Unsupervised and Weakly-Supervised Learning of Localized Texture Patterns of Lung Diseases on Computed Tomography

Jie Yang

Submitted in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Unsupervised and Weakly-Supervised Learning of Localized Texture Patterns of Lung Diseases on Computed Tomography

Jie Yang

Computed tomography (CT) imaging enables *in vivo* assessment of lung parenchyma and several lung diseases. CT scans are key in particular for the diagnosis of 1) chronic obstructive pulmonary disease (COPD), which is the fourth leading cause of death worldwide, and largely overlaps with pulmonary emphysema; and 2) lung cancer, which is the first leading cause of cancer-related death, and manifests in its early stage with the presence of lung nodules.

Most lung CT image analysis methods to-date have relied on supervised learning requiring manually annotated local regions of interest (ROIs), which are slow and labor-intensive to obtain. Machine learning models requiring less or no manual annotations are important for a sustainable development of computer-aided diagnosis (CAD) systems.

This thesis focused on exploiting CT scans for lung disease characterization via two learning strategies: 1) fully unsupervised learning on a very large amount of unannotated image patches to discover novel lung texture patterns for pulmonary emphysema; and 2) weakly-supervised learning to generate voxel-level localization of lung nodules from CT whole-slice labels.

In the first part of this thesis, we proposed an original unsupervised approach to learn emphysema-specific radiological texture patterns. We have designed dedicated spatial and texture features and a two-stage learning strategy incorporating clustering and graph

partitioning. Learning was performed on a cohort of 2,922 high-resolution full-lung CT scans, which included a high prevalence of smokers and COPD subjects. Experiments lead to discovering 10 highly-reproducible spatially-informed lung texture patterns and 6 quantitative emphysema subtypes (QES). Our discovered QES were associated independently with distinct risk of symptoms, physiological changes, exacerbations and mortality. Genome-wide association studies identified loci associated with four subtypes.

Then we designed a deep-learning approach, using unsupervised domain adaptation with adversarial training, to label the QES on cardiac CT scans, which included approximately 70% of the lung. Our proposed method accounted for the differences in CT image qualities, and enabled us to study the progression of QES on a cohort of 17,039 longitudinal cardiac and full-lung CT scans.

Overall, the discovered QES provide novel emphysema sub-phenotyping that may facilitate future study of emphysema development, understanding the stages of COPD and the design of personalized therapies.

In the second part of the thesis, we have designed a deep-learning method for lung nodule detection with weak labels, using classification convolutional neural networks (CNNs) with skip-connections to generate high-quality discriminative class activation maps, and a novel candidate screening framework to reduce the number of false positives. Given that the vast majority of annotated nodules are benign, we further exploited a data augmentation framework with a generative adversarial network (GAN) to address the issue of data imbalance for lung cancer prediction. Our weakly-supervised lung nodule detection on 1,000s CT scans achieved competitive performance compared to a fully-supervised method, while requiring 100 times less annotations. Our data augmentation

framework enabled synthesizing nodules with high fidelity in specified categories, and is beneficial for predicting nodule malignancy scores and hence improving the accuracy / reliability of lung cancer screening.

Contents

List of Figures

List of Tables

Acknowledgements

This work was supported by NIH R01-HL121270. Data used from the MESA Lung Study is funded by NIH/NHLBI R01-HL130506, R01-HL077612, R01-HL093081, R01-HL112986, RC1HL100543, RD831697, N01-HC-95159, N01-HC-95160, N01-HC-95169, N01-HC-95160, N01-HC-95162, N01-HC-95164, N01-HC-95164, N01-HC-95169, U01-HL114494; and from N01-HC95159 to HC95169. Data used from SPIROMICS was supported by contracts from NIH/NHLBI, from HHSN268200900013C to HHSN268200900020C, which were supplemented by contributions made through the Foundation for the NIH and COPD Foundation from AstraZeneca; Bellerophon Pharmaceuticals; Boehringer-Ingelheim Pharmaceuticals, Inc; Chiesi Farmaceutici SpA; Forest Research Institute, Inc; GSK; Grifols Therapeutics, Inc; Ikaria, Inc; Nycomed GmbH; Takeda Pharmaceutical Company; Novartis Pharmaceuticals Corporation; Regeneron Pharmaceuticals, Inc; and Sanofi.

The author thank the investigators, the staff, and the participants of the MESA Lung Study and SPIROMICS for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://NES.mesa-nhlbi.org. The authors thank the SPIROMICS participants and participating physicians, investigators and staff for making this research possible. More information about the study and how to access SPIROMICS data is at NES.spiromics.org. We would like to acknowledge the following

current and former investigators of the SPIROMICS sites and reading centers: Neil E Alexis, PhD; Wayne NES Anderson, PhD; R Graham Barr, MD, DrPH; Eugene R Bleecker, MD; Richard C Boucher, MD; Russell P Bowler, MD, PhD; Elizabeth E Carretta, MPH; Stephanie A Christenson, MD; Alejandro P Comellas, MD; Christopher B Cooper, MD, PhD; David J Couper, PhD; Gerard J Criner, MD; Ronald G Crystal, MD; Jeffrey L Curtis, MD; Claire M Doerschuk, MD; Mark T Dransfield, MD; Christine M Freeman, PhD; MeiLan K Han, MD, MS; Nadia N Hansel, MD, MPH; Annette T Hastie, PhD; Eric A Hoffman, PhD; Robert J Kaner, MD; Richard E Kanner, MD; Eric C Kleerup, MD; Jerry A Krishnan, MD, PhD; Lisa M LaVange, PhD; Stephen C Lazarus, MD; Fernando J Martinez, MD, MS; Deborah A Meyers, PhD; John D Newell Jr, MD; Elizabeth C Oelsner, MD, MPH; Wanda K O'Neal, PhD; Robert Paine, III, MD; Nirupama Putcha, MD, MHS; Stephen I. Rennard, MD; Donald P Tashkin, MD; Mary Beth Scholand, MD; J Michael Wells, MD; Robert A Wise, MD; and Prescott G Woodruff, MD, MPH. The project officers from the Lung Division of the National Heart, Lung, and Blood Institute were Lisa Postow, PhD, Thomas Croxton, PhD, MD, and Antonello Punturieri, MD, PhD.

Preface

This dissertation represents the culmination of dedicated collaboration, support and encouragement from many exceptional individuals. The doctoral program has given me the opportunity to work with wonderful people from numerous research groups, and has lead to one of the most exciting and important periods in my life. For that, I would like to thank all the people who have made my dissertation work possible.

First, I would like to thank Dr. Andrew F. Laine for providing me with the opportunity of conducting research at Heffner Biomedical Imaging Lab at Columbia University. The help, guidance, and mentorship from Dr. Laine have always been important to me. Thank you, Dr. Laine, for encouraging me to speak out when I was shy at the beginning; thank you for advising me to explore projects across domains to build up a wide skill-set; and thank you for endorsing my ideas and supporting me to configure GPU station for deep learning without a hesitation. I benefit from all these help in my Ph.D. journey and life.

I would like to thank Dr. Elsa D. Angelini for teaching me how to conduct research projects and perform scientific writing. Elsa, your systematic thinking, and attention to details have always impressed me. I learned a great deal of knowledge from you. There were numerous moments when I got stuck into research questions, and I would always appreciate the fact that I could just go and discuss with you. I enjoyed all of our discus-

sions, and constantly feel grateful for the timely response and advice from super Elsa.

I would like to thank Dr. R. Graham Barr for the invaluable feedback, clinical insights, and data access. Being my most important clinical collaborator, Dr. Barr led me to see the fabulous clinical society, and also provided many constructive suggestions on the technical aspects of my work. Thank you, Dr. Barr, for the infinite positivity and enthusiasm for the emphysema projects. Your encouragement has always been a great motivation for my research work.

I would also like to thank our collaborators from the Columbia University Irving Medical Center who helped me analyze data and discuss results. Particularly, I would like to thank Dr. Pallavi P. Balte for the well-structured analysis and the clear explanations on the clinical context to me. Pallavi, it has been a great pleasure to work with you on the emphysema projects. Also, I would like to thank Dr. John H.M. Austin, for teaching me how to read emphysema on CT, for lending me the books, and for the encouragement and help when we prepared publications; thank Dr. Benjamin M. Smith, for providing the data and clinical insights from his previous work; thank Dr. Wei Shen, for the help and suggestions, and our enjoyable ATS stay at San Diego; and thank Dr. Yifei Sun, for providing statistical insights, and always being helpful and accessible.

Thanks to collaborators from other institutes who have been directly involved with my work. The expertise and knowledge of Dr. Eric A. Hoffman from the University of Iowa have been valuable to me in planning research and writing publications. Mark Escher from Dr. Hoffman's group has kindly iterated with me to transfer high quality data and preprocessing results that have enabled efficient research on my side. Dr. Ani Manichaikul from University of Virginia helped us to perform GWAS analysis and enabled

us to study the genetic factors in the emphysema projects.

I have had the privilege of getting to know great friends at Heffner Biomedical Imaging Lab. First, I would like to thank Dr. Yrjö Häme, for establishing solid previous work for the emphysema projects. Yrjö gave me plenty of suggestions and explanations so that I can start from the right track. Many thanks to Dr. Guillaume David, Dr. Viktor Gamarnik, Dr. Frank Provenzano, Dr. Jia Guo and Dr. Arthur Mikhno. They are wonderful senior Ph.D. students in the lab, and have given me numerous helpful suggestions and fun talks regarding doctoral study and life. I would also like to thank Dr. Ming Jack Po. Jack graduated before I came to the lab. We met very occasionally in life but he in fact provided me with most constructive suggestions when I sought for my future career. Thanks Dr. Jingkuan Song and Dr. Yu Gan. They are experienced researchers and guided me to view machine learning from broader perspectives. Thanks Thomas Vetterli. He established helpful data processing pipeline and gave me many inspirations in the domain adaptation project. A good dose of fun is always needed. For that, I would like to thank Zhuowei Li and Xin Yu, for the amusing coffee breaks and anxiety relieving activities.

