framework for evaluating the performance of individual sites. These findings have implications for designing and managing multicenter emphysema clinical trials.

3.6 References

[1] Hayhurst MD, Flenley DC, McLean A, Wightman AJA, MacNec W, Wright D, et al. Diagnosis of pulmonary emphysema by computerized tomography. Lancet 1984; 2:320-322.

[2] Muller NL, Staples CA, Miller RR, Abboud RT. Density mask: An objective method to quantitate emphysema using computed tomography. Chest 1988; 94:782-787.

[3] Goldin JG. Quantitative CT of emphysema and the airways. J Thorac Imaging 2004; 19(4):235-240.

[4] Spouge D, Mayo JR, Cardoso W, Muller NL. Panacinar emphysema: CT and pathologic findings. J Comput Assist Tomogr 1993; 17(5):710-713.

[5] Gevenois PA, De Vuyst P, de Maertelaer V, Zanen J, Jacobovitz D, Cosio MG, Yernault JC. Comparison of computed density and microscopic morphometry in pulmonary emphysema. Am J Respir Crit Car Med 1996; 154(1):187-192.

[6] Pauls S, Gulkin D, Feuerlein S, Muche R, Kruger S, Schmidt SA, Dharaiya E, Brambs HJ, Hetzel M. Assessment of COPD severity by computed tomography: correlation with lung functional testing. Clin Imaging 2010; 34:172-178.

[7] Robinson PJ, Kreel L. Pulmonary tissue attenuation with computed tomography: comparison of inspiration and expiration scans. J Comput Assist Tomogr 1979; 3(6):740-748.

[8] Shaker SB, Dirksen A, Laursen LC, Skovgaard LT, Holstein-Rathlou NH. Volume adjustment of lung density by computed tomography scans in patients with emphysema. Acta Radiol 2004; 45:417-423.

[9] Chong D, Brown MS, Kim HJ, van Rikxoort EM, Guzman L, McNitt-Gray MF, Khatonabadi M, Galperin-Aizenberg M, Coy H, Yang K, Jung Y, Goldin JG. Reproducibility of volume and densitometric measures of emphysema on repeat computed tomography with an interval of 1 week. Eur Radiol 2012; 22(2):287-94.

[10] Brown MS, McNitt-Gray MF, Mankovich NJ, Hiller J, Wilson LS, Goldin JG, Aberle DR. Method for segmenting chest CT image data using an anatomical model: preliminary results. IEEE Trans Med Imaging. 1997; 16(6):828-839.

[11] Dirksen A, Friis M, Olesen KP, Skovgaard LT, Sorensen K. Progress of emphysema in severe alpha 1-antitrypsin deficiency as assessed by annual CT. Acta Radiologica 1997; 38: 826-832.

[12] Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45(1):255-68.

[13] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476):307-10.

[14] Bland JM, Altman DG. Statistics Notes: Measurement error. BMJ 1996; 313:744.1.

[15] Barnhart HX, Lokhnygina Y, Kosinski AS, Haber M. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. J Biopharm Stat 2007; 17(4):721-38.

[16] Guyatt G, Walter S, Normal G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987; 40(2):171-8.

[17] Parr DG, Sevenoaks M, Deng C, Stoel BC, Stockley RA. "Detection of emphysema progression in alpha 1-antitrypsin deficiency using CT densitometry; Methodological advances." Respiratory Research 2008; 9:21.

[18] Gierada DS, Yusen RD, Pilgram TK, Crouch L, Slone RM, Bae KT, Lefrak SS, Cooper JD. Repeatability of quantitative CT indexes of emphysema in patients evaluated for lung volume reduction surgery. Radiology 2001; 220(2):448-54.

[19] Brown MS, Kim HJ, Abtin F, Da Costa I, Pais R, Ahmad S, Angel E, ni C, Kleerup EC, Gjertson DW, McNitt-Gray MF, Goldin JG. Reproducibility of lung and lobar volume measurements using computed tomography. Acad Radiol 2010; 17(3):316-22.

4. Comparison of Multiclass Imbalanced Data Learning Techniques in Classification of Interstitial Lung Disease on CT

4.1 Introduction

Many classification problems suffer from imbalanced data. One challenging problem is computer-aided CT assessment of fibrotic interstitial lung disease (FILD). FILD is highly heterogeneous and subtle to the eye, manifesting in CT through three primary disease patterns: pulmonary fibrosis, ground-glass opacity, and honeycombing. In particular, honeycombing, which represents end-stage irreparable lung disease, can occur very infrequently in some patient populations. Thus, there is a role for imbalanced data learning techniques in developing a computer-aided system for FILD.

Imbalanced data learning (IDL) techniques tend to fall under two broad categories: weighting and resampling [1]. In the first, training examples are assigned differing weights according to their class, while in the second, undersampling or oversampling is employed to balance minority and majority classes. While these techniques have been shown to be effective in addressing the class imbalance problem, the majority of published studies on IDL focus on binary classification, and relatively little attention has been paid to the problem in the setting of multiclass classification [2].

The purpose of this study is twofold: first, to examine the impact of class imbalance on the performance of a support vector classifier on a multiclass FILD dataset; and second, to evaluate and compare the effectiveness of several different IDL techniques applied to this problem.

4.2 Materials and methods

4.2.1 CT Imaging Data

The study population consisted of adult subjects with an established clinical diagnosis of either idiopathic pulmonary fibrosis or FILD associated with systemic scleroderma. Imaging data was made available through an anonymized database, with signed consent from subjects and approval of a local institutional review board. A total of 45 subjects (12 male, 6 female, 27 unknown) were included in the study. Imaging data was accessed in compliance with the Health Insurance Portability and Accountability Act.

Volumetric high-resolution CTs were performed under the auspices of a central imaging core that provided training and standardized guidelines for imaging. A variety of imaging devices were used from Siemens Healthcare (Erlangen, Gemany), GE Healthcare (Little Chalfont, UK), Philips Healthcare (Andover, MA, USA), and Toshiba America Medical Systems (Tustin, CA, USA). Images were acquired at full inspiration in the prone position at 120 kVp and a tube current ≥100 mAs, then reconstructed with a slice thickness between 1.0 to 1.25 mm and a medium-sharp kernel.

Two experienced thoracic radiologists provided cubic volumes of interest (VOIs) corresponding to six classes: pulmonary fibrosis (PF), ground-glass opacity (GG), honeycombing (HC), normal lung parenchyma (NL), airways (AIR), and vessels (VES) (see Fig. 4.1). A two-pass independent reading paradigm was followed. First, each reader placed VOIs with the appropriate class labels throughout the lungs. Next, each reader independently reviewed unlabeled copies of the other reader's VOIs and assigned class labels according to their best judgment. Finally, all VOIs with matching labels from both readers were retained, resulting in a total of 1798 VOIs consisting of 564 PF, 272 GG, 42 HC, 272 NL, 294 AIR, and 354 VES.

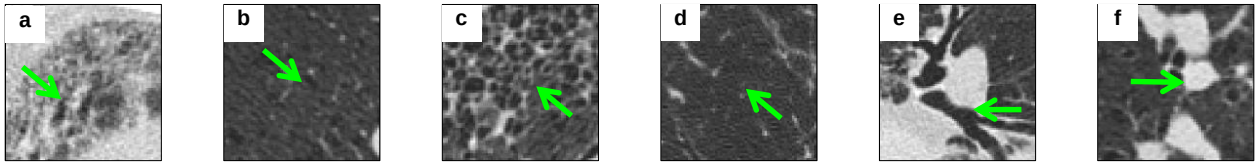

Figure 4.1 Illustration of lung textural and structural classes for classification task. (a) Pulmonary fibrosis (PF); (b) Ground-glass opacity (GG); (c) Honeycombing (HC); (d) Normal lung parenchyma (NL); (e) Airway (AIR); (f) Vessel (VES)

4.2.2 Multiscale Feature Extraction

For each CT image, a Gaussian blurring filter with 0.5 mm radius was applied, followed by isotropic resampling of the image volume at 0.5 mm. Next, for each VOI in the image, a small cubical subimage of size 9 voxels was extracted centered on the VOI. An image intensity histogram was computed on this subimage, and descriptive features were calculated on the histogram. Next, the subimage was adaptively rebinned to 16 graylevels, and gray-level co-occurrence matrix [3] and run-length matrix [4] texture features were extracted from the rebinned subimage. This process was repeated for subimage sizes 9, 11, and 13 and Gaussian radii 0.5, 1.0, 2.0, 4.0 mm for a total of twelve combinations of scalespace parameters, resulting in 792 features.

4.2.3 Support Vector Machine Classification Pipeline

The dataset of labeled VOIs was used to train and evaluate a multiclass support vector machine (SVM) classifier using a radial basis function. First, all feature values were rescaled to zero mean and unit variance. Five-fold stratified partitioning was applied to create five overlapping training folds and corresponding testing folds. An optional data resampling step was performed on each training fold. Next, feature selection was performed on each training fold using the SVM recursive feature elimination (SVMRFE) method to obtain the top 50 features [5]. Next, the optimal SVM parameter C and radial basis function parameter γ were selected for each training fold via gridsearch. Finally, an SVM classifier was trained on each training fold and evaluated against its

corresponding testing fold, and the confusion matrices of the five folds were aggregated to obtain the overall confusion matrix.

From this confusion matrix, classification performance was evaluated via the extended g-mean (EGM) evaluation measure, which is an extension of the g-mean measure to a multiclass setting [6]. Let k denote the total number of classes and REC_i be the recall of the ith class. Then, the extended g-mean is defined as:

$$EGM = \left(\prod_{i=1}^{k} REC_i \right)^{1/k} \tag{4.1}$$

Furthermore, since classification performance with respect to the minority class is a priority in this study, we also report the precision and recall of the honeycombing class ($PREC_{HC}$, REC_{HC}), which are defined as:

$$REC_i = \frac{TP_i}{TP_i + FN_i} \tag{4.2}$$

$$PREC_i = \frac{TP_i}{TP_i + FP_i} \tag{4.3}$$

where TP_i, FP_i, and FN_i denote the number of true positives, false positives, and false negatives, respectively, with respect to the ith class. We also examined the corresponding F-measure (F_{HC}), which is the harmonic mean of recall and precision:

$$F_i = \frac{2REC_i \times PREC_i}{REC_i + PREC_i} \tag{4.4}$$

4.2.4 Class Imbalance Approaches

Numerous IDL techniques have been proposed in the literature. In this study, we evaluated four different approaches: Weighted Costs (WC), Synthetic Minority Oversampling TEchnique

(SMOTE), SMOTE Different Costs (SDC), and Granular SVM - Repetitive Undersampling (GSVM-RU). Additionally, we examined a Naïve approach which used no IDL techniques.

In the Naïve approach, we simply trained the classifier as-is in order to establish a baseline classification performance with no adjustment for class imbalance.

In the WC approach, different costs are assigned to each class's support vectors in order to balance false negatives and false positives [7-8]. We set weights to be inversely proportional to the one-vs-all ratio for the number of examples of each class.

The SMOTE approach is an oversampling technique in which synthetic minority class examples are generated by taking linear combinations of existing ones [1]. For this study, we oversampled our honeycombing class by a factor of 500%.

The SDC approach combines oversampling via SMOTE with the weighted class costs of the WC method [8]. The advantage of the combined approach is to compensate for the weaknesses of the individual methods: WC alone cannot account for the sparseness of the minority class in the feature space, whereas SMOTE alone cannot bridge the gap between the minority and majority examples.

The GSVM-RU approach is an undersampling technique in which the majority class is undersampled by selecting examples near the decision boundary [9-10]. Briefly, the SVM is trained on the full dataset, and the majority class support vectors are selected. This process is iterated with previously selected examples removed from the dataset on each iteration. Finally, all selected examples are combined to form the final undersampled majority set. For this study, we fixed the number of iterations to two.

4.2.5 Statistical Analysis

Due to random variation introduced in the stratified partitioning, data resampling, and parameter selection steps of the SVM pipeline, we evaluated each IDL approach a total of seven times using different random seeds. We then performed Kruskal-Wallis non-parametric ANOVA to test for a difference in ranks between the five different approaches, focusing on EGM for overall and F_{HC} for minority performance, followed by Least Squared Difference (LSD) for multiple comparisons.

4.3 Results

A representative confusion matrix for the Naïve approach is shown in Table 4.1. Although the overall performance for this model was good (EGM=0.847), the performance of the minority class HC suffered due to class imbalance, with a relatively large number false positives and false negatives. This resulted in minority class performance measures of REC_{HC}=0.619, $PREC_{HC}$=0.703, and F_{HC}=0.658.

For comparison, a representative confusion matrix for the SDC approach is shown in Table 4.2. The performance of this model is given by EGM=0.888, REC_{HC}=0.786, $PREC_{HC}$=0.647, and

Table 4.1. Representative confusion matrix for Naïve approach

	Predicted →	PF	GG	HC	NL	AIR	VES
Truth	PF	529	14	9	0	7	5
	GG	18	216	1	32	5	0
	HC	14	0	26	0	2	0
	NL	0	27	0	242	3	0
	AIR	6	13	1	2	268	4
	VES	1	0	0	0	3	350

Table 4.2 Representative confusion matrix for SDC approach

	Predicted →	PF	GG	HC	NL	AIR	VES
Truth	PF	508	25	16	0	10	5
	GG	10	230	1	26	5	0
	HC	6	3	33	0	0	0
	NL	0	26	0	243	3	0
	AIR	4	9	0	3	273	5
	VES	2	0	1	0	1	350

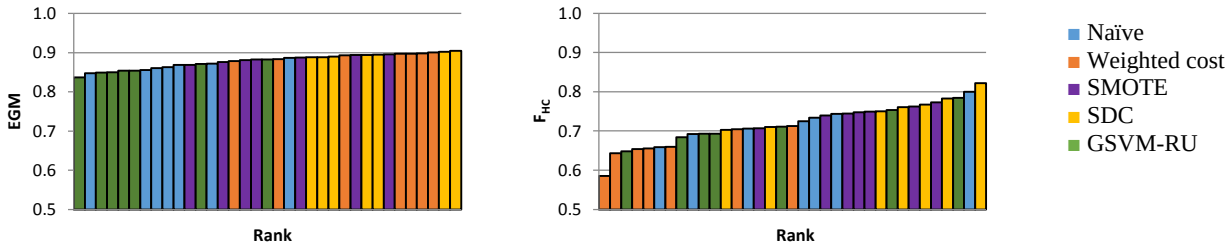

Figure 4.2. Classification performance of imbalanced data learning approaches, sorted by rank.

(Left) Extended G-Mean; (right) Honeycombing F-measure.

Table 4.3. Median classification performance between IDL approaches

Approach	*n*	EGM	REC_{HC}	$PREC_{HC}$	F_{HC}
Naïve	7	0.863	0.690	0.784	0.725
WC	7	0.897	0.857	0.538	0.656
SMOTE	7	0.883	0.762	0.727	0.748
SDC	7	0.895	0.833	0.700	0.761
GSVM-RU	7	0.854	0.643	0.788	0.693

F_{HC}=0.710. Note that while EGM, REC_{HC}, and F_{HC} all increased, there is a corresponding decrease in $PREC_{HC}$, reflecting the increased number of false positives in this model. The other IDL approaches (with the exception of GSVM-RU) exhibited similar behavior.

A summary of the classification performance for the five approaches is illustrated in Table 4.3 and Fig. 4.2. The EGM of the five approaches were significantly different (P<0.001). Additional post-hoc comparisons are described in Table 4.4. In summary, WC + SDC outperformed SMOTE, which outperformed Naïve + GSVM-RU.

Similarly, the F_{HC} of the five approaches were significantly different (P=0.006). Additional post-hoc comparisons are described in Table 4.5. In summary, SMOTE + SDC outperformed GSVM-RU + Naïve, which outperformed WC.

Table 4.4. Comparisons of extended g-mean between IDL approaches

EGM comparison			P value
ANOVA			<0.001*
Naïve + GSVM-RU	vs	SMOTE + WC + SDC	<0.001*
SMOTE	vs	WC + SDC	0.021*

P value is from Kruskal-Wallis test for ANOVA or Mann-Whitney U test for two samples. *Statistically significant at 0.05.

Table 4.5. Comparisons of honeycombing F-measure between IDL approaches

F_{HC} comparison			P value
ANOVA			0.006*
WC	vs	GSVM-RU + SMOTE + Naïve + SDC	0.002*
GSVM-RU + Naïve	vs	SMOTE + SDC	0.023*

P value is from Kruskal-Wallis test for ANOVA or Mann-Whitney U test for two samples. *Statistically significant at 0.05.

4.4 Discussion

In many machine learning applications such as fraud detection, information security, and medical examinations, it is often the least common class that may be the most interesting. For example, in FILD, honeycombing represents end-stage disease and has significance with regards to patient prognosis. In these applications, the tendency of many classifiers to prioritize majority class performance at the expense of the minority class may result in unacceptable performance. These classifiers may systematically undercall the minority class, leading to a large number of false negatives but comparatively fewer false positives; in other words, low recall but high precision. IDL techniques are therefore necessary in order to achieve balanced performance.

In our study, the Naïve approach, which used no IDL techniques, demonstrated low REC_{HC} and high $PREC_{HC}$ compared to the other approaches, which is consistent with our expectations for class imbalance. With the exception of GSVM-RU, the IDL approaches improved EGM, REC_{HC}, and F_{HC} by sacrificing some $PREC_{HC}$ (Table 4.3). The tendency of Naïve to undercall HC is evident in the confusion matrices for Naïve and SDC (Tables 4.1, 4.2). Similarly, an examination

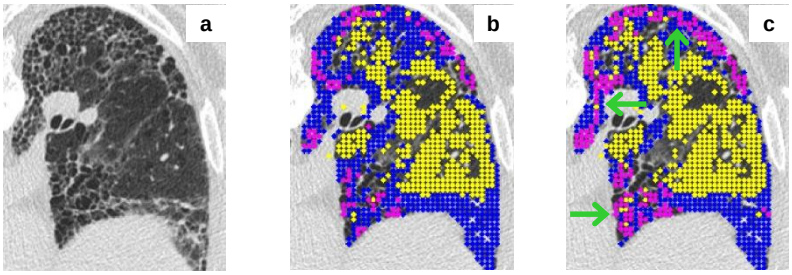

Figure 4.3. Representative images illustrating voxel-wise classification of diseased lung. (a) No overlays; (b) Naïve approach; (c) SDC approach. Only disease classes are shown. Blue denotes pulmonary fibrosis, yellow denotes ground-glass opacity, and magenta denotes honeycombing. In (c), arrows denote additional honeycombing found by SDC that were missed by Naïve.

of CT voxel-wise classification reveals that Naïve is much less likely than SDC to classify a voxel as HC (Fig. 4.3).