My internship at Siemens Corporate Research provided me the opportunity to learn from fantastic personnels. Particularly, I would like to thank Dr. Sasa Grbic, my manager, for recognizing my capability and alway being supportive and encouraging. And I would like to thank my great colleagues, Dr. Siqi Liu and Dr. Zhoubing Xu, for the engaging discussions, the brainstorming, and the educational lunch breaks; thank Dr. Bernhard Geiger, for the amazing annotation tools, and the patience when we iterated on the datasets; and thank Dr. Eli Gibson, for the enlightening and constructive suggestions and always being rigorous in scientific thinking.

I would like to thank the members of my proposal and dissertation committees: Dr. Paul Sajda for chairing the dissertation defense and for his constructive critique, and Dr. Andreas H. Hielscher for his valuable advice and feedback. I learned a great deal of knowledge from Dr. Sajda's and Dr. Hielscher's courses, and greatly appreciate their timely response, support and encouragement for my dissertation.

I would like to thank my family for the endless support that they have given me throughout my life. Thank you, Papa and Mama, for teaching me to be brave, to stay one step ahead in life, and to eventually be a useful person in the society.

And finally, thank you, Xinyang, for making me laugh every day; thank you for your incredible patience and accompany; and thank you for teaching me critical thinking, and for all the memorable days. It was such a fortune that we witnessed each other's doctoral life, and that we strived together to become better ourselves.

I owe a great deal of gratitude to each and everyone mentioned above. I thank these people from the bottom of my heart.

New York, NY, Dec 7th, 2018

Jie Yang

Chapter 1

Introduction

1.1 Computed Tomography for Lung Imaging

X-ray is electromagnetic radiation that can traverse through relatively thick objects (Klug and Alexander 1974). Radiography is an imaging technique that uses X-rays to view the internal structure of a subject. A radiography machine typically consists of an X-ray generator and sensors, between which the subjects are placed. The X-rays are absorbed by the tissues while they pass through the subject. Soft tissue (e.g., muscle) absorbs fewer X-rays than hard tissue (e.g., bone). The varying energy patterns which were not absorbed by the subject are detected by the sensors and a projection image is obtained.

Computed tomography (CT) is a specific radiography imaging procedure that creates volumetric scans of areas inside the body. In a modern CT scanner, the X-ray generator and sensors continuously rotate around the subject while the subject slides through the scanner. Thereafter, a series of 2D CT slices are reconstructed from the projection images. The 2D slices form a 3D scan and can be visualized in the three orthogonal planes.

The value of a voxel in a CT scan represents the radiodensity of a tissue and is measured on the Hounsfield unit (HU). In a voxel with a mean attenuation coefficient μ, the

corresponding HU value is:

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

where μ_{water} and μ_{air} are the attenuation coefficients of water and air. A radiodensity of distilled water at standard temperature and pressure (STP) is defined as 0 HU while the air at STP is defined as -1,000 HU (Buzug 2008).

The lungs are well suited to be imaged with CT, because they consist of air with density values close to -1,000 HU, and other tissues with higher density values and thus exhibiting large intensity contrast. The development of CT imaging provides clinicians with high-quality information of the lung parenchyma and related pulmonary pathologies. Since the introduction of the first commercially available systems in the 1980s, CT has enabled *in vivo* assessment of lung diseases at the macroscopic level. Modern multidetector-row CT (MDCT) scanners enable fast imaging ($<$12 s), so that the entire lungs can be imaged in a single breath-hold (Hoffman, Simon, and McLennan 2006).

1.2 COPD and Pulmonary Emphysema

Pulmonary emphysema is defined morphologically by the enlargement of airspaces with destruction of alveolar walls distal to the terminal bronchioles (Mets et al. 2012). Emphysematous lung destruction decreases the elastic recoil force that drives air out of the lung, causing a reduction in the maximum expiratory flow (Hogg 2004). A mixture of emphysema and small airways disease contributes to chronic airflow limitation, characteristic of chronic obstructive pulmonary disease (COPD). Emphysema and COPD are, jointly, the

fourth leading cause of death in the world in 2017 and are projected to be the third leading cause of death in 2020. More than 3 million people died of COPD in 2012, accounting for 6% of all deaths globally[1].

A major contributor to emphysema is the inhalation of particles from smoking or other sources, causing an inflammatory response in the lungs (Vestbo et al. 2013). A chronic inflammatory response may then induce parenchymal tissue destruction, although the exact mechanism of the process remains unknown. Recent research has associated changes in microvascular blood flow dynamics with structural and physiological changes leading to emphysema (Hoffman, Simon, and McLennan 2006). Emphysema can develop without smoking. At autopsy, pulmonary emphysema occurs in 30% to 50% of cigarette smokers, 8% of cigar smokers and 3% of never-smokers (Auerbach et al. 1972; Leopold and Gough 1957; Thurlbeck 1963). Genetics have been shown to affect the development of the disease. Specifically, alpha1-antitrypsin deficiency has been associated with younger patients ($<$45 yr) and lower lobe emphysema (McElvaney et al. 2017).

Since the alveolar wall destruction in emphysema is irreversible, the disease cannot be fully cured. However, the progression of the disease can be slowed down. Also, for patients with COPD, there are several ways to reduce symptoms. The therapeutic options include smoking cessation, pharmacological therapy, rehabilitation, oxygen therapy, ventilatory support, and surgical treatments. An example of a surgical treatment is lung volume reduction surgery, where parts of the lung are resected to reduce hyperinflation. This operation has been shown to improve survival in some patients with severe upper-lobe emphysema, but it is not suitable for all types of emphysema (Vestbo et al. 2013).

[1] http://www.goldcopd.org/

1.3 Challenges with Emphysema Quantification and Subtyping on CT

An early study of eleven patients (Hayhurst et al. 1984) reported that patients with emphysema on pathology had significantly more low-density values on CT than the group of patients without emphysema. This finding inspired the development of objective and reproducible emphysema quantitation on CT.

The most widely used measure for assessing emphysema severity is obtained using a density measure, called percent emphysema ($\%emph$), also referred to as emphysema index or percent low attenuation area ($\%LAA$), which quantifies the proportion of voxels with intensity values below a fixed threshold within the lung region. The $\%emph$ measure is currently used commonly in clinical studies (Galbán et al. 2012; Gevenois et al. 1995), and it has been shown to be able to predict mortality in COPD (Ceresa et al. 2011). However, there is no consensus on the intensity threshold value that should be used (Mets et al. 2012), and typical threshold values range from −950 to −910 HU (Hoffman, Simon, and McLennan 2006).

Emphysema was subtyped into centrilobular and panlobular emphysema by Leopold and Gough's on 140 autopsies (Leopold and Gough 1957); a third subtype, paraseptal emphysema, was reported on only two autopsies (Edge, Simon, and Reid 1966). The three emphysema subtypes can be visually assessed on lung CT images, using the following definitions:

1. *Centrilobular emphysema* (CLE), which is commonly characterized by low-attenuation

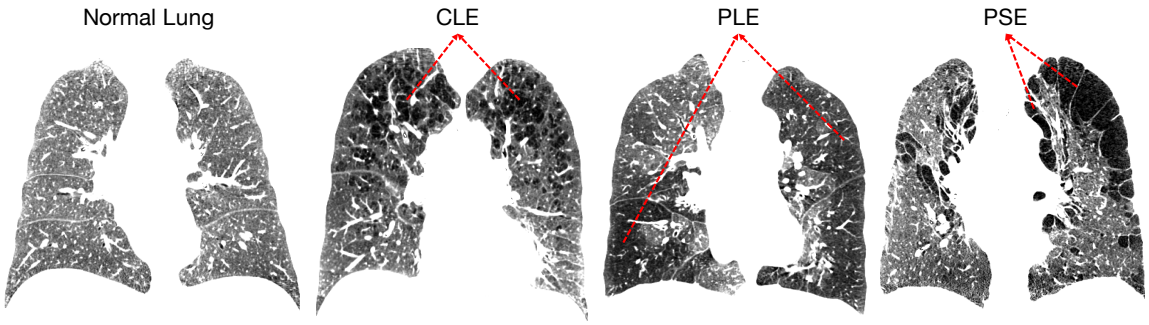

Figure 1.1: Illustration of the three standard emphysema subtypes on CT. From left to right: coronal views of a normal lung, a sample lung predominantly affected by centrilobular emphysema (CLE), panlobular emphysema (PLE), and paraseptal emphysema (PSE). The intensity window is [-1000, -700] HU.

regions surrounded by normal lung attenuation, and located centrally in the secondary pulmonary lobules (Lynch et al. 2015). Classically, its distribution is predominantly in the apical regions of the lungs;

2. *Panlobular emphysema* (PLE), which is commonly characterized by low-attenuation regions uniformly diffused in the secondary pulmonary lobules (Smith et al. 2014), and is associated with alpha1-antitrypsin deficiency. Classically, its distribution is predominantly in the basal regions of the lungs;
3. *Paraseptal emphysema* (PSE), which is commonly characterized by low-attenuation regions adjacent to pleura and to intact interlobular septa, typically found in juxtapleural lobules adjacent to mediastinal and costal pleura (Lynch et al. 2015). Classically, its distribution is predominantly in the upper and middle lung zones.

Illustrations of coronal views of a normal lung CT scan, and CT scans predominantly affected by the three emphysema subtypes are provided in Fig. 1.1.

A previous study (Smith et al. 2014) evaluated emphysema subtypes that were assessed visually on 321 CT scans by multiple readers. The study found that on patients with any

type of emphysema, 57% had multiple subtypes present, with CLE and PSE appearing together most frequently. Compared to controls, patients with CLE and PLE had greater dyspnea, reduced walk distance, greater hyperinflation and lower diffusing capacity, but patients with PSE were similar to controls. CLE was associated with an extensive smoking history, but the other two subtypes were not. Only PLE was associated with reduced body mass index. In addition, 17% of smokers without COPD on spirometry had emphysema.

Given differing risk factors (Dahl et al. 2002; Shapiro 2000), it is likely that the three standard emphysema subtypes represent different diseases; however, pathologists disagreed on the very existence of them (Anderson et al. 1964) and the large study on emphysema - of 1,800 autopsies - ignored them completely, largely for practical reasons (Auerbach et al. 1972). Moreover, radiologists' interpretation of these subtypes on CT images is labor-intensive, with substantial intra- and inter-rater variability, even in expert hands (Barr et al. 2012). Basic emphysema quantification methods (e.g. thresholding based $\%emph$) provide reproducible measures of emphysema in population study (Hoffman et al. 2009), but discards most information in individual scans and provides limited information on emphysema subtypes.

1.4 Lung Cancer and Pulmonary Nodule

Lung cancer is the leading cause of cancer death worldwide. It is the second most common cancer in men, after prostate cancer, and in women, after breast cancer (Siegel, Miller, and Jemal 2016). It is estimated that 222,500 new cases of lung cancer will be diagnosed in 2017. With an approximated 155,870 deaths, lung cancer accounts for 1 in 4 mortalities caused by cancer. The 5-year survival rate of subjects diagnosed with lung cancer is only 18.1%

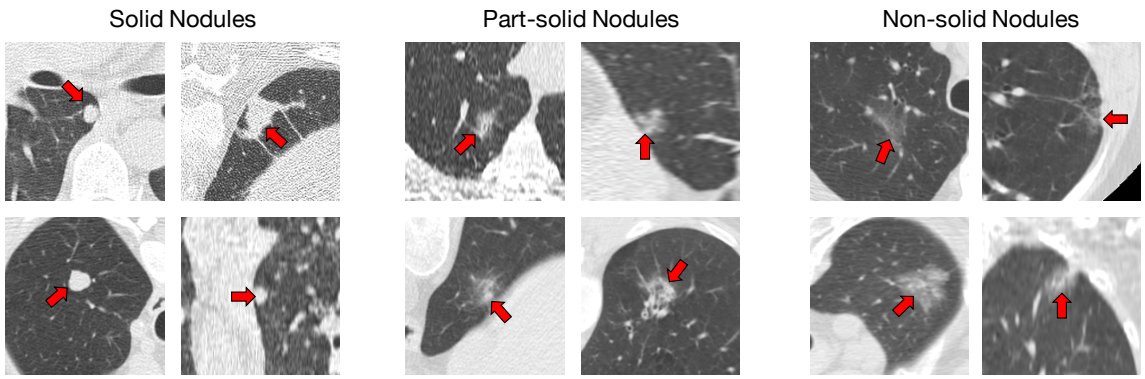

Figure 1.2: Illustration of the three main categories of lung nodules on CT. From left to right: four axial patches with solid nodules, part-solid nodules, and non-solid nodules. The figure is adapted from (Setio et al. 2016).

(Henschke et al. 1999).

The stage of the cancer at diagnosis determines treatment options, and is strongly correlated to survival rate. Most of the lung cancers are diagnosed at a late stage (57%), by which they have already metastasized (5-year survival rate is 4.5%) (Siegel, Miller, and Jemal 2016). The diagnosis is usually only made at a late stage because symptoms, such as a persistent cough, sputum with blood, chest pain, or recurrent pneumonia, typically do not occur until the cancer is already several centimeters in size (Ellis and Vandermeer 2011). Only when lung cancers are diagnosed at a localized stage, treatment options are better, and the 5-year survival rate is 55%. Therefore, to reduce the high mortality rate, there is a strong need to detect subjects with lung cancer as early as possible.

Early stage lung cancer generally manifests in the form of pulmonary nodules. A pulmonary nodule is defined as a rounded opacity, well or poorly defined, measuring up to 3 cm in diameter (Raghu et al. 2011). They can be grouped into three main categories:

1. *Solid nodules*: nodules with homogeneous soft tissue attenuation;
2. *Part-solid nodules*: also known as ground-glass nodules, manifest as hazy increased

attenuation in the lung that does not obliterate the bronchial and vascular margins;

3. *Non-solid nodules*: consist of both ground-glass regions and a solid core with soft-tissue attenuation.

Illustrations of the three main categories of lung nodules on CT are provided in Fig. 1.2. Predictors of cancer include larger nodule size, part-solid nodule type, upper lobe location, spiculated morphology, and presence of emphysema (Horeweg et al. 2014). Automated detection and prediction systems that locate and classify nodules of various sizes can assist radiologists in lung cancer diagnosis, and can therefore facilitate early lung cancer detection and timely surgical intervention (Setio et al. 2016).

1.5 Challenges with Lung Cancer Screening

Lung cancer screening has been approved and is being implemented in the United States. However, challenges present and discussions remain about the cost-effectiveness of lung cancer screening. Three major challenges were identified.

The first challenge is the need to screen many high-risk individuals. Using the recommended screening criteria, it is estimated that 8.6 million Americans are potentially eligible for screening (Siegel, Miller, and Jemal 2016). In addition, the interpretation of lung CT screening scans is tedious, error-prone, and can take up to 10 minutes per scan (Rubin et al. 2005). Therefore, national lung cancer screening can lead to a substantial increase in reading efforts for radiologists.

The second challenge is the high rate of false positives. In the the large National Lung Screening Trial (NLST) (National Lung Screening Trial Research Team 2011), the vast

majority of the nodules identified to be potentially cancerous were eventually benign. A total of 96.4% of positive examinations and 24.2% of all examinations did not result in a lung cancer diagnosis.

The third challenge is the inter-rater variability among radiologists. In the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) study (Armato III et al. 2011), four radiologists reviewed 1,018 CT scans and marked nodules larger than 3 mm. It was shown that complete agreement on what should be considered as a nodule between all four radiologists was only reached on 928 out of 2,669 nodules. Substantial variability in radiologists' false positives rate due to difference in interpretations was also found in a retrospective analysis of the NLST data (Pinsky et al. 2013).

To enable the implementation of a cost-effective lung cancer screening program, an accurate and robust interpretation of the large volume of CT scans is needed. To minimize the number of biopsies, it would be helpful if the lung nodules can be robustly detected and their malignancy can be predicted from the CT scans.

1.6 Proposed Alternatives

1.6.1 Unsupervised and Weakly-Supervised Machine Learning

Texture analysis on CT has received increasing interest recently (Depeursinge et al. 2014) for computer-aid diagnosis (CAD) of lung diseases. However, most existing methods are limited to supervised approaches (Anthimopoulos et al. 2016; Gangeh et al. 2010; Sørensen, Shaker, and De Bruijne 2010; Xu et al. 2006) relying on manually annotated regions of interest (ROIs) as ground truth, which are slow and labor-intensive to obtain.

In the era of big data, relying on massive ground truth labels to develop machine learning algorithms will become less practical, as expert time is scarce and expensive and as scanners continue to evolve significantly. A machine learning model requiring less annotations and suitable for heterogeneous scans is key for a sustainable development of CAD systems.

In this thesis, we developed unsupervised (without annotation) and weakly-supervised (with weak annotations, such as binary labels indicating the presence of disease tissue in a large field of view, rather than voxel-level delineation) methods for lung texture learning, to enable the usage of a vast amount of unannotated and weakly annotated CT scans. More specifically, we used unsupervised machine learning and discovered novel quantitive emphysema subtypes that went beyond the current definition of three standard emphysema subtypes for better understanding of the disease. And we used weakly-supervised machine learning for lung nodule detection, to address the high expense and reading efforts for more cost-effectiveness of lung cancer screening.

1.6.2 Spatial Information for Lung Texture Learning

Preliminary CT-based clinical studies suggest that regional analysis will be instrumental in advancing the understanding of multiple pulmonary diseases (Murphy et al. 2012).

In the case of pulmonary emphysema, it is suspected that different emphysema subtypes affect the lungs in preferential anatomical regions. But epidemiological understanding of how many subtypes exist, how they evolve in time and how they vary with spatial localization is still unsolved. Categorization of emphysema on CT images to date has relied on analysis of local textural patterns, using gray-level intensity-based features

(Binder et al. 2016; Gangeh et al. 2010; Sørensen, Shaker, and De Bruijne 2010; Xu et al. 2006), without consideration of spatial localization.

In the case of pulmonary nodules, their location has been shown to be a useful predictor of malignancy (Swensen et al. 1997), in addition to texture-based and shape-based features. Incorporating radiological predictors for quantitative lung cancer prediction is an active research field (Liu et al. 2017).

In this work, towards better understanding the importance of disease localization, we proposed a standardized lung shape spatial mapping, and incorporated the spatial information for novel lung texture learning and analysis, with specific validations on pulmonary emphysema and nodules.

1.6.3 Usage of Large-Scale Longitudinal Cardiac CT Datasets for Emphysema Quantification

Cardiac CT scans, which are commonly used for the assessment of coronary artery calcium scores to predict cardiac events (Detrano et al. 2008), include approximately 70% of the lungs. Despite missing apical regions, emphysema quantification on cardiac CT was shown to have high reproducibility, high correlation with full-lung measures (Hoffman et al. 2009), and correlate well with risk factors of lung disease and mortality (Oelsner et al. 2014), in *population-based* studies.

Large datasets of cardiac CT scans are readily available. The longitudinal cohort Multi-Ethnic Study of Atherosclerosis (MESA, 2000-2012) (Bild et al. 2002) study contains more than 20,000 cardiac scans, providing an invaluable opportunity for a large-scale longitudinal evaluation of emphysema quantification and texture learning, as the participants in

MESA also underwent gold-standard full-lung scanning in the most recent follow-up visit (Exam 5, 2010-2012).