Based on the median performances (Table 4.3) and rank sum analysis (Table 4.4), the WC and SDC approaches demonstrated the best overall performance (EGM). However, SDC and SMOTE had the best minority performance (F_{HC}), with WC a distant last (Table 4.5). Given these results, it appears that SDC offers the best balance of overall and minority performance, with SMOTE a close second.

The rank orders of EGM closely mirror REC_{HC}, but not $PREC_{HC}$. This indicates that EGM is inadequate to capture the performance of the classifier with respect to minority false positives. This is because only a relatively small number of false positives are needed to drive down minority class precision, and a majority class can "afford" these classification errors without sacrificing much recall. Ultimately, it is important to keep a close eye on minority false positives when evaluating classification performance. For example, the WC approach achieved very strong overall performance (EGM), but its $PREC_{HC}$ and F_{HC} were unacceptably low (Table 4.3).

It is interesting that GSVM-RU was ineffective at addressing class imbalance. Unlike the other IDL approaches, GSVM-RU's performance is nearly indistinguishable from Naïve. Although each IDL approach considered in this study was initially proposed for binary

classification, GSVM-RU appears to suffer uniquely in a multiclass setting. This is perhaps because it relies on identification of support vectors to drive the undersampling process, and multiclass SVM classification complicates the role of the support vector. The SVM is inherently a binary classifier, and when multiple classes are involved, the problem is typically decomposed into multiple binary classification tasks aggregated via a voting scheme. Each training example may therefore be a support vector in all, some, or none of these binary classifiers, and hence the power of GSVM-RU to sample the immediate neighborhood of the decision boundary is diluted. Further research is needed to adapt GSVM-RU to the multiclass paradigm.

In conclusion, imbalanced data is a significant challenge in many applications including computer-aided CT assessment of fibrotic interstitial lung disease. We have demonstrated that classifier models that explicitly account for data imbalance can outperform those that do not, both in terms of overall and minority class performance. Based on our results, SDC achieves the best balance of overall performance and minority performance. However, false positives remain a concern, and more research is necessary to further refine classification performance.

4.5 References

[1] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. J. Artificial Intelligence Research 2002. 16:321-57.

[2] Wang S, Yao X. Multiclass Imbalance Problems: Analysis and Potential Solutions. IEEE Trans Syst Man Cybern B Cybern 2012. 42:1119-30.

[3] Haralick RM, Bosley R. Texture features for image classification. Third ERTS Symp 1973. pp. 1929-1969.

[4] Tang X. Texture information in run-length matrices. IEEE Transactions on Medical Imaging. 7(11):1602-9.

[5] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002. 46:389-422.

[6] Sun Y, Kamel MS, Wang Y, Boosting for Learning Multiple Classes with Imbalanced Class Distribution. Proc. Int'l Conf. Data Mining 2006. pp. 592-602.

[7] Veropoulos K, Campbell C, Cristianini N. Controlling the Sensitivity of Support Vector Machines. Proceedings of the International Joint Conference on AI 1999. pp. 55-60.

[8] Akbani R, Kwek S, Japkowicz N. Applying Support Vector Machines to Imbalanced Datasets. Lecture Notes in Computer Science 2004. 3201:39-50.

[9] Tang Y, Zhang Y. Granular SVM with repetitive undersampling for highly imbalanced protein homology prediction. Proc of GrC-IEEE 2006. pp. 457-61.

[10] Tang Y, Zhang Y, Chawla N, and Krasser S. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B Cybern 2009. 39(1):281-8.

5. Robustness-driven feature selection in classification of fibrotic interstitial lung disease patterns in computed tomography using 3D texture features

The following chapter is adapted from the manuscript "Robustness-driven feature selection in classification of fibrotic interstitial lung disease patterns in computed tomography using 3D texture features" by D Chong, HJ Kim, P Lo, S Young, MF McNitt-Gray, F Abtin, JG Goldin, MS Brown. The manuscript has been submitted to IEEE Transactions on Medical Imaging for consideration for publication.

5.1 Background

Computer-aided techniques are playing an increasing role in quantitative computed tomography (CT). Compared to traditional manual reader approaches, computer-aided approaches offer numerous advantages such as reducing inter- and intra-reader variability. Unfortunately, one drawback is that these systems frequently suffer from lack of generalizability; an algorithm that performs well in a limited environment (for example, a relatively homogeneous dataset collected in a research setting) often fails to achieve the same level of performance when applied to a wider dataset (for example, all comers across multiple hospitals). This limitation has been a barrier to the widespread adoption of computer-aided diagnosis (CAD) systems in clinical practice.

There are numerous reasons why computer-aided systems may fail to generalize. In computed tomography, one major cause is that quantitative algorithms are often not robust to the choice of technical factors used during image acquisition and reconstruction. For example, in chronic obstructive pulmonary disease (COPD), it has been demonstrated that densitometric emphysema index measures based on CT Hounsfield Units can change significantly when computed under different kernels or slice thicknesses [1-3]. Therefore, in order to produce a truly

generalizable computer-aided system, it is necessary to ensure that the underlying algorithms be robust to variations in image acquisition and reconstruction parameters.

In fibrotic interstitial lung disease (FILD), CT classification approaches have been popular for characterizing lung parenchymal abnormalities. In particular, several studies have successfully applied gray-level co-occurrence matrix and run-length matrix texture features in quantitative 2D and 3D analysis of FILD [4-7]. These and other similar approaches have been shown to be useful for predicting survival and for assessing treatment efficacy in multicenter clinical trials [8, 9].

CT classification of FILD is a challenging task that is made even more difficult by variations across different CT imaging devices or different technical parameters (such as slice thickness, reconstruction kernel, and tube current), which can greatly influence the appearance of textures on the resulting images (Fig. 5.1). One possible solution to this issue is to minimize these variations by prescribing a uniform CT imaging protocol; however, this approach is often impractical, for example in the setting of a large-scale study with multiple scanners across different sites, each with differing levels of compliance to the desired protocol. An alternate solution is to collect a very large number of images into an all-encompassing training set reflecting a wide variety of conditions, which will in principle allow a classifier to be trained in a comprehensive fashion. However, there are numerous technical, logistical, and economic obstacles to collecting such a large and comprehensive dataset, which is why many researchers develop CAD systems in a relatively limited environment. In contrast to these approaches, we propose a solution based on improving classifier robustness with respect to CT technical factors by using a novel feature selection scheme that prioritizes robust features.

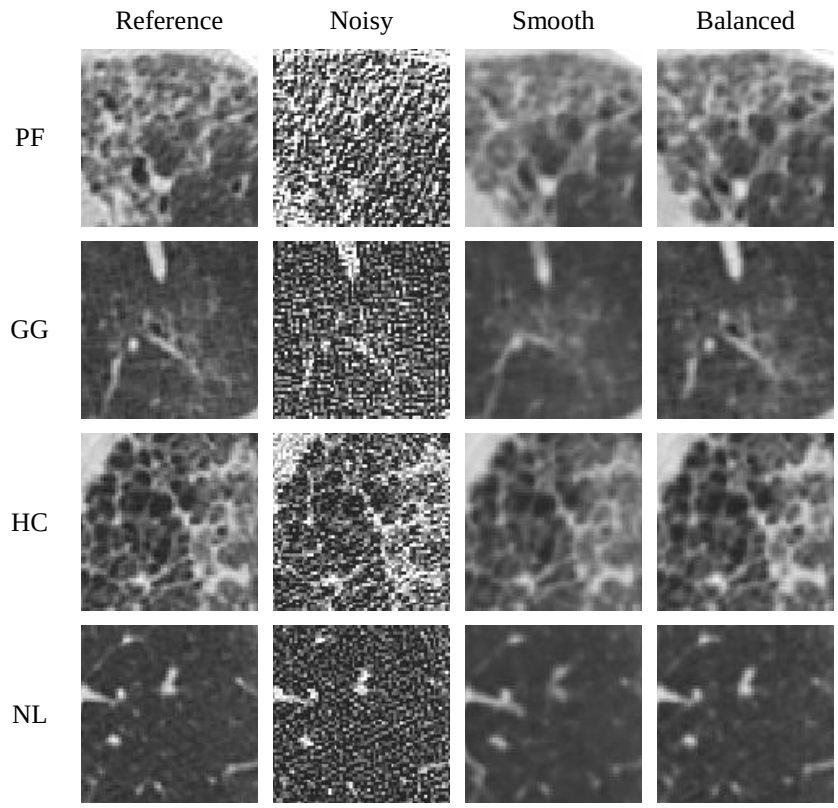

Figure 5.1. Visual illustration of impact of technical factors on texture patterns found in fibrotic interstitial lung disease (FILD). From top to bottom: (PF) pulmonary fibrosis, (GG) ground-glass opacity, (HC) honeycombing, (NL) normal lung parenchyma. Four representative combinations of technical factors are shown: (Reference) 1.0 mm, B45f, Original tube current; (Noisy) 0.6 mm, B70f, 50 mAs tube current; (Smooth) 2.0 mm, B30f, Original tube current; (Balanced) 0.6 mm, B30f, Original tube current.

There has been some prior work in the area of CAD robustness in CT. In 2003, Armato, *et al*, investigated the robustness of a previously published automated CT lung nodule detection system on a dataset consisting of images reconstructed with two different reconstruction kernels, reporting that the performance of the CAD system remained similar regardless of which images were used in training and testing the system [10]. These promising results provide a framework for evaluating the robustness of other similar algorithms; however, because the investigation involved a preexisting CAD system, they provide no insight into how such a system might be designed to be robust in the first place. Furthermore, only two different reconstruction kernels were considered, which is insufficient to capture real-world variations in CT technical factors.

More recently, Balagurunathan, *et al*, working in the area of CT imaging features for non-small cell lung cancer, proposed a multi-step feature selection method to identify highly reproducible, non-redundant features with a large biological range, using repeat-scan CT images in order to assess feature reproducibility [11]. Repeat-scan robustness is different from CT technical factor robustness which is the focus of the present study; nevertheless, their method of using a quantitative robustness measure as a criterion for feature selection is highly relevant, and we will be extending this further by building a classifier model based on the selected features.

This paper presents a novel feature selection algorithm called Robustness-Driven Feature Selection (RDFS), with the objective of producing a CT texture classifier that is robust to variations in slice thickness, reconstruction kernel, and tube current. We evaluate our proposed methodology in the setting of classification of CT disease patterns in fibrotic interstitial lung disease. We hypothesize that applying RDFS will yield a reduced but robust feature subset which will in turn produce a classifier that gives more consistent results when CT technical parameters are varied. Furthermore, we hypothesize that the reduced feature subset will not substantially reduce classifier accuracy on an independent dataset with a narrower range of CT technical parameters.

5.2 Robustness-driven feature selection

We propose a novel feature selection algorithm called Robustness-Driven Feature Selection (RDFS) in order to improve the robustness of classifiers with respect to CT technical factors without requiring a comprehensive training dataset. We define robustness as the ability of a quantitative measurement or algorithm to maintain a stable result when presented with different inputs subject to spurious sources of variation, such as image acquisition or reconstruction parameters. The key assumption underlying RDFS is that robust features produce robust classifiers.

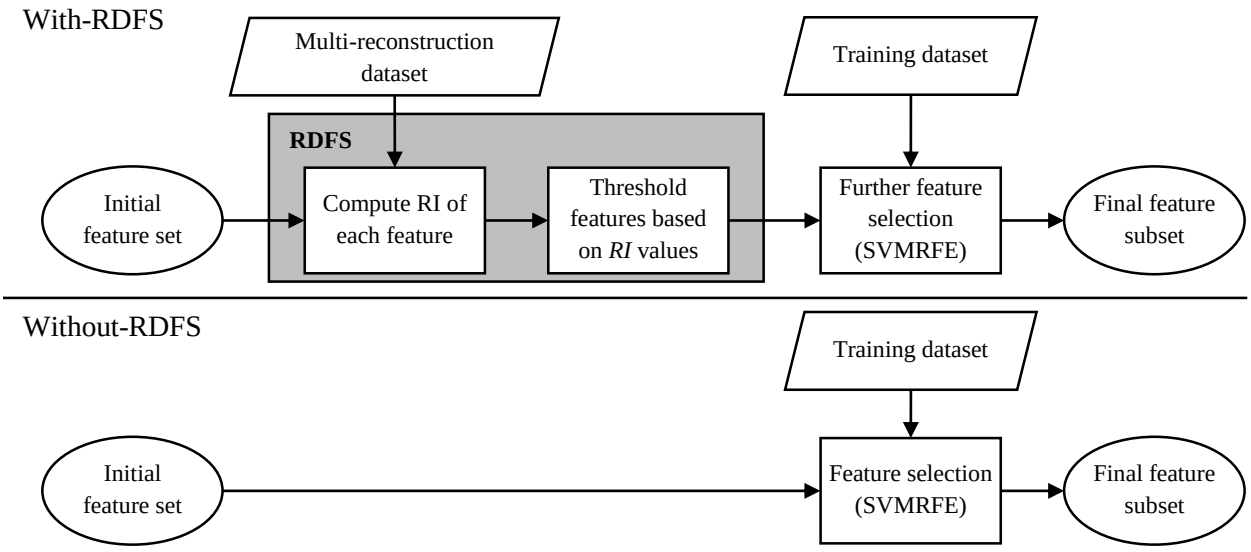

Figure 5.2. Flowchart illustrating the feature selection stage of the classification pipeline for the (top) with-RDFS and (bottom) without-RDFS classifier models. The with-RDFS model employs Robustness-Driven Feature Selection (RDFS), which makes use of a multi-reconstruction dataset consisting of images that have been systematically reconstructed from CT raw sinogram data with a range of slice thicknesses, reconstruction kernels, and simulated reduced tube currents. This dataset is used to compute a quantitative robustness index (RI) for each feature, which measures the variation in feature values due to changing CT technical parameters.

5.2.1 RDFS algorithm

The aim of Robustness-Driven Feature Selection is to produce a feature subset that is robust to CT technical factors. We achieve this objective through the following procedure. First, we assess the robustness of the features that comprise the initial feature set, assigning each feature a quantitative robustness index (*RI*). Next, we prune the feature set by discarding all features that exceed a certain *RI* threshold. This results in a robust feature subset that can then be further processed through other feature selection methods. A summary of this procedure is illustrated in Fig. 5.2.

In order to characterize the robustness of features, we rely on CT raw sinogram data, which allows us to take a single acquisition and reconstruct it using different combinations of technical parameters. By systematically reconstructing raw sinogram data collected across multiple subjects, we create a dataset which we designate as the multi-reconstruction dataset. This dataset is annotated by an expert reader with cubic volumes of interest (VOIs) that are representative of the

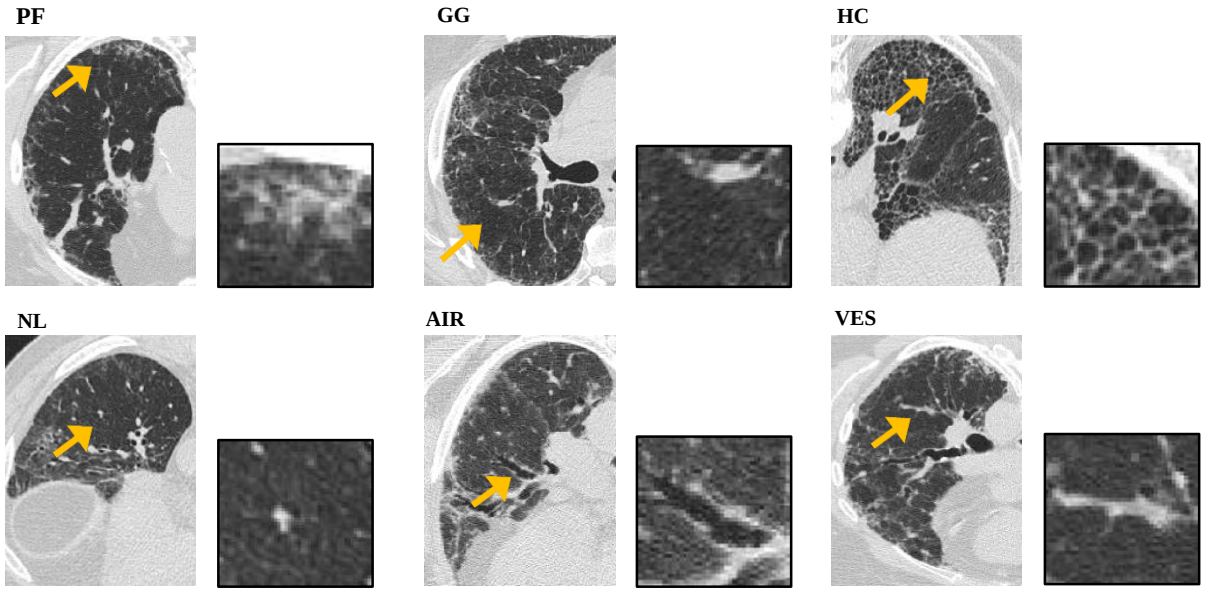

Figure 5.3. Illustration of textural and structural classes for classification of fibrotic interstitial lung disease. From left to right and top to down: (PF) pulmonary fibrosis; (GG) ground-glass opacity; (HC) honeycombing; (NL) normal lung parenchyma; (AIR) airways; (VES) vessels.

different classes that comprise the classification task (summarized in Fig. 5.3 for FILD). Lastly, of the various combinations of technical parameters that comprise this dataset, we designate one reconstruction in particular as the "reference reconstruction".