However, MESA cardiac CT scans involve heterogeneous scanner types, and their imaging protocols are different from full-lung scans, which can cause variations in the image intensity distribution and texture appearance, and thus hinder the visual characteristics of texture patterns in emphysema-like lung.

In this work, in addition to exploiting gold-standard full-lung CT scans, we utilized the large and longitudinal cardiac CT dataset in MESA, and developed robust emphysema segmentation and texture learning methods, accounting for variabilities across image domains and subjects. The proposed work enabled us to study CT image patterns of emphysema on cardiac scans from different sites, and over 10 years of longitudinal follow-up data, and would potentially advance the understanding of progression of emphysema and emphysema subtypes.

1.7 Potential Impact and Thesis Outline

The aim of this work is to significantly advance the CT-based lung texture learning methods by: 1) Exploiting unsupervised and weakly-supervised learning requiring less (or no) annotations; 2) Incorporating spatial information to study lung disease locations; 3) Extending lung texture learning to large cardiac CT datasets for longitudinal study. Our work focused on the understanding and diagnosis of two major types of lung abnormalities: emphysema (associated with COPD) and nodule (associated with lung cancer).

The proposed unsupervised learning method enabled us to discover a set of novel quantitative emphysema subtypes that were highly-reproducible. Going beyond the cur-

rent definitions of three standard emphysema subtypes, which provide an imprecise and non-biologically based disease definition that prevents the development of effective prevention and treatment strategies for COPD, our novel radiological emphysema subtypes have distinct CT representations and structures, are associated independently with unique patterns of respiratory symptoms and clinical events, have varying physiologic characteristics, and may have non-overlapping genetic associations, hence may facilitate personalized therapies.

The proposed methods for emphysema quantification and texture learning on cardiac CT scans accounted for the domain differences in CT imaging protocols and qualities, and would enable large-scale longitudinal studies over 10 years of follow-up, for better understanding of the disease progression. To our knowledge, this is the first study on longitudinal subtyping of emphysema patterns on cardiac CT scans.

The proposed weakly-supervised learning for lung nodule detection achieved competitive performance compared to a fully-supervised method, yet requiring 100 times less annotations. Based on that, we further proposed novel method to estimate lung cancer risks with scan-level diagnostic labels, which are easy to acquire in clinical scenarios. Automated methods that enable estimating early lung cancer risks based on CT images are important for a more cost-effective lung cancer screening program with less reading efforts and minimal number of biopsies.

Overall, the proposed work would enable the usage of a vast amount of unannotated and weakly annotated CT scans. Successful applications would potentially have a tremendous impact in the field, for diseases that affects millions around the world.

Hence, this thesis is organized to present the three main components: 1) Unsuper-

vised learning to discover novel quantitative emphysema subtype on CT; 2) Extending emphysema lung texture learning to cardiac CT scans; 3) Weakly-supervised learning for lung nodule detection and lung cancer prediction.

In Chapter 2, we will overview the CT datasets that are used in this thesis. In Chapter 3, we will present the unsupervised learning framework to discover novel emphysema subtypes, which incorporates spatial and texture features. We call the discovered patterns the spatially-informed lung texture patterns (sLTPs). The proposed method is first evaluated on the MESA COPD dataset (Thomashow et al. 2013) for proof-of-concept, and is then evaluated on a large full-lung CT cohort of COPD and normal controls, SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS) (Couper et al. 2013). To extend emphysema texture learning to the MESA cardiac CT scans, we will first present, in Chapter 4, an emphysema segmentation method based on the hidden Markov measure field (HMMF) model (Häme et al. 2014) to handle scanner and subject variability in MESA. Then we will present, in Chapter 5, a deep-learning method based on unsupervised domain adaptation (Ganin et al. 2016) to learn domain-invariant features across full-lung versus cardiac imaging scanners and protocols (the "domains"). In Chapter 6, we will present a weakly-supervised learning method for lung nodule detection and cancer prediction based on convolutional neural networks (CNNs). Finally, Chapter 7 will provide the summary and discussions, and Chapter 8 will present the publication list and competition performance related to this thesis.

Chapter 2

Data and Preprocessing

2.1 Available CT Data

This work includes CT scans (full-lung and/or cardiac CT) and related demographic/clinical measures from the following cohorts:

1. Multi-ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002), including longitudinal cardiac CT scans at baseline and four follow-up visits, and full-lung CT scans in the most recent visit;
2. MESA COPD (Thomashow et al. 2013), including cross-sectional full-lung CT scans;
3. SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS) (Couper et al. 2013) with cross-sectional full-lung CT scans;
4. The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) (Armato III et al. 2011) with cross-sectional full-lung CT scans;
5. Kaggle Data Science Bowl 2017 (DSB2017)[1] with cross-sectional full-lung CT scans.

We summarize the image data being used in Table 2.1.

The participants in the studies and the imaging protocols used to acquire the CT scans are described in the sections below. More detailed information can be found in their liter-

[1] https://www.kaggle.com/c/data- science-bowl-2017

Table 2.1: Overview of image data used in this work, including three cohorts for emphysema and COPD study (MESA, MESA COPD and SPIROMICS) and two cohorts for nodule and lung cancer study (LIDC-IDRI and Kaggle DSB2017).

Study	N_{ppt}	N_{scan}	N_{type}	N_{year}	Lung Disease
MESA†	6,814	31,228‡ cardiac 3,131 full-lung	11 4	12	Emphysema & COPD
MESA COPD	321	317§ full-lung	3	-	Emphysema & COPD
SPIROMICS	3,200	3,200 full-lung	9	-	Emphysema & COPD
LIDC-IDRI	1,010	1,018 full-lung	17	-	Nodule & Lung Cancer
Kaggle DSB2017	2,101	2,101 full-lung	n/a	-	Nodule & Lung Cancer

N_{ppt} = number of participants in the study;

N_{scan} = number of CT scans;

N_{type} = number of scanner types;

N_{year} = years of follow-up.

† MESA cardiac CT scans were acquired at visits 1-4 in 2000-08 with axial CT scanners, and at visit 5 in 2010-2012 with helical CT scanners; full-lung CT scans were acquired at visit 5 in 2010-12 with helical CT scanners.

‡ Most subjects had two repeated cardiac scans per visit at visit 1-4, while there is one full-lung scan and one cardiac scan in visit 5.

§ Four CT scans are discarded due to data corruption or incomplete lung field of view.

atures for study designs respectively. The subsequent chapters describing the performed studies include additional information that was specific to each experiment.

The CT scans in MESA study, MESA COPD study and SPIROMICS are available in the Heffner Biomedical Imaging Lab. The CT scans in LIDC-IDRI and Kaggle DSB2017 are publicly available online.

2.1.1 MESA

MESA is a prospective cohort study, which recruited 6,814 men and women in 2000-2002 at six US field centers from four racial/ethnic groups, who were aged 45-84 years and free of clinical cardiovascular disease. All 6,814 of these MESA participants underwent cardiac CT scanning at enrollment with either electron beam CT (EBT, Imatron C-150 scanners) or multi-detector CT (MDCT, GE LightSpeed or Siemens S4+ Volume Zoom scanners). Scans were performed under a standardized protocol by designated, MESA-certified, experienced radiology technologists under the supervision of the reading center co-investigator. Axial images were reconstructed with an isotropic pixel resolution in the range [0.44, 0.78] mm, and a slice thickness of 2.5 or 3.0 mm.

The MESA Lung study performed spirometry tests, quantitative lung measures, and assessed cotinine levels on all MESA cardiac CT scans in addition to acquiring gold-standard full-lung CTs for 3,200 participants on 64-slice helical scanners in 2010-12, following the MESA-Lung/SPIROMICS full-inspiration protocol (Sieren et al. 2016). Full-lung images were reconstructed with an in-plane pixel resolution in the range [0.47, 0.92] mm and a slice thickness of 0.625 or 0.75 mm.

2.1.2 MESA COPD

The MESA COPD study includes 321 subjects who were aged 50-79 years, with 10 or more pack-year smoking history and who did not have clinical cardiovascular disease, stage IIIb-V kidney disease, asthma prior to age 45 years, other lung disease, prior lung resection, cancer, allergy to gadolinium, claustrophobia, metal in the body, pregnancy or

weight > 300 lbs. Among the 321 subjects, 192 were recruited from the MESA study, and the rest were recruited from the EMCAP study (Barr et al. 2007), which is a cohort study of smokers. Full-lung CTs were acquired with Siemens and GE 64-slice scanners, at 120 kVp, 0.5 seconds, with 200 mA for the EMCAP participants, and current (mA) set by body mass index (BMI) for the MESA participants, following the MESA-Lung/SPIROMICS full-inspiration protocol (Sieren et al. 2016). Images were reconstructed with an in-plane pixel resolution within the range [0.58, 0.88] mm, and a slice thickness of 0.625 mm.

2.1.3 SPIROMICS

The SPIROMICS recruited 3,200 participants (2,400 patients with COPD, 600 smokers without COPD and 200 non-smokers without COPD) aged 40-80 years old in six US field centers. All participants underwent the CT scan at baseline and one-year follow-up. The lung CT scanning protocol in SPIROMICS is identical to that in MESA Lung. Images were reconstructed with an in-plane pixel resolution within the range [0.48, 0.98] mm, and a slice thickness of 0.625 or 0.75 mm.

2.1.4 LIDC-IDRI

The LIDC-IDRI dataset is a web-accessible international resource for development, training, and evaluation of CAD methods for lung cancer detection and diagnosis. Seven academic centers and eight medical imaging companies collaborated to create this resource. The dataset contains 1018 cases, each of which includes a clinical thoracic CT scan and an associated record of lung nodule annotation process performed by four experienced thoracic radiologists. Four scanner manufacturers and 17 models were represented, and four

types of convolution kernels were used for image reconstruction. The in-plane pixel resolution ranged from 0.46 to 0.98 mm (mean = 0.69 mm), and the slice thicknesses ranged from 0.45 to 5.0 mm (mean = 1.74 mm).

2.1.5 Kaggle DSB2017

The Kaggle DSB2017 dataset is provided by the public Kaggle Data Science Bowl 2017 international competition for early lung cancer detection, which involves 2,101 patients who were high-risk of lung cancer. Each patient is associated with a full-lung CT scan and a pathological diagnosis of with/without early lung cancer. While the CT scans vary in scanner, acquisition time and image quality, only necessary information such as patient ID, in-plane pixel resolution ([0.49, 0.98] mm, mean = 0.68 mm) and slice thickness ([0.625, 3.0] mm, mean = 1.71 mm) is provided to encourage the development of CAD methods that are invariant to acquisition conditions.

2.2 Preprocessing

2.2.1 Lung Mask Segmentation

Along with CT scans in MESA study, MESA COPD study and SPIROMICS study, lung mask files were generated by the APOLLO® software (VIDA Diagnostics, Iowa). For CT scans in LIDC-IDRI and Kaggle DSB2017, lung masks were segmented with a classic lung segmentation method, by:

1. Applying an intensity threshold of -400 HU and locating the largest connected objects in the resulting binary mask (Hu, Hoffman, and Reinhardt 2001);

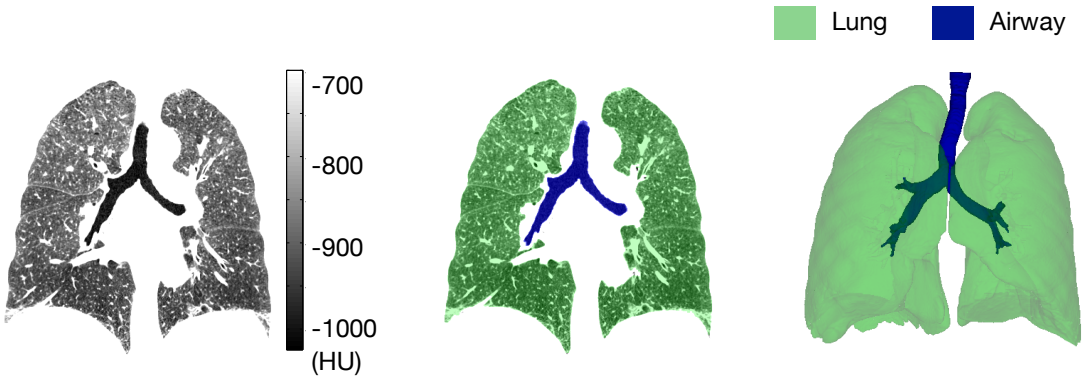

Figure 2.1: Illustration of lung and airway segmentation results on a sample CT scan in the preprocessing stage. (Left) Coronal view of a sample lung CT intensity image; (Middle) Coronal view of the segmentation of lung (green) and airway (blue); (Right) 3D view of the segmentation of lung (green) and airway (blue).

2. Removing trachea and some of the large airways using closed space dilation (Masutani, Masamune, and Dohi 1996).

The segmentation results of lung and airway are illustrated in Fig. 2.1 for a sample full-lung CT scan.

2.2.2 Emphysema Segmentation

In the proposed work, emphysema was segmented from the lungs using two approaches: a traditional thresholding-based approach, and a preliminary segmentation tool based on the hidden Markov measure field (HMMF) (Häme et al. 2014).

In the thresholding-based segmentation, regions of lung with attenuation $<$ -950 HU were segmented as emphysema (denoted as $\%emph_{-950}$). This threshold has been previously validated against autopsy specimens and is commonly used in large clinical studies (Oelsner et al. 2014; Yang et al. 2016a).

The HMMF segmentation (denoted as $\%emph_{\text{HMMF}}$) method is a prior work in the Heffner Biomedical Imaging Lab. It provides robust emphysema segmentation, which has

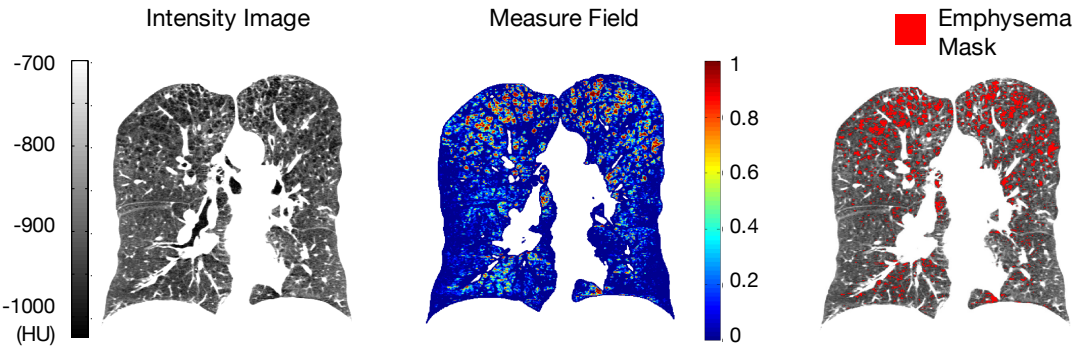

Figure 2.2: Illustration of HMMF-based emphysema segmentation result on a sample full-lung CT scan in the preprocessing stage. (Left) Coronal view of a sample lung CT intensity image; (Middle) Coronal view of the intermediate measure field values, in the range of [0,1], in the HMMF segmentation model; (Right) Coronal view of final HMMF-based binary segmentation result. This figure is adapted from (Häme et al. 2014).

been demonstrated in heterogeneous full-lung CT scans in (Häme et al. 2014). It enforces spatial coherence of the labeled emphysematous regions within neighborhood cliques via Markovian regularization weight, and relies on parametric modeling of intensity distributions within emphysematous and normal lung tissue to adapt to individual and scanner variability. The HMMF model first optimizes an intermediate Markov measure field output, which estimates the probability of each voxel belonging to the emphysema class, and then computes the final emphysema segmentation. A Gaussian distribution is used to characterize the intensity of emphysema class, and a skew-normal distribution is used to characterize the intensity for normal lung tissue. The intermediate Markov measure field values and emphysema segmentation result using HMMF are illustrated in Fig. 2.1 for a sample full-lung CT scan.

In Chapter 4, we will also present a dedicated framework which extends the HMMF-based model for robust emphysema segmentation on heterogeneous and longitudinal cardiac CT scans in MESA.

Chapter 3

Unsupervised Learning of Spatially-Informed Lung Texture Patterns for Pulmonary Emphysema

3.1 Introduction

Pulmonary emphysema is morphologically defined by the enlargement of airspaces with destruction of alveolar walls distal to the terminal bronchioles (Aoshiba, Yokohori, and Nagai 2003). Emphysema overlaps considerably with chronic obstructive pulmonary disease (COPD), which is currently the 4th leading cause of death worldwide, and is projected to be the 3rd leading cause of death in 2020[1].

Based on small autopsy series, pulmonary emphysema is traditionally subcategorized into three standard subtypes: centrilobular emphysema (CLE), panlobular emphysema (PLE) and paraseptal emphysema (PSE). The three standard emphysema subtypes are associated with different risk factors and clinical manifestations (Dahl et al. 2002), and are likely to represent different diseases.

However, given that these subtypes were initially defined at autopsy before the availability of CT scanning, there have been disagreements among pathologists on the very existence of such pure subtypes (Anderson et al. 1964), and a large emphysema study on

[1] http://www.goldcopd.org/

1,800 autopsies (Auerbach et al. 1972) ignored them completely, mainly for practical reasons. Radiologists' interpretation of these subtypes on CT scans is labor-intensive, with substantial intra- and inter-rater variability (Smith et al. 2014).

Automated CT-based analysis enables *in vivo* study of emphysema patterns, and has received increasing interest recently (Depeursinge et al. 2014; Mets et al. 2012), either via supervised learning for replicating emphysema subtype labeling as in (Asherov, Diamant, and Greenspan 2014; Gangeh et al. 2010; Ginsburg et al. 2012; Sørensen, Shaker, and De Bruijne 2010), or via unsupervised learning for the discovery of new emphysema subtypes as in (Binder et al. 2016; Häme et al. 2015b; Yang et al. 2016b).

Preliminary CT-based clinical studies suggest that regional analysis will be instrumental in advancing the understanding of multiple pulmonary diseases (Murphy et al. 2012). In the case of pulmonary emphysema, it is suspected that different subtypes of emphysema affect the lungs in preferred anatomical region. But physiological understanding of how many subtypes exist, how they evolve in time and how they vary with spatial location is still unsolved.

To date, categorization of emphysema on CT images has relied only on analysis of local textural patterns, using either grey-level co-occurrence matrix (GLCM) features (Ginsburg et al. 2012), texton features (Gangeh et al. 2010), or local binary pattern (LBP) features (Sørensen, Shaker, and De Bruijne 2010). All these approaches use intensity information without consideration of spatial location.

In two previous studies (Häme et al. 2015b; Yang et al. 2016b), we proposed to use local textural patterns to generate unsupervised lung texture patterns (LTPs) followed by LTP-grouping based on their spatial co-occurrence in local neighborhoods. Such separate use

of intensity and spatial information cannot guarantee spatial and textural homogeneity of the final LTPs. Therefore, we propose to perform discovery of LTPs via unsupervised clustering of joint spatial and textural information of local patterns on CT. Spatial information can be inferred from crude partitioning of the lung with subdivisions of Cartesian coordinates or by segmenting the lung into zones (e.g. upper, lower) (Smith et al. 2014) or lobes (Hoffman et al. 2003). However, such approaches have limited spatial precision and lack relative information such as peripheral versus central positioning, which is important in defining paraseptal emphysema and subpleural bullae.