5.2.2 Feature robustness index

The robustness index (*RI*) of a feature is computed directly from the multi-reconstruction dataset. First, the value of the feature is computed across all VOIs and reconstructions. Let $x_{i,j}$ represent the value of the feature when computed for the *i*th VOI under the *j*th reconstruction, where *j*=0 represents the reference reconstruction. We first compute the standard deviation of feature values within the reference reconstruction only:

$$SD_{within} = \sqrt{\frac{\sum_i (x_{i,0} - \bar{x}_0)^2}{m-1}} \tag{5.1}$$

where $\bar{x}_0$ represents the mean feature value within the reference reconstruction, and *m* = # of VOIs.

Next, the feature values are taken as paired measurement data ($x_{i,0}$, $x_{i,j}$), with the reference reconstruction being the first measurement and the remaining reconstructions comprising the second. Let $d_{i,j} = x_{i,j} - x_{i,0}$ be the paired difference of feature values between the jth reconstruction and the reference reconstruction. We then compute the standard deviation of paired differences:

$$SD_{across} = \sqrt{\frac{\sum_{i}\sum_{j\neq 0}(d_{i,j} - \bar{d})^2}{N-1}} \tag{5.2}$$

where $\bar{d}$ represents the mean difference between measurement pairs and N = (# of VOIs) × (# of reconstructions – 1) is the total number of measurement pairs ($x_{i,0}$, $x_{i,j}$).

The robustness index is then defined as the ratio of the two standard deviations:

$$RI = \frac{SD_{across}}{SD_{within}} \tag{5.3}$$

The robustness index evaluates the robustness of a feature by comparing the variation due to changing technical factors across multiple reconstructions (SD_{across}) against the natural biological variation found within a single reference reconstruction (SD_{within}). A large value of *RI* indicates that technical factor variation is large relative to biological variation, and that the feature is therefore non-robust. Inversely, a small value of *RI* implies that the feature is robust.

The next step is to determine the appropriate *RI* threshold for pruning non-robust features. We determined this threshold experimentally by adopting a two-fold cross-evaluation approach using the multi-reconstruction dataset. RDFS was performed on the first fold using a sliding scale of different *RI* thresholds. The resulting classifier model was evaluated on the second fold (details under section 5.4.2 Evaluation of classifier models, below). Finally, we swapped the role of the two folds and aggregated the classification results between the two folds. The appropriate *RI* threshold was chosen by examining the performance and the robustness of each of these models.

5.3 Materials

5.3.1 CT imaging data

The study population consisted of 99 adult subjects separated into three distinct datasets: training, testing, and multi-reconstruction. The training and testing datasets consisted of 45 and 42 subjects, respectively, with an established clinical diagnosis of either interstitial lung disease associated with systemic scleroderma (SSc) or idiopathic pulmonary fibrosis (IPF). The multi-reconstruction dataset consisted of 12 subjects with diffuse lung disease. All imaging data were anonymized and made available through a central imaging core lab with the approval of a local institutional review board and was accessed in compliance with the Health Insurance Portability and Accountability Act.

For the training and testing datasets, volumetric high-resolution CT images were collected for all subjects. The CTs were performed under the auspices of a central imaging core that provided training and prescribed standardized guidelines for image acquisition and reconstruction. A variety of imaging devices were used from Siemens (Siemens Healthcare, Forchheim, Germany), General Electric (GE Healthcare, Waukesha, WI, USA), Philips (Philips Healthcare, Cleveland, OH, USA), and Toshiba (Toshiba America Medical Systems, Tustin, CA, USA). Images were acquired at full inspiration in the prone position at 120 kVp. For the training dataset, the CT technical parameters were as follows: an average tube current between 100 mAs to 150 mAs; slice thickness between 1.0 mm and 1.25 mm, inclusive; and a medium-sharp reconstruction kernel of B45f, BONE, D, or FC52. For the testing dataset, average tube currents ranged from 50 mAs to 350 mAs; slice thicknesses ranged from 0.625 mm to 2.5 mm; and a wider range of medium-smooth to sharp reconstruction kernels was used: B40f, B45f, B60f, B70f, BONE, B, D, and FC86.

For the multi-reconstruction dataset, CT raw sinogram data were collected for all subjects. The CTs were performed as a part of standard clinical practice using a diffuse lung disease protocol. The subjects were imaged at full inspiration in the supine position at 120 kVp using a multidetector CT device (Definition Flash, Siemens Healthcare, Forchheim, Germany). Tube current modulation was used, resulting in average tube currents ranging from 211 mAs to 328 mAs per subject. The raw sinogram data were used to create additional reconstructions by systematically varying slice thickness and reconstruction kernel. In addition, simulated reduced-tube-current images were generated via synthetic noise, which was added to the CT raw sinogram data using a previously-validated algorithm described in [13-15]. Three slice thicknesses (0.6, 1.0, 2.0 mm), three kernels (B30f-smooth, B45f-medium-sharp, B70f-sharp), and three tube currents (original tube current, 100 mAs, 50 mAs) were used for a total of 27 reconstructions per subject. Of these, the reconstruction corresponding to a slice thickness of 1.0 mm, kernel of B45f, and the original tube current was designated as the "reference reconstruction" since these parameters corresponded most closely with the parameters from the training dataset. Fig. 5.1 gives a representative illustration of some of these reconstructions and their effect on FILD texture classes.

5.3.2 Small volumes of interest for classifier development and assessment

For the training and testing datasets, two experienced thoracic radiologists (JGG, FGA) provided a total of 4088 (2120 training, 1968 testing) cubic volumes of interest (VOIs) corresponding to six visually-based textural and structural classes (Fig. 5.3): pulmonary fibrosis (PF), ground-glass opacity (GG), honeycombing (HC), normal lung parenchyma (NL), airways (AIR), and vessels (VES). The readers followed a two-pass independent reading paradigm. In the first pass, each reader independently placed VOIs throughout the lungs corresponding to each of the above classes, assigning their VOIs with the appropriate class labels. In the second pass, each reader was

independently presented with unlabeled copies of the other reader's VOIs and asked to assign class labels according to their best judgment. At the end of this reading process, each VOI had two labels, one from each reader. Of the total 4088 VOIs, 3443 (1798 training, 1645 testing) VOIs were assigned identical class labels by the two readers, resulting in an overall Cohen's kappa of agreement of 0.805 (0.810 training, 0.800 testing). Training and testing of the classifier was performed on this agreement subset only.

For the multi-reconstruction dataset, VOIs were provided by one experienced thoracic radiologist (JGG), who annotated a total of 238 cubic VOIs corresponding to the same six classes as above (PF, GG, HC, NL, AIR, VES). For each subject, the VOIs were annotated on the reference reconstruction only, then they were propagated to the remaining 26 reconstructions.

In summary, the agreement subset for the training dataset consisted of 1798 VOIs, with individual class counts of 564, 272, 42, 272, 294, and 354 for PF, GG, HC, NL, AIR, and VES, respectively. Similarly, the agreement subset for the testing dataset consisted of 1645 VOIs, with individual class counts of 433, 266, 115, 206, 314, and 311, respectively. Finally, the multi-reconstruction dataset had no agreement subset as it was only annotated by a single radiologist, and it had individual class counts of 75, 37, 23, 27, 43, and 33, respectively.

5.4 Methods

5.4.1 Feature extraction and support vector machine classification

For each CT image, a Gaussian blurring filter with 0.5 mm radius was applied, followed by isotropic resampling of the image volume, using trilinear interpolation to produce a resampled image volume with 0.5 mm × 0.5 mm × 0.5 mm voxels. Next, for each radiologist-provided VOI in the image, a small cubical subimage of size 9 voxels was extracted centered on the VOI. An image intensity histogram was computed on this subimage, and first-order descriptive features

(mean, median, first, and third quartiles) were calculated on the histogram. Next, the subimage was adaptively rebinned to 16 graylevels, and 3D graylevel co-occurrence matrix (GLCM) [16-17] and run-length matrix (RLM) [18-19] texture features were extracted from the rebinned subimage. These GLCM and RLM features were computed in three dimensions across 13 directions with single-voxel spacing, and the mean and range across these directions were retained for each feature. This process was repeated for subimage sizes 9, 11, and 13 voxels and Gaussian radii 0.5, 1.0, 2.0, 4.0 mm for a total of twelve combinations of scalespace parameters, resulting in 792 features in all.

The underrepresentation of honeycombing (HC) examples was identified as a potential limitation in the training dataset (only 42 of 1798 VOIs). In order to account for this problem, we compared several different imbalanced data learning approaches, including weighted SVM costs [20], the synthetic minority oversampling technique (SMOTE) [21], SMOTE with different costs [22], and granular SVM – repetitive undersampling [23]. Of these approaches, the SMOTE with different costs method demonstrated the best performance when evaluated on the training dataset, so this is the method that we adopted for all of our experiments. Briefly, SMOTE generates additional synthetic examples of a minority class by taking linear combinations of existing examples in the feature space. In order to mitigate the impact of class imbalance we boosted the HC examples by 500%, bringing the number of HC examples to 252 (and increasing the total number of training VOIs to 2008).

A support vector machine (SVM) classifier was trained on our data as follows. First, all feature values were standardized to zero mean and unit variance. Next, feature selection was performed through the use of Robustness-Driven Feature Selection (RDFS) as described above. After RDFS, further feature selection was performed using the Support Vector Machine Recursive

Feature Elimination method (SVMRFE) [24] to produce a final feature subset of 50 features (Fig. 5.2). The size of the final feature subset was chosen by evaluating a range of subset sizes from 5 to 50, then selecting the subset size that yielded the best performance on the training dataset. The optimal SVM cost parameter C and radial basis function parameter γ were selected via gridsearch with 5-fold cross validation on the training dataset [25]. Additionally, in accordance with the SMOTE with different costs method, we assigned individual class weights in the SVM model to be inversely proportional to the one-vs-all ratio for the number of training examples of each class as described in [22]. Finally, the SVM classifier model was trained on the training dataset using the selected features and parameters. Classification of multiple classes was performed using the one-against-one method described in [26]. Briefly, separate binary SVM classifiers were constructed for each pairwise combination of classes, and a voting strategy was employed to predict new instances.

5.4.2 Evaluation of classifier models

In many medical applications, sensitivity and specificity are popular measures for assessing classification performance. These measures can be generalized to multiple classes in a fairly straightforward fashion by taking each class in turn as the positive class; however, specificity in particular is inadequate in a multiclass setting because it depends on the total number of true negatives. Indeed, assuming a non-degenerate distribution of classes, it can be shown that as the number of classes increases, the specificity of each class approaches 1 regardless of the actual performance of the classifier.

Instead, we relied on recall and precision as our measures of class performance. Recall, which is identical to sensitivity, is an indication of the false negative performance of a classifier. In a multiclass setting, the recall of any class can be computed by considering that class to be

positive and taking all of the remaining classes as negatives. This class-specific recall is defined as

$$REC_i = \frac{TP_i}{TP_i + FN_i} \tag{5.4}$$

where TP_i and FN_i represent the number of true positives and false negative with respect to class i. Precision, like specificity, is an indication of the false positive performance of a classifier; however, unlike specificity, it depends on the number of true positives rather than true negatives:

$$PREC_i = \frac{TP_i}{TP_i + FP_i} \tag{5.5}$$

where FP_i represents the number of false positives with respect to class i. It is easy to see that recall and precision approach 1 as the number of false negatives and false positives, respectively, approach 0.

Next, we required a measure which would summarize the overall performance of the classifier in a balanced fashion. We adopted the extended g-mean (EGM) measure, which was suggested in [27] as a generalization of the g-mean measure to multiple classes. Let m denote the total number of classes. Then the extended g-mean is defined as

$$EGM = \sqrt[m]{\prod_{i=1}^{m} REC_i} \tag{5.6}$$

Lastly, we required a measure which would reflect classifier robustness, that is, whether classification results changed as a result of the varying technical parameters found in the multi-reconstruction dataset. Since this is essentially a question of agreement, we decided to use Cohen's kappa measure for inter-reader agreement [28]. The first reader was defined to be the classification results with respect to the reference reconstruction, and the classification results with respect to

each of the 26 remaining reconstructions in turn comprised the second reader. Then, Cohen's kappa is given by

$$\kappa = \frac{O - E}{1 - E} \tag{5.7}$$

where O = (# of agreed results / total # of results) is the observed agreement between the first and second reader, and E is the expected agreement between the two readers due to chance. Note that while recall, precision, and extended g-mean all report classifier performance with respect to the ground truth, the kappa measure disregards ground truth and only reflects the robustness of the classifier as compared to the reference reconstruction.

5.4.3 Experimental design

We constructed two different classifier models, one that used RDFS and one that did not. We will refer to these classifier models as "with-RDFS" and "without-RDFS", respectively. Aside from the step of pruning non-robust features via RDFS (the gray box in Fig. 5.2), every other step of the classification pipeline was performed in an identical fashion between the two models.

We performed two different experiments to compare the two classifier models. In the first experiment, the models were evaluated on the multi-reconstruction dataset. The purpose of this experiment was to demonstrate the improved robustness of the with-RDFS model when presented with images from a wide range of technical parameters. The robustness was evaluated by computing the EGM (with respect to the ground truth) for each of the 27 reconstructions, then by computing the kappa (with respect to the reference reconstruction) for each of the 26 non-reference reconstructions. We also examined agreement confusion matrices to assess how classification of VOIs was influenced by varying technical parameters.

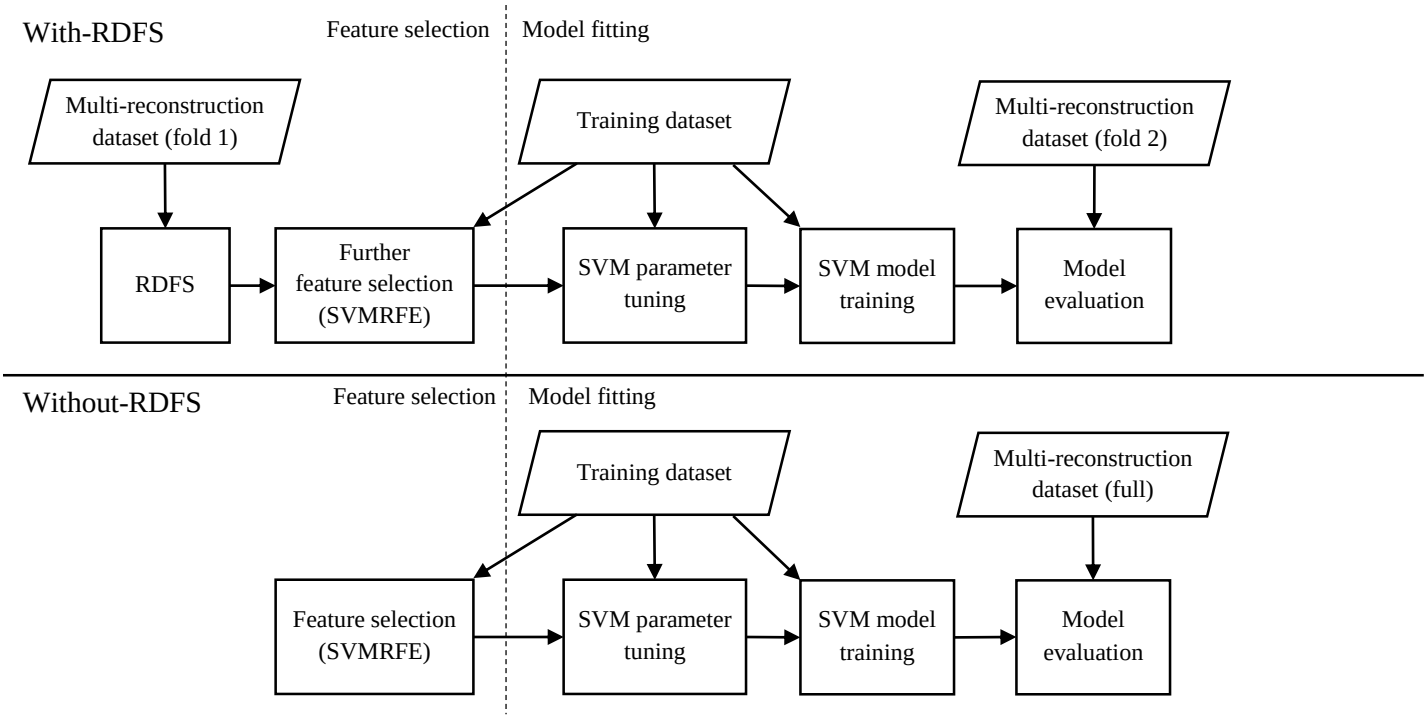

Figure 5.4. Flowchart illustrating classification pipeline for first experiment (evaluation on multi-reconstruction dataset). The steps to the left of the dashed line indicate feature selection, which is depicted in greater detail in Fig. 5.2.

The classification pipeline for the first experiment was as follows. Two-fold cross evaluation was performed for with-RDFS. Briefly, RDFS was applied using one fold of the multi-reconstruction dataset to produce a subset of robust features. Next, further feature selection (via SVMRFE), parameter tuning, and model training were done using the training dataset. Finally, the resulting classifier model was evaluated on the second fold of the multi-reconstruction dataset, then the roles of the two folds were swapped, and classifier output was aggregated between both folds. For without-RDFS, feature selection (via SVMRFE only), parameter tuning, and model training were all done on the training dataset, and the resulting classifier model was evaluated on the multi-reconstruction dataset. This pipeline is illustrated in Fig. 5.4.

In the second experiment, the two classifier models were evaluated on the standalone testing dataset. The purpose of this experiment was to examine whether the reduced but robust feature subset of RDFS resulted in decreased performance on a novel dataset consisting of a

relatively narrow range of technical parameters. The performance was evaluated by computing the recall and precision of the models with respect to each of the classes, and the extended g-mean (EGM) measure was used to summarize overall performance. Additionally, the EGM was computed on a per-subject basis for both models, and EGM scores were compared pairwise across subjects via Westlake's one-sided test for equivalent proportions, using an 80% limit of equivalence and a 0.05 level of significance. Lastly, exact binomial confidence intervals were computed at the 95% level for subject-wise EGM for both with-RDFS and without-RDFS.

The classification pipeline for the second experiment was as follows. For with-RDFS, RDFS was performed by using the entire multi-reconstruction dataset, then further feature selection (via SVMRFE), parameter tuning, and model training were done using the training dataset. The resulting classifier model was evaluated on the standalone testing dataset. For without-RDFS, feature selection (via SVMRFE only), parameter tuning, and model training were all done on the training dataset, and the resulting classifier model was evaluated on the standalone testing dataset.

Support vector classification was performed within the environment of Weka 3.7.11 (Waikato Environment for Knowledge Analysis, The University of Waikato, Hamilton, New Zealand) [29] using the implementation provided in the software package LibSVM (National Taiwan University, Taipei, Taiwan) [30]. Statistical analysis was performed using Microsoft Excel 2013 (Redmond, WA, USA) and Stata (StataCorp, College Station, TX, USA) .

5.5 Results

5.5.1 Characterization of feature robustness

Table 5.1 summarizes the results of the feature robustness analysis, which involved computing the robustness index (*RI*) for each feature by computing the feature across the various reconstructions

that comprise the multi-reconstruction dataset. Neither the training nor the standalone testing datasets were involved in this analysis.

Of the three types of features in our feature space, first-order descriptive features exhibited the smallest *RI*, indicating that they were the most robust with respect to variations in slice thickness, reconstruction kernel, and tube current. Furthermore, there was a trend of decreasing *RI* with increasing subimage window size and especially with increasing Gaussian blurring radius, indicating that features extracted at higher levels of scale were more robust. A heatmap illustrating *RI* values across the entire feature space of 792 features is shown in Fig. 5.5. This diagram represents *RI* values when computed across the entire multi-reconstruction dataset; similar trends were observed for both folds in two-fold cross-evaluation.

Fig. 5.6 shows the results of the experiment to determine an appropriate threshold for robustness-driven feature selection (RDFS). Both classifier performance (according to EGM) and agreement (according to kappa) demonstrated a decreasing trend as the *RI* threshold was increased (allowing less robust features to be included), although EGM exhibited much noisier behavior than kappa. Based on these results for the entire multi-reconstruction dataset, an *RI* threshold of 0.29 was chosen in order to retain as many features as possible before performance and agreement start to decay. This threshold resulted in a robust feature subset of 292 features (37% of the initial feature space). Similar results were obtained for two-fold cross-evaluation.

5.5.2 Evaluation against multi-reconstruction dataset

The results of evaluating the with-RDFS and without-RDFS classifier models on the multi-reconstruction dataset are summarized in Tables 5.2 and 5.3. Table 5.2 lists the classification performance (with respect to ground truth) as measured via the extended g-mean (EGM). On the reference reconstruction, the classifier models achieved a performance of 0.849 for with-RDFS

and 0.778 for without-RDFS. For both models, there was a trend of decreasing EGM when a 2.0 mm slice thickness was used. Furthermore, we observed a relatively larger decrease in EGM for certain combinations of parameters, such as B70f and 50 mAs or 2.0 mm and B30f.

Table 5.3 lists the classification agreement (with respect to the reference reconstruction) as measured via Cohen's kappa. Agreement varied between 0.899 to 0.989 for with-RDFS and 0.827 to 0.968 for without-RDFS. As with the EGM measure, there was a trend of decreasing kappa for the 2.0 mm slice thickness as well as for certain combinations of parameters, such as B70f and 50 mAs or 2.0mm and B30f. The with-RDFS model demonstrated higher kappa than without-RDFS for 24 of the 26 non-reference reconstructions.

Table 5.4 presents classifier agreement confusion matrices for two representative non-reference reconstructions. Note that unlike traditional confusion matrices, the matrices shown here do not reflect classifier accuracy with respect to ground truth. Instead, off-diagonal entries in these matrices indicate VOIs which were classified differently on the indicated reconstruction than they were on the reference reconstruction. It can be seen that these classification disagreements occurred less frequently for the with-RDFS model compared to without-RDFS.

Fig. 5.7 summarizes the major disagreements between reference and non-reference reconstructions for the two classifier models. Each cell in these diagrams corresponds to a specific combination of technical parameters in the multi-reconstruction dataset. The entries within the cells represent off-diagonal entries in the corresponding classifier agreement confusion matrix. Only classification disagreements of three or more are depicted in this diagram. For both the with-RDFS and without-RDFS models, the majority of disagreements occur at a slice thickness of 2.0 mm. However, the with-RDFS model exhibits substantially fewer disagreements overall, indicating that it is more robust to these technical factors than the without-RDFS model.

Table 5.1. Robustness index of features

Features	*RI* mean (SD)
First-order descriptive (n = 48)	0.179 (0.005)
Texture-GLCM (n = 480)	0.387 (0.155)
Texture-RLM (n = 264)	0.361 (0.130)

	Gaussian radius (mm)			
Subimage size (voxels)	0.5	1.0	2.0	4.0
9	0.550 (0.186)	0.423 (0.116)	0.346 (0.079)	0.283 (0.062)
11	0.505 (0.182)	0.366 (0.107)	0.315 (0.069)	0.253 (0.045)
13	0.482 (0.184)	0.339 (0.093)	0.286 (0.060)	0.242 (0.043)

Robustness index (*RI*) summarized by (top) feature category and (bottom) feature scale. For feature scale, each cell represents n = 66 features. Smaller values of *RI* indicate increased robustness.

Table 5.2. Classification performance of SVM models on multi-reconstruction dataset

With-RDFS	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.853	0.843	0.832	0.844	0.848	0.839	0.839	0.834	0.774
1.0 mm	0.845	0.845	0.843	0.849	0.843	0.844	0.849	0.827	0.762
2.0 mm	0.798	0.792	0.792	0.791	0.788	0.804	0.849	0.824	0.821
Without-RDFS	**B30f**			**B45f**			**B70f**		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.770	0.782	0.782	0.793	0.779	0.789	0.798	0.747	0.733
1.0 mm	0.733	0.769	0.773	0.778	0.773	0.771	0.796	0.774	0.719
2.0 mm	0.675	0.675	0.675	0.679	0.682	0.700	0.710	0.693	0.693

Performance is measured via the extended g-mean (EGM) measure. The reference reconstruction is denoted with a gray background.

.**Table 5.3**. Classification robustness of SVM models on multi-reconstruction dataset

With-RDFS	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.974 (0.951, 0.997)	0.963 (0.936, 0.990)	0.958 (0.929, 0.987)	0.979 (0.958, 0.999)	0.968 (0.943, 0.993)	0.984 (0.966, 1.000)	0.921 (0.883, 0.960)	0.952 (0.922, 0.983)	0.899 (0.855, 0.942)
1.0 mm	0.952 (0.922, 0.983)	0.953 (0.922, 0.983)	0.963 (0.936, 0.990)	---	0.979 (0.958, 0.999)	0.989 (0.975, 1.000)	0.947 (0.915, 0.979)	0.947 (0.915, 0.979)	0.909 (0.868, 0.951)
2.0 mm	0.931 (0.895, 0.967)	0.920 (0.881, 0.959)	0.920 (0.881, 0.959)	0.931 (0.895, 0.967)	0.926 (0.888, 0.963)	0.947 (0.915, 0.979)	0.968 (0.943, 0.993)	0.952 (0.922, 0.983)	0.942 (0.908, 0.975)
Without-RDFS	**B30f**			**B45f**			**B70f**		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.968 (0.943, 0.993)	0.968 (0.943, 0.993)	0.963 (0.935, 0.99)	0.942 (0.908, 0.975)	0.936 (0.901, 0.972)	0.905 (0.863, 0.947)	0.880 (0.833, 0.927)	0.894 (0.850, 0.939)	0.872 (0.823, 0.920)
1.0 mm	0.936 (0.900, 0.971)	0.952 (0.921, 0.983)	0.946 (0.914, 0.979)	---	0.968 (0.943, 0.993)	0.952 (0.921, 0.983)	0.905 (0.863, 0.947)	0.926 (0.888, 0.963)	0.876 (0.828, 0.924)
2.0 mm	0.855 (0.804, 0.907)	0.861 (0.810, 0.911)	0.850 (0.797, 0.902)	0.865 (0.815, 0.915)	0.865 (0.815, 0.915)	0.871 (0.822, 0.920)	0.860 (0.810, 0.911)	0.844 (0.790, 0.897)	0.827 (0.771, 0.883)

Robustness is measured via Cohen's kappa by comparing classifier output for each non-reference reconstruction against the reference reconstruction (denoted with a gray background). Numbers in parentheses represent 95% confidence intervals.

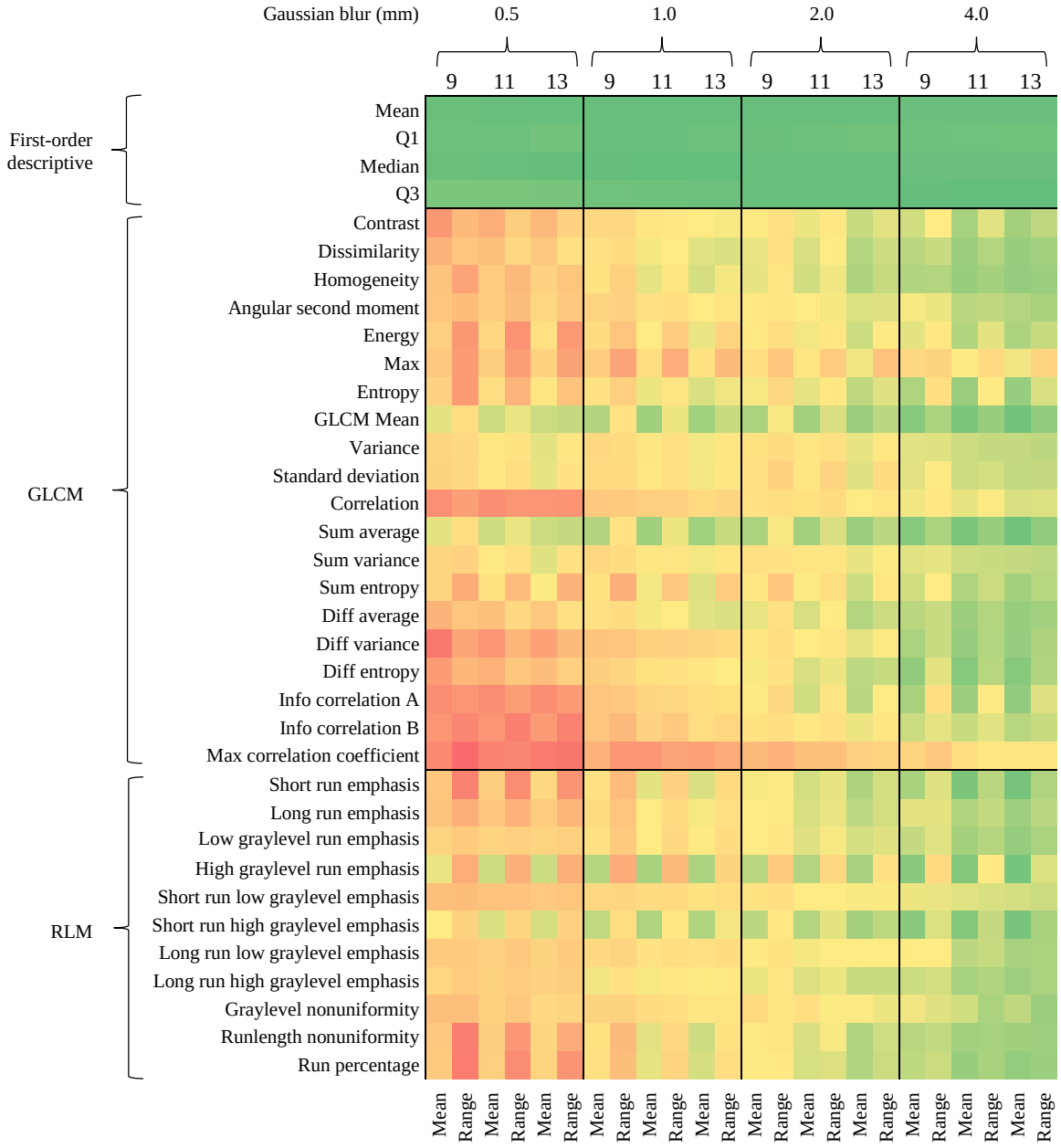

Figure 5.5. Heatmap of robustness index (RI), which measures the variation in feature values due to changing CT technical parameters. The columns represent levels of scale (Gaussian blurring and subimage size) while the rows represent the features that were calculated for each level of scale. The color map covers the range of observed RI values, with green (RI=0.17) indicating very good robustness, red (RI=0.98) indicating very poor robustness, and yellow (RI=0.32) indicating the median value. RI values computed using the multi-reconstruction dataset (n = 12 subjects).

Table 5.4. Confusion matrices for SVM model robustness in multi-reconstruction dataset

← REFERENCE	With-RDFS "Noisy" → (0.6 mm, B70f, 50 mAs tube current)						Without-RDFS "Noisy" → (0.6 mm, B70f, 50 mAs tube current)					
	PF	GG	HC	NL	AIR	VES	PF	GG	HC	NL	AIR	VES
PF	76						73		5		1	
GG	4	28			3		5	33	1		1	
HC	2	1	17				1		10			
NL		4		18	3			4		22		
AIR	2				48		2	2	1	1	44	
VES						32						32

← REFERENCE	With-RDFS "Smooth" → (2.0 mm, B30f, Original tube current)						Without-RDFS "Smooth" → (2.0 mm, B30f, Original tube current)					
	PF	GG	HC	NL	AIR	VES	PF	GG	HC	NL	AIR	VES
PF	75	1					72	7				
GG		35						40				
HC	5		14		1		2	1	8			
NL		2		23				7		15	4	
AIR	1	1		1	46	1	3	3			44	
VES						32						32

Robustness confusion matrices comparing classifier output between reference reconstruction and indicated non-reference reconstructions. Note that unlike traditional confusion matrices, these matrices do not reflect classifier accuracy with respect to ground truth. Instead, rows indicate the class labels assigned by each classifier (with-RDFS or without-RDFS) when evaluated on the reference reconstruction. For example, using the with-RDFS classifier model, 4 VOIs that were classified as NL on the reference reconstruction were instead classified as GG on the "Noisy" reconstruction (top left matrix). Cells with a value of 0 were left blank for readability.

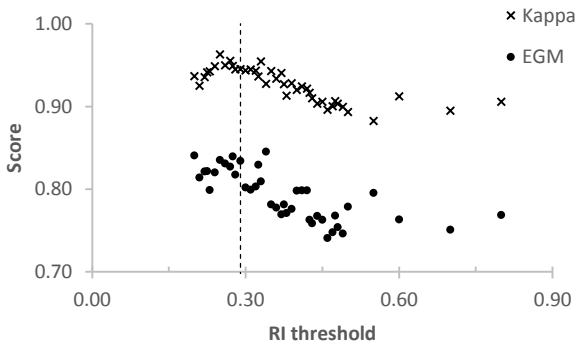

Figure 5.6. Result of experiment for determining appropriate robustness index (*RI*) threshold for Robustness-Driven Feature Selection across entire multi-reconstruction dataset. Extended g-mean (EGM) is measured by assessing classification output on reference reconstruction against ground truth. Kappa is measured by comparing classification output on all non-reference reconstructions against reference reconstruction. Dashed line indicates the selected *RI* threshold of 0.29.

Table 5.5. Confusion matrices for SVM model performance in standalone testing dataset

With-RDFS	CLASSIFIED AS →							
		PF	GG	HC	NL	AIR	VES	REC	PREC
↓ TRUTH	PF	385	9	11		26	2	0.889	0.867
	GG	16	190		44	16		0.714	0.876
	HC	26		56		33		0.487	0.812
	NL		8		194	4		0.942	0.678
	AIR	1	10	2	48	253		0.806	0.757
	VES	16				2	293	0.942	0.993
								EGM	0.778

Without-RDFS	CLASSIFIED AS →							
		PF	GG	HC	NL	AIR	VES	REC	PREC
↓ TRUTH	PF	405	13	4	1	7	3	0.935	0.890
	GG	18	191		50	7		0.718	0.868
	HC	25	3	55	1	29	2	0.478	0.917
	NL		9		186	10	1	0.903	0.655
	AIR	5	4	1	46	257	1	0.818	0.824
	VES	2				2	307	0.987	0.978
								EGM	0.785

Performance confusion matrices indicate agreement between classifier output and ground truth. Rows indicate the ground truth class labels assigned by expert radiologists. Cells with a value of 0 were left blank for readability. REC, PREC, and EGM indicate recall, precision, and extended g-mean, respectively.