In this work, we first propose a new standardized lung shape spatial mapping, called Poisson distance conformal mapping (PDCM), which enables detailed, precise and standardized mapping of voxel positions with respect to the lung surfaces. And we exploit the proposed mapping for the study of emphysema spatial patterns across populations of CLE-, PLE- and PSE-predominant subjects, without registration being required besides orientation alignment. Then we propose a two-stage unsupervised learning framework to discover *emphysema-specific* lung texture patterns, which we call the spatially-informed LTPs (sLTPs).

For a proof-of-concept, we first exploit the proposed the method using a cohort of 317 full-lung CT scans from the MESA COPD study (Thomashow et al. 2013), and 22 longitudinal CT scans from the EMCAP study (Barr et al. 2007). The discovered sLTPs are evaluated in terms of their reproducibility, and ability to encode standard emphysema subtypes. Then we apply the unsupervised learning framework to a large cohort, the SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS) (Couper et al. 2013), which contains CT scans of 2,922 individuals of COPD subjects and

normal controls. Then we use the discovered patterns to label all CT scans in SPIROMICS and another large cohort, the MESA Lung Study (Bild et al. 2002), and evaluate the discovered patterns extensively in terms of their associations with respiratory symptoms, clinical events, physiologic characteristics, and genetic variants.

3.2 Method

3.2.1 Overview

The proposed framework is structured in four main steps to model the spatial and texture features within emphysema-like lung, and generate the emphysema-specific sLTPs:

1. Generate spatial mapping of the lung masks: mapping voxels within the lung masks into a custom Poisson distance map (PDM) to encode the "peel to core" distance, and a conformal mapping to distinguish superior versus inferior, anterior versus posterior and medial versus lateral voxel positions;
2. Encode regions of interest (ROIs) within emphysema-like lung: sampling ROIs from emphysema segmentation masks, and generating spatial features (based on spatial mapping) and texture features of each ROI;
3. Discover an initial set of LTPs: clustering training ROIs into a large number of clusters, based on texture features, and then iteratively augment the LTPs with spatial information via customized regularization and penalty terms;
4. Generate the final set of sLTPs: measure the similarity between LTPs in the initial set, and then group similar / redundant LTPs and generate the final set of sLTPs via partitioning the similarity graph.

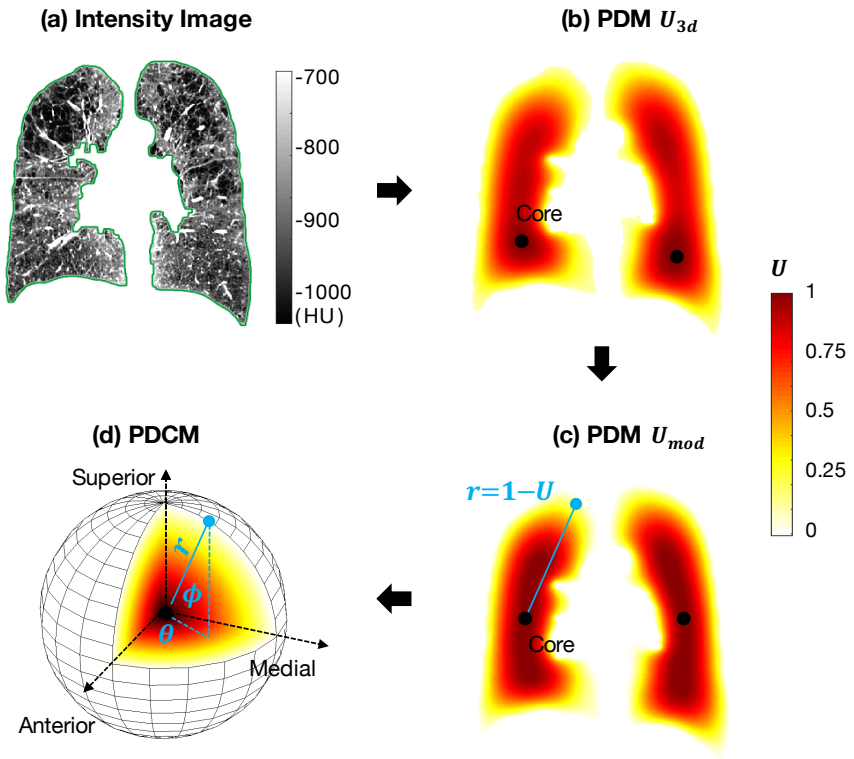

Figure 3.1: Illustration of the proposed lung shape spatial mapping. (a) Coronal slice of a sample CT image (green contour indicates the boundary of lung mask); (b) Corresponding Poisson distance map (PDM) U_{3d} with values in range $[0, 1]$ that measure the "peel to core" distance to the lung mask external surface; (c) Modified PDM U_{mod}, for comparable core locations between subjects; (d) Conformal mapping of the lung PDM to a sphere leading to a Poisson distance conformal map (PDCM) where pixels are assigned three coordinate values (r, θ, ϕ) which enable to distinguish superior vs. inferior, anterior vs. posterior and medial vs. lateral positions, in addition to "peel to core" distance.

We now detail these steps individually in sections below.

3.2.2 Spatial Mapping of the Lung Masks

To generate spatial mapping of the lung masks, we first use the concept of Poisson distance map (PDM), introduced in (Gorelick et al. 2006), to encode the shape of individual lung masks V. PDM is commonly used for characterizing the silhouette of an object via continuous labeling of voxel positions with scalar field values U in the range of $[0, 1]$. In our case, the field value U encodes the "peel to core" distance between a given voxel and

the external lung surface ∂V. This field is computed by solving the following Poisson equation:

$$\begin{aligned} &\Delta U(x,y,z) = -1, \text{ for } (x,y,z) \in V \\ &\text{subject to } U(x,y,z) = 0, \text{ for } (x,y,z) \in \partial V \end{aligned} \tag{3.1}$$

where $\Delta U = U_{xx} + U_{yy} + U_{zz}$.

The solution for U proposed in (Gorelick et al. 2006) is guaranteed to be smooth according to (Haidar et al. 2006). It has the advantage of generating distance values that are sensitive to global shape characteristics, unlike other distance metrics (e.g. Euclidian or Metropolis distances) which exploit single contour points. PDM can therefore reflect rich shape properties of the lung.

The core of the PDM is the set of voxels (one or very few) with the largest U value. The PDM generated from a lung surface generally exhibits nice star-shaped profiles when viewed in axial cuts, with a unique maxima in the center. On the other hand, core positions can vary greatly among subjects along superior-inferior axis, due to variable morphologies of the lungs, especially near the heart and at the base. We illustrate an example in Fig. 3.1 (b) where the PDM generated with Equation (3.1) has core point(s) located very low within the lung rather than concentrated toward the middle of the longitudinal axis. We propose the following approach to calibrate lung PDMs targeting high values of U concentrated near the skeleton of the lung shapes and in the mid-level slices.

We denote $U^{max}(S_i)$ the maximal in-slice value of U, where S_i is the axial slice level with i in ascending order from the apex. We denote $S_{V\%}$ the slice level with $V\%$ of total lung volume above. A normalized version (denoted as U_{2d}), of the original PDM (denoted

as U_{3d}), is then defined, per axial slice S_i, as $U_{2d}(S_i) = U_{3d}(S_i)/U_{3d}^{max}(S_i)$.

We further modify U by combining U_{3d} and U_{2d} values. First, two axial slice levels $S_{i'_u}$ and $S_{i'_d}$, corresponding to the most apical and basal slice levels of local maxima in U_{3d}, are identified as:

$$\begin{aligned} i'_u &= \underset{x}{\text{argmax}}\big[U_{3d}^{max}(S_i) < U_{3d}^{max}(S_x), \forall\, i < x\big] \\ i'_d &= \underset{x}{\text{argmin}}\big[U_{3d}^{max}(S_i) < U_{3d}^{max}(S_x), \forall\, i > x\big] \end{aligned} \tag{3.2}$$

We then define two reference slice levels S_{i_u} and S_{i_d} as:

$$\begin{aligned} S_{i_u} &= \min(S_{25\%}, S_{i'_u}) \\ S_{i_d} &= \max(S_{75\%}, S_{i'_d}) \end{aligned} \tag{3.3}$$

The reference levels S_{i_u} and S_{i_d} are exploited to ensure that the modified core regions reach at least extremal levels $S_{25\%}$ and $S_{75\%}$, with the following modification of the U values into the modified PDM (denoted as U_{mod}):

$$\begin{aligned} U_{mod}(S_i) &= U_{2d}(S_i),\ \forall\, i_u \leqslant i \leqslant i_d \\ U_{mod}(S_i) &= U_{3d}(S_i)/U_{3d}^{max}(S_{i_u}),\ \forall\, i < i_u \\ U_{mod}(S_i) &= U_{3d}(S_i)/U_{3d}^{max}(S_{i_d}),\ \forall\, i > i_d \end{aligned} \tag{3.4}$$

We illustrate in Fig. 3.1 (c) an example of U_{mod} which takes similar maximal values (equal to 1) over a large mid-level extent along the superior-inferior axis and exhibits decreasing values when moving toward the apex or the base of the lung.

This simple calibration enables us to equip the PDM with a coordinate system centered

at a core localized on axial slice level $S_{50\%}$ (ensuring a balanced numbers of voxels above and below), where the core is defined as the point with $U_{mod} = 1$, and closest to the 2D center of mass, for the sake of simplicity.