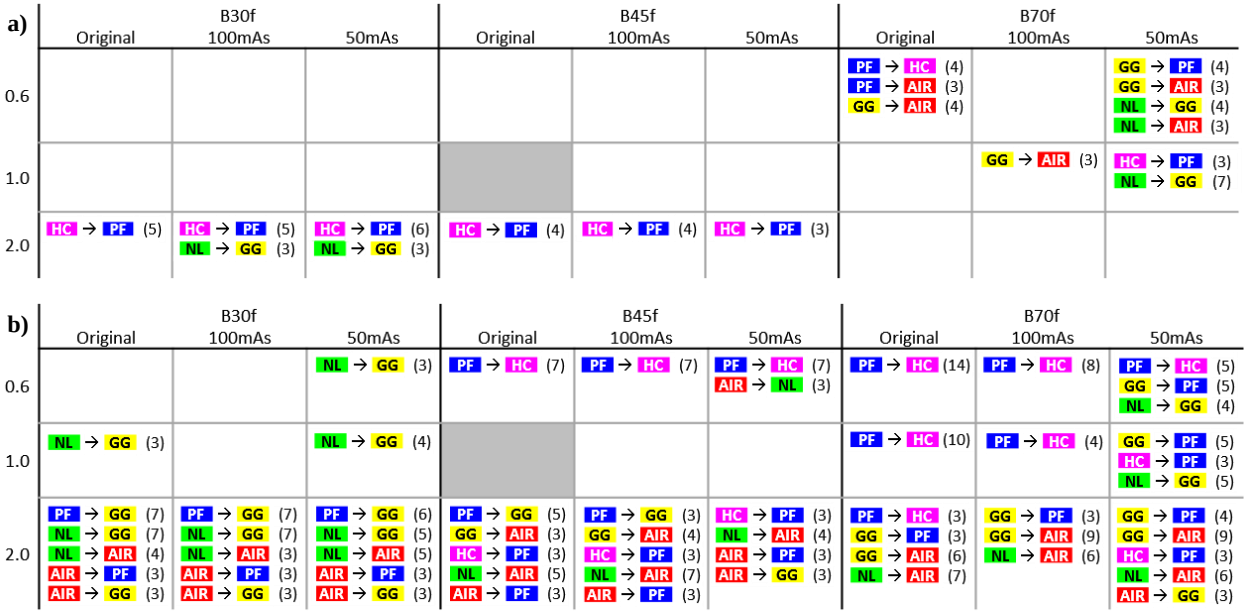

Figure 5.7. Summary of classification disagreements between reference reconstruction (denoted in gray) and each other reconstruction for (a) with-RDFS and (b) without-RDFS classifier models. The numbers in parentheses represent the number of disagreements between indicated classes. For example, the bottom-left-most entry of (a) indicates that five VOIs which were classified as HC on the reference reconstruction were instead classified as PF on the 2.0 mm, B30f, Original tube current reconstruction. Disagreements consisting of fewer than three cases are not listed in the diagram

Table 5.6. Final feature subsets for standalone testing dataset evaluation

With-RDFS	Without-RDFS
Q3_0.5_09	MEAN_0.5_09
Q1_0.5_09	Q1_0.5_09
MEAN_0.5_09	Q3_0.5_09
Q3_0.5_11	Q1_0.5_11
Q1_0.5_11	Q3_0.5_11
Q3_0.5_13	Q1_0.5_13
MEAN_0.5_13	Q3_0.5_13
Q1_0.5_13	Q1_1.0_09
Q1_1.0_09	Q3_1.0_09
Q3_1.0_09	Q3_1.0_11
Q3_1.0_11	Q1_1.0_13
Q3_1.0_13	Q3_1.0_13
MEDIAN_1.0_13	Q3_2.0_13
Q3_2.0_11	MEAN_4.0_09
Q1_2.0_13	Q3_4.0_09
Q3_2.0_13	MEAN_4.0_11
Q3_4.0_09	MEDIAN_4.0_11
Q3_4.0_11	Q3_4.0_11
MEAN_4.0_11	MEAN_4.0_13
Q3_4.0_13	MEDIAN_4.0_13
Q1_4.0_13	Q3_4.0_13
MEAN_4.0_13	MEAN_GLCM_MAX_0.5_09
MEDIAN_4.0_13	MEAN_GLCM_CONTRAST_0.5_11
RANGE_GLCM_SUM_AVERAGE_0.5_13	RANGE_GLCM_INFO_CORRELATION_B_0.5_11
MEAN_GLCM_SUM_VARIANCE_0.5_13	MEAN_GLCM_INFO_CORRELATION_B_0.5_13
RANGE_GLCM_DIFF_AVERAGE_1.0_13	MEAN_GLCM_MAX_0.5_13
MEAN_GLCM_HOMOGENEITY_1.0_13	MEAN_GLCM_SUM_VARIANCE_0.5_13
MEAN_GLCM_ENTROPY_1.0_13	MEAN_GLCM_VARIANCE_0.5_13
RANGE_GLCM_MEAN_1.0_13	RANGE_GLCM_ANGULAR_SECOND_MOMENT_0.5_13
MEAN_GLCM_SUM_ENTROPY_1.0_13	RANGE_GLCM_ENERGY_0.5_13
MEAN_GLCM_DIFF_AVERAGE_1.0_13	RANGE_GLCM_MAX_0.5_13
MEAN_GLCM_DISSIMILARITY_1.0_13	RANGE_GLCM_SUM_ENTROPY_0.5_13
MEAN_GLCM_DIFF_AVERAGE_2.0_11	RANGE_GLCM_SUM_VARIANCE_0.5_13
MEAN_GLCM_MEAN_2.0_13	MEAN_GLCM_STANDARD_DEVIATION_1.0_09
MEAN_GLCM_MEAN_4.0_11	MEAN_GLCM_SUM_AVERAGE_4.0_13
RANGE_GLCM_SUM_AVERAGE_4.0_13	MEAN_GLCM_SUM_VARIANCE_4.0_13
MEAN_GLCM_MEAN_4.0_13	RANGE_GLCM_DIFF_ENTROPY_4.0_13
RANGE_GLCM_DIFF_ENTROPY_4.0_13	RANGE_RLM_LONG_RUN_HIGH_GRAYLEVEL_EMPHASIS_0.5_11
MEAN_GLCM_SUM_AVERAGE_4.0_13	RANGE_RLM_SHORT_RUN_HIGH_GRAYLEVEL_EMPHASIS_0.5_11
MEAN_RLM_HIGH_GRAYLEVEL_RUN_EMPHASIS_0.5_13	MEAN_RLM_GRAYLEVEL_NONUNIFORMITY_0.5_13
MEAN_RLM_RUNLENGTH_NONUNIFORMITY_1.0_11	MEAN_RLM_LONG_RUN_HIGH_GRAYLEVEL_EMPHASIS_0.5_13
MEAN_RLM_SHORT_RUN_HIGH_GRAYLEVEL_EMPHASIS_1.0_13	MEAN_RLM_RUN_PERCENTAGE_0.5_13
MEAN_RLM_RUN_PERCENTAGE_1.0_13	MEAN_RLM_RUNLENGTH_NONUNIFORMITY_0.5_13
MEAN_RLM_RUNLENGTH_NONUNIFORMITY_2.0_13	MEAN_RLM_SHORT_RUN_EMPHASIS_0.5_13
RANGE_RLM_LONG_RUN_EMPHASIS_2.0_13	MEAN_RLM_SHORT_RUN_LOW_GRAYLEVEL_EMPHASIS_0.5_13
MEAN_RLM_RUN_PERCENTAGE_4.0_09	RANGE_RLM_GRAYLEVEL_NONUNIFORMITY_0.5_13
MEAN_RLM_SHORT_RUN_EMPHASIS_4.0_09	RANGE_RLM_LONG_RUN_HIGH_GRAYLEVEL_EMPHASIS_0.5_13
RANGE_RLM_SHORT_RUN_LOW_GRAYLEVEL_EMPHASIS_4.0_11	MEAN_RLM_RUNLENGTH_NONUNIFORMITY_2.0_13
MEAN_RLM_GRAYLEVEL_NONUNIFORMITY_4.0_13	MEAN_RLM_LONG_RUN_HIGH_GRAYLEVEL_EMPHASIS_4.0_13
MEAN_RLM_LONG_RUN_HIGH_GRAYLEVEL_EMPHASIS_4.0_13	MEAN_RLM_LONG_RUN_LOW_GRAYLEVEL_EMPHASIS_4.0_13

Final feature subsets used for standalone testing dataset evaluation, obtained by retaining the top 50 features according to SVMRFE with and without performing RDFS first. RDFS was performed on the entire multi-reconstruction dataset. For readability, features are divided into categories then presented in order of increasing scale. For GLCM and RLM features, MEAN and RANGE specify the method of resolving feature values across 13 directions in 3D.

5.5.3 Evaluation against standalone testing dataset

Table 5.5 summarizes the evaluation of the with-RDFS and without-RDFS classifier models across the entire standalone testing dataset. These confusion matrices represent classifier accuracy against ground truth. The with-RDFS model exhibited slightly worse performance than without-RDFS, with EGM measures of 0.778 and 0.785, respectively. Class-specific performance of the two classifier models was very similar, with a few notable differences such as REC_{PF} and $PREC_{HC}$. Westlake's equivalence test between the models was significant with an 80% limit (P=0.01), indicating equivalence in EGM scores across subjects. 95% confidence intervals (using the exact binomial method) were as follows: (0.606, 0.879) for with-RDFS; (0.659, 0.914) for without-RDFS.

Table 5.6 lists the final feature subsets used by with-RDFS and without-RDFS for standalone testing dataset evaluation.

5.6 Discussion

This study focused on evaluating the novel technique of Robustness-Driven Feature Selection (RDFS) in the setting of CT texture classification of fibrotic interstitial lung disease (FILD). The primary contributions of this investigation are twofold. First, we used multiple systematic reconstructions of CT raw sinogram data to assess the robustness of the features that comprise our feature set, then used this knowledge to drive the feature selection step of the classification pipeline. Second, we used multiple raw sinogram reconstructions to directly evaluate the impact of varying CT technical factors on the output of a CT texture classifier. The use of CT raw sinogram data is key to this study because it allowed us to isolate variation due to technical factors without introducing additional scan-rescan variation (and without increasing radiation dose to our subjects).

Due to the impracticality of collecting a comprehensive dataset that accurately and inclusively reflects real-world scenarios, most research and development of CAD systems for medical imaging takes place in environments of limited data. Nevertheless, it is important for these systems to be able to perform well not only within their own limited development environment, but also when exposed to the variability and unpredictability of real-world use. The RDFS technique offers a method to mitigate the problem of limited data by leveraging CT sinogram reconstructions to mimic anticipated real-world variation in CT technical factors.

The feature robustness investigation revealed some interesting findings. First, we showed that our first-order descriptive features are substantially more robust than our texture features, with a mean robustness index (*RI*) less than half that of the texture features (Table 5.1). To put in another way, although descriptive features comprise only a little over 6% of our total feature space, they make up 92% of the top 50 most robust features according to *RI*. This result is not entirely surprising; although all three of the technical factors considered in this study influence the distribution of graylevel values in the image, they are all mean-preserving. Therefore, first-order descriptive features such as mean and median are not likely to be influenced as strongly as measures of texture.

Another significant finding of the feature robustness analysis is that the robustness of features improved at higher levels of scale (Table 5.1). This result is twofold. First, high levels of Gaussian blurring improve feature robustness. It is likely that this is because the changes in graylevel distribution introduced by technical factors are simply getting blurred out. Second, larger subimage sizes improve feature robustness. Because these subimages are cubical, the 11- and 13-voxel subimages contain 83% and 201% more voxels, respectively, than the 9-voxel subimage. It appears that the smaller subimages are simply too susceptible to the graylevel distribution changes

introduced by different technical factors, and a more macro-level scale is required to achieve stability in feature values.

It is important to note that having more robust features does not mean that these features will be of any value in a particular classification task. For example, applying an extremely high level of Gaussian blurring will certainly improve the robustness of the resulting features, but at the same time it will necessarily blunt their ability to discriminate between classes. In the end, what we desire for any classification task is a set of features that are robust to spurious sources of variation while retaining their sensitivity towards true physiological differences. In our study, we were able to achieve this balance by first applying RDFS then following up with a further feature selection step using SVMRFE.

A classifier can be considered robust if its classifier output remains stable despite spurious changes in its inputs. In practice, it is important for a CAD system to have both good classification accuracy and robustness. We captured these characteristics through the use of two metrics: the extended g-mean (EGM), which measures classification accuracy; and Cohen's kappa, which measures classification agreement with respect to changing technical factors. A classifier could have good accuracy but poor robustness, for example, if it had high EGM values for only one or a few combinations of technical parameters, but low EGM and kappa for other combinations.

The without-RDFS classifier model demonstrated a substantial lack of robustness, particularly when presented with thicker slices of 2.0 mm. This result is not surprising as partial voluming combined with the smoothing effect of thicker slices can result in the loss of textural information. This phenomenon is evidenced by the relatively lower EGM and kappa measures at the 2.0 mm reconstructions (Tables 5.2 and 5.3), and by the large number of classification disagreements in Fig. 5.7. Interestingly, there was no corresponding decrease of robustness at 0.6

mm, indicating that the classifier model was only adversely affected by thicker slices, not thinner, at least within the range of slice thicknesses examined in this investigation.

The with-RDFS classifier model was substantially more robust than without-RDFS. This is indicated by the larger values of EGM and kappa (Tables 5.2 and 5.3) and by the smaller number of classification disagreements in Fig. 5.7. Some trends of reduced agreement for certain reconstructions were observed, but in each of these cases the impact was less than what was observed for without-RDFS.

Compared to slice thickness, reconstruction kernel and tube current did not have as strong of an impact on classifier agreement. However, there appears to be an interaction effect among the technical factors, resulting in reduced EGM and kappa for certain combinations of technical parameters. For example, the combination of thinner slice, sharper kernel, and/or lower tube current (e.g. 0.6 mm, B70f, 50 mAs) had a noticeable negative impact on both classifier models. Each of the technical parameters in this combination introduces more image noise, which interferes with the features' efforts to capture texture information (Fig. 5.1, "Noisy" column). Correspondingly, the opposite combination of thicker slice, smoother kernel, and/or higher tube current (e.g. 2.0 mm, B30f, original tube current), which produces a smoother image with less image noise, also had a negative impact on classifier agreement (Fig. 5.1, "Smooth" column). By contrast, it can be seen that when technical parameters that have opposing effects on image noise are combined, they balance each other, and classifier agreement is not as strongly impacted under these conditions, especially for with-RDFS (Fig. 5.1, "Balanced" column).

The summary of disagreements (Fig. 5.7) reveals that the vast majority of disagreements for the with-RDFS classifier model occur between NL and GG, GG and PF, or PF and HC. An examination of reader variability indicated that the expert readers who provided the VOIs for this

study also disagreed substantially among these same pairs of classes. Furthermore, there is evidence suggesting that normal lung, ground-glass opacity, and fibrosis (and potentially honeycombing) represent a continuum of gradual change in interstitial lung disease [31], making strict differentiation between these classes an inherently difficult problem. It appears that this difficulty is reflected in the behavior of our with-RDFS classifier model.

When evaluated on the standalone testing dataset, the without-RDFS classifier model slightly outperformed the with-RDFS model (Table 5.5). This result is not surprising because the RDFS algorithm discards features on the basis of their robustness, inviting the possibility that some highly informative features may be discarded. In fact, a comparison of the final feature subsets (Table 5.6) reveals some systematic differences between the two models. In particular, without-RDFS relies heavily upon GLCM and RLM features extracted at the 0.5 mm level of Gaussian blurring, which we have shown to be highly nonrobust (Table 5.1, Fig. 5.5). Furthermore, we emphasize that the difference in performance between the two models is small, and that the performance was statistically equivalent when compared on a per-subject basis.

It is important to acknowledge the composition of the standalone testing dataset. Although the standalone testing dataset contains images from outside the range of technical parameters of the training dataset, in particular for reconstruction kernel and slice thickness, these cases represent a minority of the dataset. For example, only 12 out of 42 cases use a nonstandard kernel (where "standard" is defined as a kernel found in the training dataset), and only 2 out of 42 cases use a slice thickness greater than 1.25 mm. We reiterate that the standalone testing dataset evaluation is not intended to provide further evidence of the effectiveness of RDFS at improving classifier robustness, but rather to uncover any potential side effects of RDFS by serving as an independent evaluation dataset that is more similar to the training dataset.

We observe that our results illustrate a fundamental tradeoff between classifier robustness and accuracy. The same features that give the without-RDFS model a small edge on the standalone testing dataset also impair its ability to cope with the wide range of technical parameters found in the multi-reconstruction dataset. To put it more simply, some features that are highly sensitive to signal (variation between classes) are also highly sensitive to noise (variation due to technical factors). By applying RDFS, we were able to identify a subset of features that are almost as sensitive to signal while being substantially more robust to noise. The power of the RDFS approach lies in the ability to characterize feature robustness, which is essential for finding the most effective and worthwhile tradeoff for a particular classification task.

In this study, we developed a two-step methodology for feature selection, with the first step (RDFS) selecting based on feature robustness and the second step selecting based on informativeness. In this manner, our proposed methodology may be thought of as a composite feature selection method that balances two separate criteria, using the robustness index threshold (Fig. 5.6) as a parameter to determine the relative importance of each. Although we focused our investigation using SVMRFE as the second feature selection step, in general any method that selects informative features could be used, along with any classification method. We briefly examine a few other such methods in the Appendix.

There were several limitations to our study. Although the training and testing datasets are very similar to each other, the multi-reconstruction dataset has some notable differences. The multi-reconstruction dataset represents a general interstitial lung disease population, whereas the training and testing cases all come from either idiopathic pulmonary fibrosis or lung disease associated with systemic scleroderma. Furthermore, the multi-reconstruction images were acquired with supine patient positioning, unlike the training and testing images which were

acquired at prone. The multi-reconstruction images were annotated by a single reader, while the training and testing images followed a two-reader paradigm. Lastly, the reduced-tube-current images in the multi-reconstruction dataset were generated using a synthetic noise algorithm rather than true CT image noise. However, an investigation of this sort would not have been possible with true CT noise due to the additional variation introduced by multiple acquisitions at different levels of tube current and due to the ethical and practical concerns of patient radiation dose associated with such an approach. The synthetic noise approach used in this study has been previously validated [13-15].

In our study, we have evaluated robustness with respect to three technical factors: slice thickness, reconstruction kernel, and tube current. Many other sources of variation in CT imaging exist, such as scan/rescan variability, imaging device manufacturer and model, and patient demographics. Further research will need to be done in order to improve classifier robustness in these areas.

5.7 Conclusion

We have developed a novel technique for feature selection called Robustness-Driven Feature Selection. Following this method, we were able to substantially improve the robustness of a support vector classifier for fibrotic interstitial lung disease while maintaining its performance on a standalone testing dataset. These results have implications for improving the generalizability of classifier-based CT CAD systems, which is of great importance in multicenter clinical trials that rely on accurate and reproducible measures. This work will ultimately help pave the way for more widespread adoption of these systems in clinical practice.

5.A. Evaluation with other feature selection and classification methods

In the present study, we have investigated the effect of our proposed Robustness-Driven Feature Selection (RDFS) method in conjunction with Support Vector Machine Recursive Feature Elimination (SVMRFE) as a further feature selection step followed by support vector machine (SVM) classification. In practice, RDFS may be applied in combination with any further feature selection scheme and classification method. In this appendix, we briefly examine a few other combinations. This investigation is intended to be illustrative rather than exhaustive, to demonstrate the applicability of our proposed RDFS method with methods other than those previously discussed.

The experimental design for this investigation mirrors the multi-reconstruction evaluation from the main study. Briefly, we create two classifier models, one with and one without RDFS, for each combination of methods. The classification pipeline involves RDFS on the multi-reconstruction dataset (if applicable); followed by further feature selection, parameter tuning, and model training on the training dataset; followed at last by evaluation on the multi-reconstruction dataset. Two-fold cross evaluation on the multi-reconstruction dataset is employed for RDFS.

We evaluate a total of six different classifier models, which we will refer to as models A through F. Models A and B are identical to the with-RDFS and without-RDFS models, respectively, examined in the main study. Models C and D use sequential floating forward selection (SFFS) as the further feature selection step [32], followed by SVM as the classification method. Finally, models E and F use random forests (RF) as the classification method [33], with no explicit further feature selection beyond the built-in RF feature selection. These models are summarized in Table 5.7.

Table 5.7. Summary of classifier models

Model	A	B	C	D	E	F
RDFS?	yes	no	yes	no	yes	no
Further feature selection	SVMRFE	SVMRFE	SFFS	SFFS	none	none
Classification method	SVM	SVM	SVM	SVM	RF	RF

The results of evaluating the six models are presented in Table 5.8. For brevity, only Cohen's kappa measures indicating classifier agreement are reported, and the confidence intervals have been omitted. For all models, kappa values are highest for slice thicknesses 0.6 mm and 1.0 mm and for B30f and B45f kernels, demonstrating a noticeable deterioration at 2.0 mm and B70f (except perhaps when 2.0 mm and B70f are applied simultaneously). Reduced tube current does not appear to have a strong impact on classifier agreement except in conjunction with the B70f kernel. Lastly, for each pair of models (A+B, C+D, E+F), the model utilizing RDFS exhibits higher kappa values than the model without, most notably in those reconstructions suffering the most deterioration.

We conclude that RDFS is an effective tool for improving classifier robustness that may be employed in combination with a variety of feature selection schemes or classification methods.

Table 5.8. Robustness of classifier models on multi-reconstruction dataset

Model A*	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.974	0.963	0.958	0.979	0.968	0.984	0.921	0.952	0.899
1.0 mm	0.952	0.953	0.963	---	0.979	0.989	0.947	0.947	0.909
2.0 mm	0.931	0.920	0.920	0.931	0.926	0.947	0.968	0.952	0.942

Model B	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.968	0.968	0.963	0.942	0.936	0.905	0.880	0.894	0.872
1.0 mm	0.936	0.952	0.946	---	0.968	0.952	0.905	0.926	0.876
2.0 mm	0.855	0.861	0.850	0.865	0.865	0.871	0.860	0.844	0.827

Model C*	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.958	0.952	0.942	0.953	0.952	0.937	0.900	0.921	0.872
1.0 mm	0.952	0.936	0.942	---	0.968	0.968	0.942	0.947	0.909
2.0 mm	0.846	0.857	0.862	0.894	0.899	0.899	0.926	0.915	0.920

Model D	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.952	0.947	0.920	0.968	0.968	0.931	0.931	0.931	0.851
1.0 mm	0.936	0.957	0.936	---	0.989	0.942	0.931	0.904	0.850
2.0 mm	0.784	0.784	0.784	0.800	0.789	0.795	0.719	0.707	0.701

Model E*	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.951	0.973	0.962	0.962	0.968	0.957	0.930	0.930	0.857
1.0 mm	0.951	0.967	0.967	---	0.978	0.989	0.968	0.935	0.868
2.0 mm	0.907	0.907	0.896	0.907	0.913	0.918	0.935	0.935	0.902

Model F	B30f			B45f			B70f		
	Original	100mAs	50mAs	Original	100mAs	50mAs	Original	100mAs	50mAs
0.6 mm	0.956	0.956	0.957	0.967	0.957	0.957	0.908	0.897	0.809
1.0 mm	0.962	0.940	0.962	---	0.962	0.951	0.930	0.897	0.829
2.0 mm	0.890	0.874	0.890	0.868	0.852	0.841	0.837	0.793	0.757

Robustness is measured via Cohen's kappa by comparing classifier output for each non-reference reconstruction against the reference reconstruction (denoted with a gray background). The six classifier models A-F are described in Table 5.7. *Denotes a classifier model that utilizes RDFS.

5.8 References

[1] Boedeker KL, McNitt-Gray MF, Rogers MR, Truong DA, Brown MS, Gjertson DW, Goldin JG. Emphysema: Effect of reconstruction algorithm on CT imaging measures. Radiology 2004. 232(1):295-301.

[2] Gierada DS, Bierhals AJ, Choong CK, Bartel ST, Ritter JH, Das NA, Hong C, Pilgram TK, Bae KT, Whiting BR, Woods JC, Hogg JC, Lutey BA, Battafarano RJ, Cooper JD, Meyers BF, Patterson GA. Effects of CT section thickness and reconstruction kernel on emphysema quantification: Relationship to magnitude of the CT emphysema index. Acad Radiol 2010. 17(2):146-56.

[3] Bartel ST, Bierhals AJ, Pilgram TK, Hong C, Schechtman KB, Conradi SH, Gierada DS. Equating quantitative emphysema measurements on different CT image reconstructions. Med Phys 2011. 38(8):4894-902.

[4] Kim HJ, Li G, Gjertson D, Elashoff R, Shah SK, Ochs R, Vasunilashorn F, Abtin F, Brown MS, Goldin JG. Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. Acad Radiol 2008. 15(8):1004-16.

[5] Kim HJ, Tashkin DP, Clements PJ, Li G, Brown MS, Elashoff R, Gjertson DW, Abtin F, Lynch DA, Strollo DC, Goldin JG. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. Clin Exp Rheumatol 2010. 28(5 sup 62):S26-35.

[6] Wang J, Li F, Doi K, Li Q. Computerized detection of diffuse lung disease in MDCT: The usefulness of statistical texture features. Phys Med Biol 2009. 54(22):6881-99.

[7] Xu Y, van Beek EJ, Hwanjo Y, Guo J, McLennan G, Hoffman EA. Computer-aided classification of interstitial lung dis-eases via MDCT: 3D adaptive multiple feature method (3D AMFM). Acad Radiol 2006. 13(8):969-78.

[8] Maldonado F, Moua T, Rajagopalan S, Karwoski RA, Raghunath S, Decker PA, Hartman TE, Bartholmai BJ, Robb RA, Ryu JH. Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. Eur Respir J 2014. 43(1):204-12.

[9] Kim HJ, Brown MS, Elashoff R, Li G, Gjertson DW, Lynch DA, Strollo DC, Kleerup E, Chong D, Shah SK, Ahmad S, Abtin F, Tashkin DP, Goldin JG. Quantitative texture-based assessment of one-year changes in fibrotic reticular patterns on HRCT in scleroderma lung disease treated with oral cyclophosphamide. Eur Radiol 2011. 21(12):2455-65.

[10] Armato SA, Altman MB, La Riviere PJ. Automated detection of lung nodules in CT scans: Effect of image reconstruction algorithm. Med Phys 2003. 30(3):461-72.

[11] Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, Goldgof DB, Hall LO, Korn R, Zhao B, Schwartz LH, Basu S, Eschrich S, Gatenby RA, Gillies RJ. Test-retest reproducibility analysis of lung CT image features. J Digit Imaging 2014. 27(6):805-23.

[12] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986. 1(8476):307-10.

[13] Massoumzadeh P, Don S, Hildebolt F, Bae KT, Whiting BR. Validation of CT dose-reduction simulation. Med Phys 2009. 36(1):174-89.

[14] Zabic S, Wang Q, Morton T, Brown KM. A low dose simulation tool for CT systems with energy integrating detectors. Med Phys 2013. 40(3):031102.

[15] Young S, McNitt-Gray M. Estimating lesion volume in low-dose chest CT: How low can we go? SPIE Medical Imaging 2013. 9033:1-13.

[16] Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans Syst Man Cybernetics 1973. 3:610-621.

[17] Haralick RM. Statistical and structural approaches to texture. Proc IEEE 1979. 67:786-804.

[18] Galloway M. Texture analysis using gray level run lengths. Comput Graph Imaging Process 1975. 4:172-9.

[19] Tang X. Texture information in run-length matrices. IEEE Trans Image Proc 1998. 7(11):1602-9.

[20] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines. In proceedings of the International Join Conference on AI 1999. 55-60.

[21] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority oversampling technique. J Artificial Intelligence Research 2002. 16:321-57.

[22] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced data sets. Lecture Notes in Computer Science 2004. 3201:39-50.

[23] Tang Y, Zhang YQ. Granular SVM with repetitive undersampling for highly imbalanced protein homology prediction. Proc Intl Conf Granular Computing 2006. 457-460.

[24] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002. 46:389-422.

[25] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[26] Hsu CW, Lin CJ. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 2002. 13:415-25.

[27] Sun Y, Kamel MS, Wang Y. Boosting for learning multiple classes with imbalanced class distribution. Proc Intl Conf Data Mining 2006. 592-602.

[28] Cohen JA. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960. 20:37-46.

[29] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009. 11(1).

[30] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[31] Wells AU. High-resolution computed tomography and scleroderma lung disease. Rheumatology (Oxford) 2008. 47(sup 5):59-61.

[32] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters 1994. 15(11):1119-25.

[33] Breiman L. Random forests. Machine Learning 2001. 45(1):5-32.

6. Robustness-driven feature selection in CT classification of fibrotic interstitial lung disease: The effect of 3.0 mm slice thickness

6.1 Introduction

In the previous chapter, we have demonstrated that the Robustness-Driven Feature Selection (RDFS) technique is highly effective at improving CT texture classification robustness with respect to slice thickness, reconstruction kernel, and tube current. In particular, we examined slice thicknesses of 0.6, 1.0, and 2.0 mm, and showed that the 2.0 mm slice thickness was a particular challenge for classifier robustness. This result raises the question of how classifier robustness might be affected if slice thickness is increased further.

The purpose of this study is to determine the extent to which RDFS is able to improve classifier robustness against 3.0 mm slice thicknesses.

6.2 Materials

6.2.1 CT imaging data

The study population consisted of three datasets: training, multi-reconstruction, and extended multi-reconstruction. The training dataset consisted of 45 adult subjects with an established clinical diagnosis of either interstitial lung disease associated with systemic scleroderma (SSc) or idiopathic pulmonary fibrosis (IPF). The multi-reconstruction and extended multi-reconstruction datasets consisted of 12 subjects with diffuse lung disease. All imaging data were anonymized and made available through a central imaging core lab with the approval of a local institutional review board and was accessed in compliance with the Health Insurance Portability and Accountability Act.

For the training dataset, volumetric high-resolution CT images were collected for all subjects. The CTs were performed under the auspices of a central imaging core that provided training and prescribed standardized guidelines for image acquisition and reconstruction. A variety of imaging devices were used from Siemens (Siemens Healthcare, Forchheim, Germany), General Electric (GE Healthcare, Waukesha, WI, USA), Philips (Philips Healthcare, Cleveland, OH, USA), and Toshiba (Toshiba America Medical Systems, Tustin, CA, USA). Images were acquired at full inspiration in the prone position at 120 kVp. The CT technical parameters were as follows: an average tube current between 100 mAs to 150 mAs; slice thickness between 1.0 mm and 1.25 mm, inclusive; and a medium-sharp reconstruction kernel of B45f, BONE, D, or FC52.

For the multi-reconstruction dataset, CT raw sinogram data were collected for all subjects. The CTs were performed as a part of standard clinical practice using a diffuse lung disease protocol. The subjects were imaged at full inspiration in the supine position at 120 kVp using a multidetector CT device (Definition Flash, Siemens Healthcare, Forchheim, Germany). Tube current modulation was used, resulting in average tube currents ranging from 211 mAs to 328 mAs per subject. The raw sinogram data were used to create additional reconstructions by systematically varying slice thickness and reconstruction kernel. In addition, simulated reduced-tube-current images were generated via synthetic noise, which was added to the CT raw sinogram data using a previously-validated algorithm described in [1-3]. Three slice thicknesses (0.6, 1.0, 2.0 mm), three kernels (B30f-smooth, B45f-medium-sharp, B70f-sharp), and three tube currents (original tube current, 100 mAs, 50 mAs) were used for a total of 27 reconstructions per subject. Of these, the reconstruction corresponding to a slice thickness of 1.0 mm, kernel of B45f, and the original tube current was designated as the "reference reconstruction" since these parameters corresponded most closely with the parameters from the training dataset.

Table 6.1. Characterization of datasets

Dataset	Training	Multi-reconstruction	Extended multi-reconstruction
# of subjects	42	12	12
Type of disease	SSc or IPF	ILD	ILD
Patient position	Prone	Supine	Supine
Breathhold	Full inspiration	Full inspiration	Full inspiration
CT manufacturer	Siemens, GE, Phillips, or Toshiba	Siemens	Siemens
Slice thickness (mm)	{1.0, 1.25}	{0.6, 1.0, 2.0}	{0.6, 1.0, 2.0, 3.0}
Reconstruction kernel	B45f, BONE, D, or FC52	{B30f, B45f, B70f}	{B30f, B45f, B70f}
Tube current (mAs)	≥ 100	Original, 100, 50	Original, 100, 50
# of reconstructions	1	27	36

The extended multi-reconstruction dataset is a superset of the multi-reconstruction dataset.

Lastly, the extended multi-reconstruction dataset consists of all of the same subjects and reconstructions as the multi-reconstruction dataset. In addition to these, 3.0 mm slice thickness reconstructions were created as well, bringing the total number of reconstructions in the extended multi-reconstruction dataset to 36.

The three datasets training, multi-reconstruction, and extended multi-reconstruction are summarized in Table 6.1.

6.2.2 Small volumes of interest for classifier development and assessment

For the training dataset, two experienced thoracic radiologists (JGG, FGA) provided a total of 2120 cubic volumes of interest (VOIs) corresponding to six visually-based textural and structural classes: pulmonary fibrosis (PF), ground-glass opacity (GG), honeycombing (HC), normal lung parenchyma (NL), airways (AIR), and vessels (VES). The readers followed a two-pass independent reading paradigm. In the first pass, each reader independently placed VOIs throughout the lungs corresponding to each of the above classes, assigning their VOIs with the appropriate class labels. In the second pass, each reader was independently presented with unlabeled copies of the other reader's VOIs and asked to assign class labels according to their best judgment. At the end of this reading process, each VOI had two labels, one from each reader. Of the total 2120 VOIs,

1798 VOIs were assigned identical class labels by the two readers. Development and training of the classifier was performed on this agreement subset only.

For the multi-reconstruction and extended multi-reconstruction datasets, VOIs were provided by one experienced thoracic radiologist (JGG), who annotated a total of 238 cubic VOIs corresponding to the same six classes as above (PF, GG, HC, NL, AIR, VES). For each subject, the VOIs were annotated on the reference reconstruction only, then they were propagated to the remaining reconstructions.

In summary, the agreement subset for the training dataset consisted of 1798 VOIs, with individual class counts of 564, 272, 42, 272, 294, and 354 for PF, GG, HC, NL, AIR, and VES, respectively. The multi-reconstruction datasets had no agreement subset as it was only annotated by a single radiologist, and it had individual class counts of 75, 37, 23, 27, 43, and 33, respectively.

6.3 Methods

6.3.1 Feature extraction and support vector machine classification

For each CT image, a Gaussian blurring filter with 0.5 mm radius was applied, followed by isotropic resampling of the image volume, using trilinear interpolation to produce a resampled image volume with $0.5 \times 0.5 \times 0.5$ mm^3 voxels. Next, for each radiologist-provided VOI in the image, a small cubical subimage of size 9 voxels was extracted centered on the VOI. An image intensity histogram was computed on this subimage, and first-order descriptive features were calculated on the histogram. Next, the subimage was adaptively rebinned to 16 graylevels, and 3D graylevel co-occurrence matrix [4-5] and run-length matrix [6-7] texture features were extracted from the rebinned subimage. This process was repeated for subimage sizes 9, 11, and 13 voxels and Gaussian radii 0.5, 1.0, 2.0, 4.0 mm for a total of twelve combinations of scalespace parameters, resulting in 792 features in all.

The underrepresentation of honeycombing (HC) examples was identified as a potential limitation in the training dataset (only 42 of 1798 VOIs). We utilized the synthetic minority oversampling technique described in [8], which generates additional synthetic examples of a class by taking linear combinations of existing examples in the feature space. In order to mitigate the impact of class imbalance, we boosted the HC examples by 500%, bringing the number of HC examples to 252 (and increasing the total number of training VOIs to 2008).

A support vector machine (SVM) classifier was trained on our data as follows. First, all feature values were standardized to zero mean and unit variance. Next, feature selection was performed through the use of Robustness-Driven Feature Selection (RDFS) as described in Chapter 5. After RDFS, further feature selection was performed using the Support Vector Machine Recursive Feature Elimination method (SVMRFE) [9] to produce a final feature subset of 50 features. The optimal SVM cost parameter C and radial basis function parameter γ were selected via gridsearch with 5-fold cross validation [10]. Additionally, in order to further mitigate the impact of class imbalance, we assigned individual class weights in the SVM model to be inversely proportional to the one-vs-all ratio for the number of training examples of each class as described in [11]. Finally, the SVM classifier model was trained on the training dataset using the selected features and parameters. Classification of multiple classes was performed using the one-against-one method described in [12]. Briefly, separate binary SVM classifiers were constructed for each pairwise combination of classes, and a voting strategy was employed to predict new instances.

6.3.2 Experimental Design

Three different classifier models were constructed: RDFS-2.0, RDFS-3.0, and without-RDFS. The RDFS-2.0 and RDFS-3.0 models differed only in which dataset was used as an input to the RDFS algorithm; RDFS-2.0 used the multi-reconstruction dataset, while RDFS-3.0 used the extended

multi-reconstruction dataset. By contrast, the without-RDFS model did not use RDFS at all. Note that the RDFS-2.0 and without-RDFS classifier models are respectively identical to the with-RDFS and without-RDFS models described in Chapter 5. The three classifier models are illustrated in Fig. 6.1.