To uniquely encode 3D voxel positions, we define radial values $r = 1 - U_{mod}$ and add conformal mapping of voxels positions onto a sphere, generating a Poisson distance conformal map (PDCM). We encode superior versus inferior, anterior versus posterior and medial versus lateral voxel positioning via latitude and longitude angles (θ, ϕ) with respect to the PDM core defined above and standard image axis. The generation of the spatial PDCM mapping is illustrated in Fig. 3.1 (d).

The PDCM spatial mapping will be exploited for sLTP learning. Furthermore, we can use PDCM to study population-based spatial distributions of pulmonary diseases. In this chapter, we exploit PDCM to study emphysema spatial location, as reported in Section 3.3.2. Later on in this thesis (Chapter 6 Section 6.3.3), we will also exploit PDCM to study the spatial location of lung nodules.

3.2.3 Texture and Spatial Features

Prior Emphysema Segmentation and ROI Sampling

Texture and spatial analysis is performed within local ROIs centered on a subset of lung voxels. Sampling ROIs from emphysema-like lung requires prior emphysema segmentation. In this work, we exploit two training cohorts (MESA COPD and SPIROMICS) of full-lung CT scans and their associated emphysema masks, which are generated using both a thresholding-based voxel selection and a hidden Markov measure field (HMMF) segmentation (Häme et al. 2014).

For thresholding, voxels with attenuation values below -950 HU are selected. This threshold has been previously validated against autopsy specimens and is commonly used in large clinical studies (Hoffman et al. 2014; Yang et al. 2016a). The HMMF segmentation enforces spatial coherence of the labeled emphysematous regions, and relies on scanner-specific and subject-specific parametric modeling of intensity distributions within emphysematous and normal lung tissues to adapt to individual and scanner variability (more details are in Chapter 4). With the two sets of emphysema masks, percent emphysema measurements quantify the proportion of emphysematous voxels within the lung region, and are denoted $\%emph_{-950}$ and $\%emph_{\text{HMMF}}$.

We experimented several options for ROI sampling in preliminary implementations such as keypoint sampling in (Häme et al. 2015b) and regular sampling in (Yang et al. 2016b). In this study, we use the systematic uniform random sampling (SURS) strategy as suggested in (Puliyakote et al. 2016) for use on lung CT scans. Each individual lung mask is randomly sampled via dividing the bounding box of each lung into 3D regular stacks, and then selecting voxels per stack with a random shift of positions. Two parameters are used for the sampling: β_1 is used for the random shift of positions and β_2 is used to set the number of sampled voxels per stack. The SURS sampling ensures even representation of all lung regions while introducing variability in the position of sampled points with the random shift parameter β_1.

When applying the learning algorithm to the MESA COPD dataset, we select only ROIs with both percent emphysema measures $\%emph_{-950}$ and $\%emph_{\text{HMMF}}$ larger than 1% for training to ensure sufficient representation of emphysematous regions (i.e. each training ROI has a minimal proportion of emphysema but can be a mixture of normal

and emphysematous tissues). The threshold 1% here is a pre-fixed value, that we consider to be sufficient to include early emphysema signals. When applying the algorithm to the SPIROMICS and MESA Lung Study, the ROI selection is based on subject-specific threshold values computed by a reference equation (Hoffman et al. 2014). More details are provided in Section 3.4.

Texture Features

We use texton-based texture features to characterize each ROI, which model texture as the repetition of a few basic primitives (called textons), and were shown to outperform other texture features in unsupervised lung texture learning in (Yang et al. 2016b).

First, we generate small-sized random patches from the training ROIs. A texton codebook is constructed by retaining the cluster centers (textons) of intensity values from those small-sized training patches. The clustering is performed with K-means.

Then, for each ROI, we extract all small-sized patches in a sliding window manner, and compute their voxel intensity distance to the textons. By projecting all small-sized patches of a ROI onto the codebook via searching for the closest textons, the texton-based feature of this ROI is the normalized histogram of texton frequencies.

Spatial Features

To generate spatial features of individual ROIs, we divide the lung masks into lung sub-regions via discretizing our lung shape spatial mapping. For the sake of simplicity, we define lung sub-regions by dividing $r \in [0, 1]$ into 3 regular intervals to distinguish core to peel regions, dividing $\theta \in [0, 2\pi]$ into 4 regular intervals to distinguish anterior, medial, posterior and lateral regions, and dividing $\phi \in [-\pi/2, \pi/2]$ into 3 regular intervals to

distinguish inferior, mid-level and superior regions. The spatial feature of each ROI is a one-hot vector indicating the lung sub-region it belongs to. Ordering of the bins that represent the sub-regions is done via arbitrary spatial rastering as no assumption needs to be made on spatial adjacency of adjacent bins.

3.2.4 Initial Augmented LTPs

Our discovery of spatially-informed lung texture patterns (sLTPs) is formulated as an unsupervised clustering problem. One key factor in unsupervised clustering is the choice of number of clusters. The algorithm is expected to find quantitative emphysema subtypes that are finer-grained than the three standard emphysema subtypes. Therefore, the number of clusters should be large enough to handle the diversity of textures encountered in the lung volumes (i.e. good intra-cluster homogeneity), and on the other hand, be small enough to avoid redundancy (i.e. good inter-cluster differences) for better clinical interpretation. A simple one-stage clustering is suboptimal since it requires tuning or a pre-fixed number of clusters, and may not be able to preserve rare patterns. Therefore, we propose a two-stage learning strategy, where we first generate an empirically large number of fine-grained lung texture patterns (LTPs), and then group similar LTPs to produce the final set of sLTPs, according to a dedicated metric.

LTPs $\{LTP_k\}$ ($\{\cdot\}$ denotes a set of variables hereafter) are characterized by their spatial and texture feature centroids, which are encoded as histograms, and are enforced for intra-class similarity and inter-class separation. For a given LTP_k, its *texture* centroid

Algorithm 1: Generating and Augmenting LTPs

Input : N_{LTP} : Target number of LTPs;

$\{x, FT_x, FS_x\}$: Training ROIs x along with their texture features FT_x and spatial features FS_x.

Output: $\{\overline{FT}_{LTP_k}, \overline{FS}_{LTP_k}\}_{k=1,...,N_{LTP}}$: LTP texture and spatial feature centroids.

Procedure:

- Cluster training ROIs $\{x\}$ into N_{LTP} clusters with $\{FT_x\}$, using K-means.
- Set $t = 0$, and initialize $\Lambda^{(0)}_{LTP_k}$ ($k = 1, ..., N_{LTP}$) with the N_{LTP} LTPs.
- For each k, compute $\overline{FT}^{(0)}_{LTP_k}$, $\overline{FS}^{(0)}_{LTP_k}$ based on $\Lambda^{(0)}_{LTP_k}$.

while $t = 0$ **or** $\{\Lambda^{(t)}_{LTP_k}\} \neq \{\Lambda^{(t-1)}_{LTP_k}\}$ **do**

1. $t = t + 1$;
2. $\{\Lambda^{(t)}_{LTP_k}\} \leftarrow \{\Lambda^{(t)}_{LTP_k}\}^*$ following Equation (3.6);
3. Compute $\{\overline{FT}^{(t)}_{LTP_k}, \overline{FS}^{(t)}_{LTP_k}\}$ based on $\{\Lambda^{(t)}_{LTP_k}\}$.

end

$\overline{FT}_{LTP_k}$ and *spatial* centroid $\overline{FS}_{LTP_k}$ are computed as:

$$\left[\overline{FT}_{LTP_k},\ \overline{FS}_{LTP_k}\right] = \frac{1}{|\Lambda_{LTP_k}|} \sum_{x \in \Lambda_{LTP_k}} \left[FT_x,\ FS_x\right] \tag{3.5}$$

where FT_x and FS_x are respectively the texture feature and spatial feature of a ROI x, and Λ_{LTP_k} denotes the set of ROIs that are labeled as LTP_k.

An initial set of LTPs is generated by clustering with *texture* features, and is then augmented with *spatial* regularizations via iteratively updating $\{\overline{FT}_{LTP_k}, \overline{FS}_{LTP_k}\}$ and $\{\Lambda_{LTP_k}\}$. The generation and augmentation of LTPs are summarized in Algorithm 1.

Designing proper distance metrics for histograms plays a crucial role in many computer vision tasks. Two popular choices are the χ^2 and the ℓ^2 distance metrics. The latter

equally weights distances of all bins and is favored to compare one-hot vectors, while the former is a weighted distance and is favored to compare probability distributions. In our case, texture feature histograms encode distributions over textons, and the χ^2 metric is used. On the other hand, spatial features are sparse one-hot vectors for individual ROIs and we chose the ℓ^2 metric to favor spatial centroids being concentrated in specific lung sub-regions. We therefore propose a mixed χ^2-ℓ^2 similarity metric to enforce spatial concentration of LTPs while preserving their intra-class textural homogeneity:

$$\begin{aligned}\left\{\Lambda_{LTP_k}^{(t)}\right\}_{\{\lambda,W,\gamma\}}^{*} = \underset{\{\Lambda_{LTP_k}^{(t)}\}}{\operatorname{argmin}} \sum_{k} \sum_{x\in\Lambda_{LTP_k}^{(t)}} \\ \chi^2\left(FT_x, \overline{FT}_{LTP_k}^{(t-1)}\right) + \lambda \cdot W \cdot \left|\left|FS_x - \overline{FS}_{LTP_k}^{(t-1)}\right|\right|_2^2 + \\ \gamma \cdot \mathbb{1}\left[\chi^2\left(FT_x, \overline{FT}_{LTP_k}^{(t-1)}\right) > \max_{x'\in\Lambda_{LTP_k}^{(t-1)}} \chi^2\left(FT_{x'}, \overline{FT}_{LTP_k}^{(t-1)}\right)\right]\end{aligned} \tag{3.6}$$

where $\left\{\Lambda_{LTP_k}^{(t)}\right\}_{\{\lambda,W,\gamma\}}^{*}$ denotes the optimal value identified with a set of parameters $\{\lambda, W, \gamma\}$ at iteration t. The first distance metric $\chi^2(\cdot)$ measures the χ^2 distance between the textural feature of a ROI x and LTP_k. The second distance metric $||\cdot||_2^2$ measures the ℓ^2 distance between the spatial feature of a ROI x and LTP_k. A textural penalty term is then introduce as the third term, where $\mathbb{1}$ is the indicator function.