The three classifier models (RDFS-2.0, RDFS-3.0, and without-RDFS) were evaluated on the extended multi-reconstruction dataset in order to directly assess their robustness under conditions of changing technical parameters. Classifier robustness was evaluated by computing the extended g-mean (EGM) with respect to ground truth for each of the 36 reconstructions, then by computing Cohen's kappa with respect to the reference reconstruction for each of the 35 non-reference reconstructions.

Two-fold cross evaluation was used for RDFS-2.0 and RDFS-3.0. Briefly, the 12 subjects that make up the multi-reconstruction dataset were separated into two folds, and the extended multi-reconstruction dataset was separated into two folds according to the same division of subjects. RDFS was applied using one fold of the multi-reconstruction dataset (extended multi-reconstruction for RDFS-3.0). Further feature selection (via SVMRFE), parameter tuning, and model training were done using the training dataset. Finally, the resulting classifier models were evaluated using the second fold of the extended multi-reconstruction dataset. For without-RDFS, feature selection (via SVMRFE only), parameter tuning, and model training were all done on the training dataset, and the resulting classifier model was evaluated on the extended multi-reconstruction dataset.

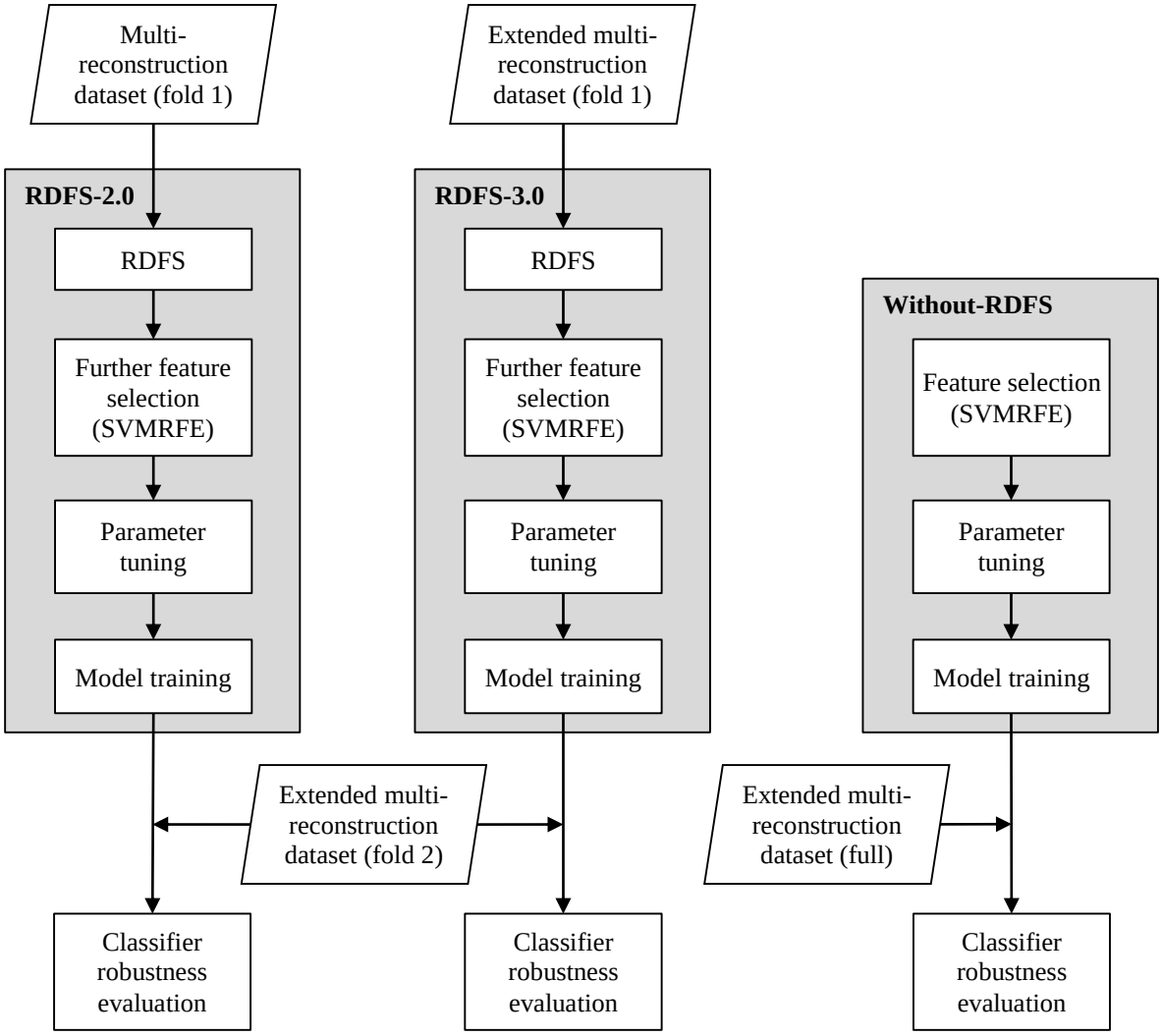

Figure 6.1. Illustration of support vector classification pipeline for the three classifier models. RDFS-2.0 and RDFS-3.0 differ only in the input to the RDFS algorithm, and both utilize two-fold cross evaluation. Note that all three models are evaluated using the extended multi-reconstruction dataset. SVMRFE, parameter tuning, and model training are all done through the training dataset (not pictured).

Note that the input to RDFS differed depending on the classifier model (multi-reconstruction for RDFS-2.0, extended multi-reconstruction for RDFS-3.0); however, model evaluation for all three models was performed using the extended multi-reconstruction dataset. The classification pipeline for the three models is illustrated in Fig. 6.1.

Finally, because of the potential sensitivity of RDFS to the choice of robustness index (RI) threshold, a range of five different RI threshold values was evaluated for both of the models. EGM

and kappa scores are reported as the mean and standard deviation across these thresholds. No corresponding threshold exists for without-RDFS, so its EGM and kappa scores are simply reported as is.

Support vector classification was performed within the environment of Weka 3.7.11 (Waikato Environment for Knowledge Analysis, The University of Waikato, Hamilton, New Zealand) [13] using the implementation provided in the software package LibSVM (National Taiwan University, Taipei, Taiwan) [14]. Statistical analysis was performed using Microsoft Excel 2013 (Redmond, WA, USA).

6.4 Results

Table 6.2 summarizes the results of the feature robustness analysis for both the RDFS-2.0 and RDFS-3.0 approaches. Of the three types of features, first-order descriptive features demonstrated the smallest robustness index (RI), indicating that these features are the most robust with respect to variations in slice thickness, reconstruction kernel, and tube current. Furthermore, there was a trend of decreasing RI with increasing subimage window size and especially with increasing Gaussian blurring radius, indicating that features extracted at higher levels of scale were more robust.

Table 6.2. Robustness index summarized by feature category, Gaussian radius, and subimage window size

Category	RDFS-2.0	RDFS-3.0
First-order descriptive (n = 48)	0.179 (0.005)	0.220 (0.007)
Texture-GLCM (n = 480)	0.387 (0.155)	0.482 (0.178)
Texture-RLM (n = 264)	0.361 (0.130)	0.459 (0.178)
Gaussian radius (mm)	RDFS-2.0	RDFS-3.0
0.5 (n = 198)	0.512 (0.185)	0.630 (0.236)
1.0 (n = 198)	0.376 (0.111)	0.472 (0.130)
2.0 (n = 198)	0.316 (0.074)	0.408 (0.087)
4.0 (n = 198)	0.259 (0.053)	0.323 (0.058)
Subimage window size (voxels)	RDFS-2.0	RDFS-3.0
9 (n = 264)	0.401 (0.156)	0.495 (0.180)
11 (n = 264)	0.360 (0.146)	0.450 (0.180)
13 (n = 264)	0.337 (0.142)	0.430 (0.183)

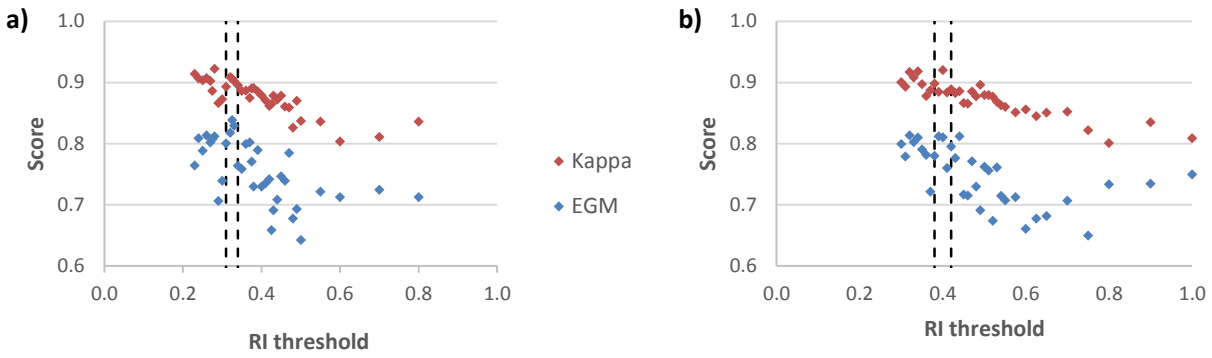

Figure 6.2. Result of experiment for determining appropriate Robustness Index (RI) threshold for Robustness-Driven Feature Selection, single fold. (a) RDFS-2.0; (b) RDFS-3.0. Dashed lines indicate range of threshold values selected for each classifier model. Similar trends were observed for other fold.

Fig. 6.2 illustrates the results of the experiment to determine an appropriate RI threshold for the RDFS-2.0 and RDFS-3.0 classifier models, using one fold of the cross-evaluation approach. Both classifier performance (according to EGM) and agreement (according to kappa) demonstrated a decreasing trend as feature robustness was increased (allowing less robust features to be included), although EGM exhibited much noisier behavior than kappa. Based on these results, a range of five threshold values centered at 0.325 was chosen for RDFS-2.0, and a range centered at 0.400 was chosen for RDFS-3.0. Similar results were obtained for the other fold.

The results of evaluating the RDFS-2.0, RDFS-3.0, and without-RDFS classifier models on the extended multi-reconstruction dataset are illustrated in Figs. 6.3 and 6.4. In summary, these results show that the RDFS-2.0 and RDFS-3.0 models are substantially more robust than without-RDFS; however, they are nearly indistinguishable from each other. All three classifier models demonstrate similar behavior with respect to changing technical parameters. EGM and kappa both decrease sharply for slice thicknesses of 2.0 or 3.0 mm, although this effect is much more pronounced for without-RDFS.

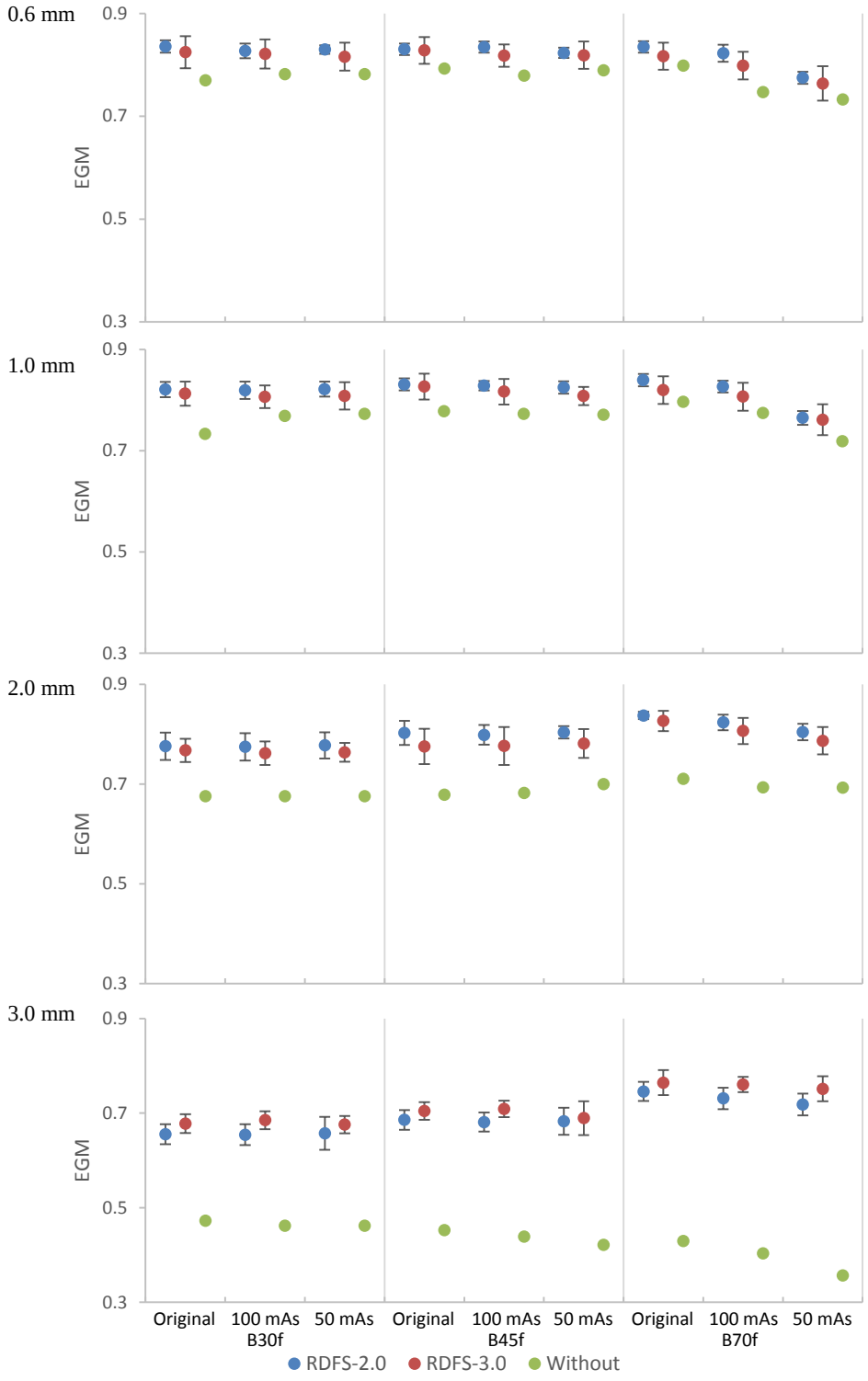

Figure 6.3. Comparison of extended g-mean (EGM) for RDFS-2.0, RDFS-3.0, and without-RDFS classifier models. Each set of three datapoints represents one combination of technical parameters. For RDFS-2.0 and RDFS-3.0, datapoints and error bars indicate mean and standard deviation across five different RI thresholds.

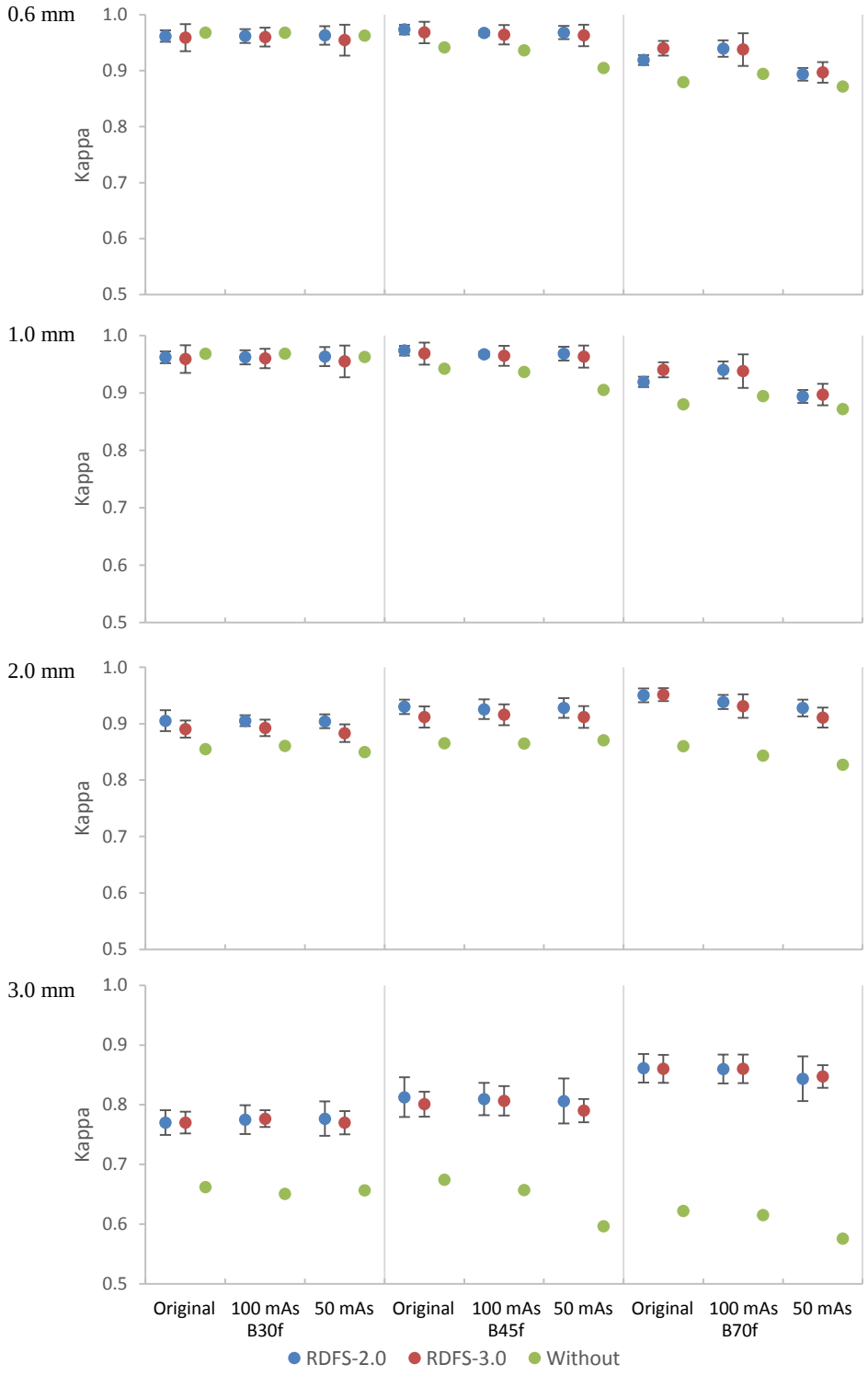

Figure 6.4. Comparison of kappa measure for RDFS-2.0, RDFS-3.0, and without-RDFS classifier models. Each set of three datapoints represents one combination of technical parameters. For RDFS-2.0 and RDFS-3.0, datapoints and error bars indicate mean and standard deviation across five different RI thresholds.

6.5 Discussion

Compared to RDFS-2.0, the RI values for RDFS-3.0 were higher for every category and level of scale (Table 6.2). This result raises an important point about the robustness index, namely that as a quantitative measure, the RI itself is subject to sources of variation. There are many factors that can influence the RI of a feature, including (but not limited to) the choice of reconstructions to include in the multi-reconstruction dataset; the anatomy and disease being studied; the demographics of the subject population; and the distribution of volumes of interest (VOIs) throughout the CT images. Our results indicate that including 3.0 mm slice thickness reconstructions increases the value of RI, and we have taken this into account by selecting different RI thresholds for both the RDFS-2.0 and RDFS-3.0 classifier models (Fig. 6.2).

The RDFS-2.0 and RDFS-3.0 classifier models are subject to variation due to the choice of the robustness index (RI) threshold. This variation is reflected in the standard deviations of the EGM and kappa scores, which are depicted as error bars in Figs. 6.3 and 6.4. The choice of RI threshold effectively determines how many features are retained leading into the SVMRFE step of the classification pipeline (Fig. 6.1). In one fold of RDFS-2.0, for example, an RI threshold of 0.325 retains 350 features, while raising the threshold to 0.330 increases the number of retained features to 365 (including the same 350). With 15 additional features to choose from, SVMRFE may select a different subset of features, which in turn influences the behavior of the resulting classifier model.

It is significant that the robustness of the RDFS-3.0 classifier model is no better than RDFS-2.0. Although RDFS-3.0 incorporates additional information (in the form of 3.0 mm slice thickness reconstructions included in the extended multi-reconstruction dataset), this additional information does not result in improved classification performance or agreement. It appears that

no significant insight is gained by introducing 3.0 mm slice thicknesses into robustness index calculations. In other words, features that are robust (or nonrobust) at 2.0 mm tend to remain robust (or nonrobust) at 3.0 mm as well.

Our results show that for all three classifier models (RDFS-2.0, RDFS-3.0, and without-RDFS), classifier performance and agreement drop sharply at 2.0 mm and especially 3.0 mm. We suspect that this behavior is due to the difficulty of computing 3D texture features as the image volume becomes increasingly non-isotropic. The CT images in the training and multi-reconstruction datasets have sub-millimeter in-plane resolution, and all images are isotropically resampled to 0.5 mm voxels. This means that for the 0.6 mm and 1.0 mm slice thickness reconstructions, each pair of consecutive slices in the original CT image volume is longitudinally sampled at most two times. However, for the 2.0 mm and 3.0 mm slice thickness reconstructions, each slice pair is longitudinally sampled many times. Under these circumstances, there is simply too much loss of information due to the thicker slices. Particularly challenging are the fine textural patterns of pulmonary fibrosis (PF) and structural patterns of honeycombing (HC), where individual cysts can be smaller than 3.0 mm in size.

Compared to the without-RDFS classifier model, the application of RDFS substantially improves classification robustness. However, the fact that RDFS-3.0 did not yield any further improvement over RDFS-2.0 suggests that we may have reached the limit of what we can accomplish through robustness-improving methods alone. The decrease in performance and agreement between 2.0 mm and 3.0 mm is much larger than the corresponding decrease between 1.0 mm and 2.0 mm, as Fig. 6.3 and Fig 6.4 show. Furthermore, examining the classification disagreements at each reconstruction for RDFS-2.0 (Fig. 6.5) reveals that the disagreements

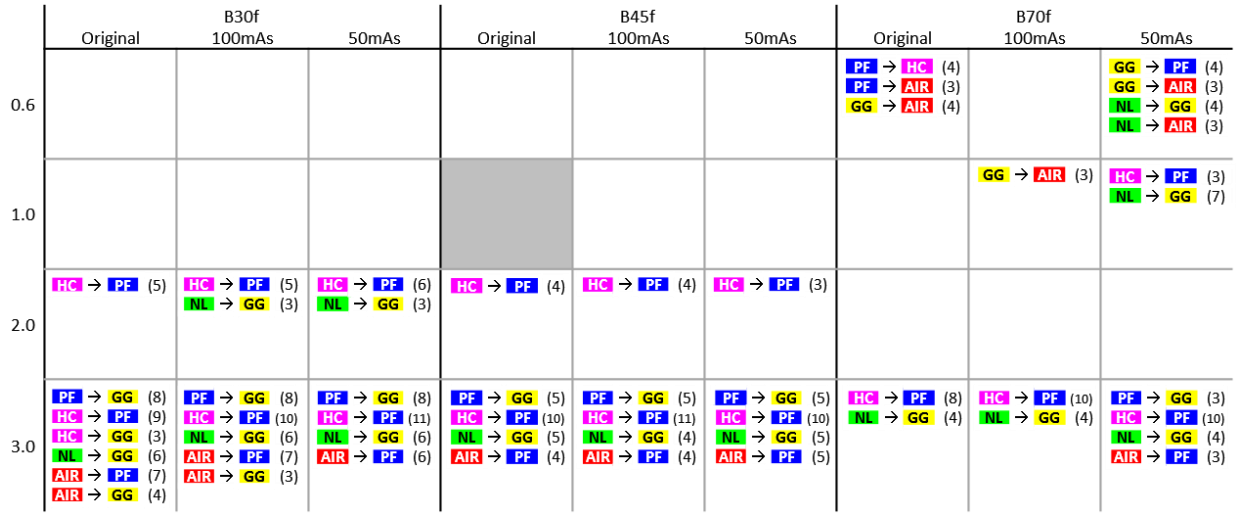

Figure 6.5. Summary of classification disagreements between reference reconstruction (noted in gray) and each other reconstruction for RDFS-2.0, using a robustness index threshold of 0.25. The numbers in parentheses represent the number of disagreements between indicated classes. For example, the top-right-most entry indicates that eight VOIs which were classified as NL on the reference reconstruction were instead classified as GG on the 0.6 mm, B70f, 50 mAs reconstruction. Disagreements consisting of fewer than three cases are not listed in the diagram.

become substantially more numerous at 3.0 mm. The 3.0 mm slice thickness appears to be simply too much for the classifier to handle. Based on these observations, we conclude that the 3.0 mm slice thickness is beyond the limits of acceptable input for our classifier.

In order to improve classification robustness with respect to the 3.0 mm slice thickness, it may be necessary to explore other types of features. In particular, 2D texture features may prove to be less sensitive to the effects of non-isotropy at thicker slices; however, this may come at the cost of reduced discriminative ability for the classification task.

One limitation of our study is that although we examined slice thickness in detail, we did not consider slice spacing. The extended multi-reconstruction dataset consists of CT images with contiguous slices, where slice spacing equals slice thickness. If the 2.0 mm and 3.0 mm images were reconstructed using overlapping slices, the increased longitudinal sampling might offset some of the information lost due to partial voluming of the thicker slices. Further research will be

necessary to understand the interaction between slice thickness and slice spacing and their effect on feature and classifier robustness.

In conclusion, using our Robustness-Driven Feature Selection framework, we have developed a support vector classifier that is robust to variations in slice thickness, reconstruction kernel, and tube current. We have demonstrated that the classifier model can handle CT images up to 2.0 mm in slice thickness, but not 3.0 mm. These results have implications for determining the limits of input for classifier-based CAD systems.

6.6 References

[1] Massoumzadeh P, Don S, Hildebolt F, Bae KT, Whiting BR. Validation of CT dose-reduction simulation. Med Phys 2009. 36(1):174-89.

[2] Zabic S, Wang Q, Morton T, Brown KM. A low dose simulation tool for CT systems with energy integrating detectors. Med Phys 2013. 40(3):031102.

[3] Young S, McNitt-Gray M. Estimating lesion volume in low-dose ch est CT: How low can we go? SPIE Medical Imaging 2013. 9033:1-13.

[4] Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans Syst Man Cybernetics 1973. 3:610-621.

[5] Haralick RM. Statistical and structural approaches to texture. Proc IEEE 1979. 67:786-804.

[6] Galloway M. Texture analysis using gray level run lengths. Comput Graph Imaging Process 1975. 4:172-9.

[7] Tang X. Texture information in run-length matrices. IEEE Trans Image Proc 1998. 7(11):1602-9.

[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority oversampling technique. J Artificial Intelligence Research 2002. 16:321-57.

[9] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002. 46:389-422.

[10] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[11] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced data sets. Lecture Notes in Computer Science 2004. 3201:39-50.

[12] Hsu CW, Lin CJ. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 2002. 13:415-25.

[13] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009. 11(1).

[14] Chang CC, Lin CJ. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Appendix A. Reader agreement investigation

A total of 4088 cubic volumes of interest (VOIs) were provided by two expert radiologists (J.G.G. and F.G.A.) for training, development, and evaluation of CT classifiers for fibrotic interstitial lung disease. These VOIs represented examples of six visually-based textural and structural classes (See Fig 4.1): pulmonary fibrosis (PF), ground-glass opacity (GG), honeycombing (HC), normal lung parenchyma (NL), airways (AIR), and vessels (VES).

The readers followed a two-pass independent reading paradigm. In the first pass, each reader independently placed VOIs throughout the lungs in regions corresponding to each of the above classes, assigning their VOIs with the appropriate class labels. In the second pass, each reader was independently presented with unlabeled copies of the other reader's VOIs and asked to assign class labels according to their best judgment. At the end of this reading process, each VOI had two labels, one from each reader. The subset of VOIs for which the two readers provided identical labels was identified. Training, development, and evaluation of classifier models was performed on this agreement subset only.

Out of the original 4088 VOIs provided by the readers, 2120 were in the training dataset and 1968 were in the testing dataset. The two readers provided identical labels for 1798 (84.8%) and 1645 (83.6%) VOIs, respectively. The distributions of class labels for the VOIs are summarized in Fig. A.1. Of particular note is the small number of HC examples, especially in the training dataset.

Agreement between the two readers is reported in Tables A.1 and A.2. According to the Cohen's kappa measure, the two readers achieved an overall agreement of 0.810 for the training dataset and 0.800 for the testing dataset. Kappa values for individual classes varied, with GG, HC, and NL below 0.800; PF between 0.800 and 0.900; and AIR and VES above 0.900.

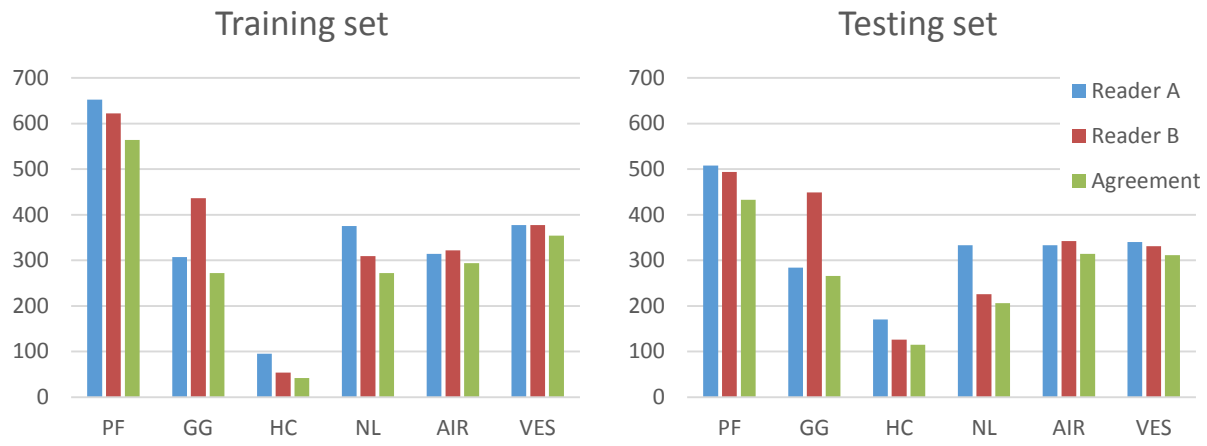

Figure A.1. Distribution of VOI class labels as provided by expert readers for training and testing datasets.

A close examination of the confusion matrices reveals systematic differences in judgment between the two expert readers, particularly between certain pairs of classes. For example, when considering NL versus GG, Reader A tended to call NL while Reader B called GG. Similarly, for GG versus PF, Reader A tended to call PF while Reader B called GG. Lastly, for PF versus HC, Reader A tended to call HC while Reader B tended to call PF. These areas of disagreement are consistent with the observation that normal lung, ground-glass opacity, and fibrosis (and potentially honeycombing) represent a continuum of gradual change in interstitial lung disease.

Because of the strategy of keeping only the agreement subset of VOIs, these areas of disagreement are discarded, leaving only the most clear, unambiguous VOIs for classifier training and evaluation. The advantage of this approach is that there is no possibility of confounding the classifier with examples which are disputed by the expert readers. At the same time, however, the classifier is denied the opportunity to learn from the differences in judgment between the readers, leaving no insight as to how to correctly discriminate between classes inside of these “in-between” zones. This observation is illustrated qualitatively in Fig. A.2.

Table A.1. Reader agreement confusion matrices for training and testing datasets

Training dataset

Reader A \ Reader B	PF	GG	HC	NL	AIR	VES	Total
PF	564	74	11	1	0	2	652
GG	3	272	0	25	4	3	307
HC	49	2	42	0	1	1	95
NL	2	82	1	272	10	8	375
AIR	3	2	0	6	294	9	314
VES	1	4	0	5	13	354	377
Total	622	436	54	309	322	377	2120

Testing dataset

Reader A \ Reader B	PF	GG	HC	NL	AIR	VES	Total
PF	433	55	10	5	3	2	508
GG	6	266	1	10	1	0	284
HC	54	1	115	0	0	0	170
NL	0	120	0	206	4	3	333
AIR	0	3	0	1	314	15	333
VES	1	4	0	4	20	311	340
Total	494	449	126	226	342	331	1968

Table A.2. Cohen's kappa measures of reader agreement for training and testing datasets

Class	Training	Testing
PF	0.836	0.818
GG	0.677	0.667
HC	0.549	0.759
NL	0.756	0.695
AIR	0.911	0.916
VES	0.926	0.912
Overall	0.810	0.800

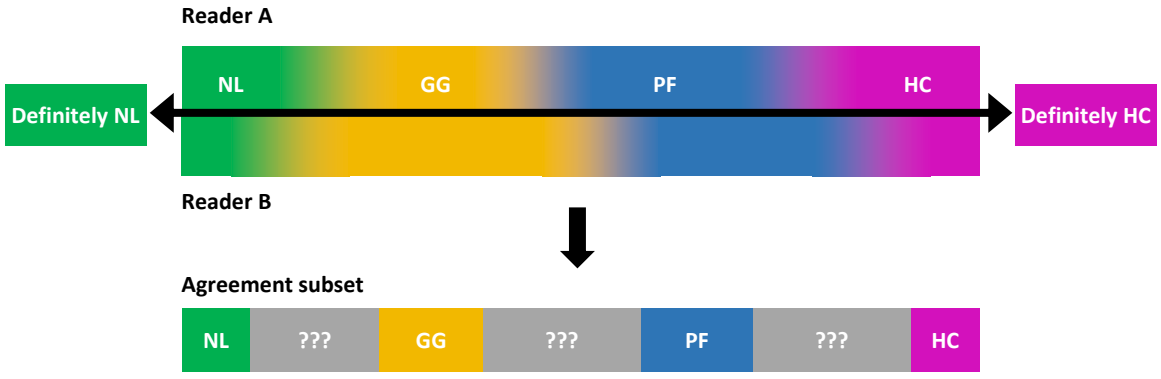

Figure A.2. Illustration qualitatively depicting the effect of discarding disagreement VOIs on the composition of the agreement subset.

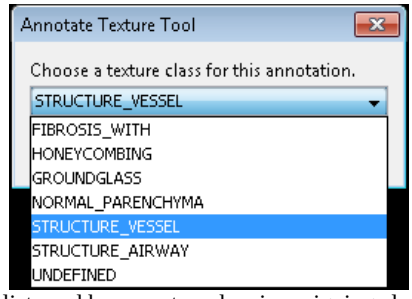

Figure A.3. Screenshot of the drop-down list used by expert readers in assigning class labels to VOIs. A misclick may result in the assignment of an unintended class label.

Another area of disagreement to consider is AIR versus VES. Compared to the disagreements discussed previously, the number of disagreements between AIR and VES is relatively small. Nevertheless, this particular disagreement is significant because it is unlikely to arise from a difference in judgment between the two expert readers. Rather, we suspect that disagreements between AIR and VES arise due to user error. The labeling task calls for the reader to select the class label from a drop-down list, and a simple misclick can result in unintended assignment (Fig. A.3). The presence of user error between AIR and VES implies that other errors in class label assignment must exist as well.

Two types of misclicks can occur when assigning class labels: unintentional agreements and unintentional disagreements. Disagreements are less harmful to classifier training or evaluation because they are simply discarded under the strategy of keeping only the agreement subset. Agreements, on the other hand, introduce noise in the dataset. Fortunately, there are two reasons why unintentional agreements must be relatively rare. First, unintentional agreement can only occur when the two readers disagree in the first place, and we know from the base rates of reader agreement (84.8% and 83.6%) that this situation arises infrequently. Second, even if a reader intends to disagree, four out of five possible misclicks will still result in disagreement, making unintentional agreement still less likely. Therefore, we conclude that although the class labeling task is subject to user error, the impact of this error on classifier performance is minimal.