Minimization of Equation (3.6) (step 1 in Algorithm 1) is performed via exhaustive search over all possible values of $\{\Lambda_{LTP_k}^{(t)}\}$. Update of LTP centroids (step 2 in Algorithm 1) is performed after relabeling each ROI to the LTP to which it has the smallest weighted feature distances without turning on the penalty.

Parameter W: This parameter is used to scale contributions between textural distance

and spatial distance terms so that λ can be tuned within a small range of values. We defined it as:

$$W = \frac{SST_T}{SST_S} = \frac{\sum_x \chi^2 \left(FT_x, \sum_x FT_x/N\right)}{\sum_x ||FS_x - \sum_x FS_x/N||_2^2} \tag{3.7}$$

where SST_T and SST_S are respectively the texture and spatial *total* sum-of-square distances, computed on the whole N training ROIs to measure the overall diversity of texture and spatial features.

Parameter λ: This parameter controls the spatial regularization which will inevitably decrease textural homogeneity of individual LTPs. The value of λ is set as follows. First we define SSW_T as the initial sum-of-square *within-cluster* homogeneity of texture features without spatial regularization:

$$SSW_T = \sum\nolimits_k \sum\nolimits_{x \in \Lambda^{(0)}_{LTP_k}} \chi^2 \left(FT_x, \overline{FT}^{(0)}_{LTP_k}\right) \tag{3.8}$$

Then we define SSW_T^λ as the SSW_T measured on augmented LTPs with spatial regularization enforced with $\lambda \in [0, 2]$. Final value of λ is set to:

$$\lambda^* = \underset{\lambda}{\text{argmax}} \left[\Delta SSW_T(\lambda) < L_T\right] \text{ where } \Delta SSW_T(\lambda) = \frac{SSW_T^\lambda - SSW_T}{SSW_T}\% \tag{3.9}$$

In the context of unsupervised discovery, we hereby spatially regularize the augmented LTPs via an empirically acceptable textural homogeneity loss with the threshold L_T (set based on data observations, as reported in Section 3.3).

Parameter γ: This parameter weights the textural penalty term which is used for ROI labeling. We set $\gamma = \infty$ to prevent a ROI from being labeled to a spatially preferred but

texturally dissimilar LTP.

3.2.5 Final Spatially-Informed LTPs (sLTPs)

In this final step, we generate sLTPs by partitioning a weighted undirected graph G where nodes are the N_{LTP} initial augmented LTPs. To define weighted edges between nodes, we rely on replacement tests. We first define N_{LTP} subsets of augmented LTPs as $\{LTP_k\}_{k\neq i}$ (i.e. without LTP_i in the subset of LTPs) for $i = 1, 2, ..., N_{LTP}$. Labeling again all training ROIs with these subsets, we defined N_{LTP} sets of labeled data $\Lambda_{LTP_{i\to j}}$ as the ROIs labeled as LTP_j when using $\{LTP_k\}_{k\neq i}$. In the replacement tests, a ROI with a textural distance to LTP_k exceeding the maximal within-cluster textural distance of LTP_k is not re-labeled. Therefore, defining $N_{i\to j} = |\Lambda_{LTP_{i\to j}}|$, we guarantee that $\sum_k N_{i\to k}/N_i \leqslant 1$ for $N_i = |\Lambda_{LTP_i}|$ when all augmented LTPs are used for labeling. We define similarity weights $G_{i,j}$ as a measure of replacement ratios of LTP_i into LTP_j and vice versa:

$$G_{i,j} = \frac{N_{i\to j} + N_{j\to i}}{N_i + N_j} \cdot E_{i,j} \tag{3.10}$$

The binary variable $E_{i,j}$ controls the existence of an edge between LTP_i and LTP_j. To prevent weak associations of LTPs that are not easily replaceable, we define this binary variable as:

$$E_{i,j} = \mathbb{1}\left(\frac{\sum_k N_{i\to k}}{N_i} > \eta\right) \cdot \mathbb{1}\left(\frac{\sum_k N_{j\to k}}{N_j} > \eta\right) \tag{3.11}$$

The threshold parameter η is set to 0.5 focusing on the elimination of LTPs via graph partitioning that are replaceable in at least 50% of the training ROIs. Indeed, graph partitioning tends to preserve nodes that are not connected, which in our case would correspond to

LTPs that are not easily replaced by other ones in the labeling task.

We use the Infomap algorithm (Rosvall and Bergstrom 2008) to partition the similarity graph G. We define the frequency of each node on G as the sum of the similarity weights of connected nodes divided by twice the total weight in G. Then, each node is encoded with Huffman coding, where short codewords are assigned to the high-frequency nodes and long codewords are assigned to the low-frequency ones. Infomap then finds an efficient description of how information flows on the network. By detecting the partition that minimizes the description length of the network, Infomap returns a final set of sLTPs with guaranteed global optimality.

Texture and spatial centroids $\{\overline{FT}_{sLTP_k}, \overline{FS}_{sLTP_k}\}$ of the sLTPs $\{sLTP_k\}$ are then computed with Equation (3.5) utilizing the ROIs labeled with $\{LTP_k\}$.

3.2.6 Labeling of CT Scans with sLTPs

In the test stage, scans in the whole dataset are labeled by extracting sample points and their ROIs $\{x\}$. Since it is computationally prohibitive to evaluate the textural and spatial features on every voxels within the lung masks, we only label centers of ROIs that are densely sampled using again SURS. Sampled ROIs with percent emphysema measurements below the previously defined thresholds will have their center labeled as no-emphysema class. Remaining sampled centers get a sLTP label, via minimization of the following cost metric:

$$\chi^2(FT_x, \overline{FT}_{sLTP_k}) + \lambda \cdot W \cdot ||FS_x - \overline{FS}_{sLTP_k}||_2^2 \tag{3.12}$$

Non-sampled voxels are labeled with the sLTP index of the nearest sampled center point via nearest neighbor search within the lung mask (i.e. using a Voronoi diagram). Labeling lung scans with the discovered sLTPs generates histograms of sLTPs, which are efficient lung texture signatures exploited for several tasks, as described in the evaluation sections.

3.2.7 Spatial Density Visualization of sLTPs

To study the spatial distribution of sLTPs, we generate spatial visualization by scatter plotting of voxels labeled with individual sLTPs in sagittal projections, as follows.

We first randomly sample a initial set of ROIs over each lung via SURS sampling. Each ROI is associated with its center point coordinates (r, θ, ϕ) in the PDCMs. To avoid artificial higher densities on the scatter plot in regions close to the core, we adapt the number of ROIs selected per radial regions. The r values are binned into N_r intervals with midpoint values $r_1, ..., r_{N_r}$ to generate isovolumetric sub-volumes of the lung. We then define the sub-sampling ratio $\alpha_i = r_i / r_{N_r}$ (which approximates the ratio of areas in the scatter plot) and set the number of ROIs sampled per r bin to $N_{\text{IsoV}_i} = \alpha_i \cdot N_{\overline{\text{IsoV}}}$ where $N_{\overline{\text{IsoV}}}$ is a pre-set number of ROIs sampled in the outermost part of the lung.

All ROI centers in the sub-sampled set are converted to (x, y, z) Cartesian image coordinates and accumulated in a sagittal single plane, by setting $x = 0$. Final density plots of sLTPs are shown in projected radial coordinates $r' = \sqrt{y^2 + z^2}$ and $\phi' = atan(z/y)$. We color-code each point on the sagittal projection with the following density measure:

$$Den_{sLTP_k}^{(r',\phi')} = \frac{|\Lambda_{sLTP_k} \cap \Lambda_{(r',\phi')}|}{|\Lambda_{sLTP_k}|} \Big/ \frac{\sum_i |\Lambda_{sLTP_i} \cap \Lambda_{(r',\phi')}|}{\sum_i |\Lambda_{sLTP_i}|} \tag{3.13}$$

where $\Lambda_{(r',\phi')}$ denotes the set of ROIs at (r', ϕ') positions. The numerator (the first term) in Equation (3.13) measures the probability of $sLTP_k$ at projected position (r', ϕ'), and the denominator (the second term) measures the observed overall probability of (r', ϕ') to host any $sLTP_i$.

3.3 Experimental Results in MESA COPD Study

3.3.1 Data

The data used for evaluation consists of full-lung CT scans of 317 subjects. All subjects had underwent CT scanning in the MESA COPD study (Smith et al. 2014), between 2009−2011. In addition, 22 out of the 317 subjects underwent CT scanning in the EMCAP study (Barr et al. 2007), between 2008−2009.

For the MESA COPD study, all CT scans were acquired at full inspiration with either a Siemens 64-slice scanner or a GE 64-slice scanner, at 120 kVp, speed 0.5 s, and current (mA) set according to body mass index following the SPIROMICS protocol (Couper et al. 2013). Images were reconstructed using B35/Standard kernels with axial pixel resolutions within the range [0.58, 0.88] mm, and 0.625 mm slice thickness.

For the EMCAP study, scans were acquired with a Siemens 16-slice scanner, at 120 kVp, speed 0.5 s, and a current between 169 mA and 253 mA. Images were reconstructed using the B31f kernel with axial resolutions within the range [0.49, 0.87] mm, and 0.75 mm slice thickness.

Emphysema subtypes and severity have previously been assessed visually in the MESA COPD study (details available in Smith et al. 2014). The raters included four experienced

chest radiologists from two academic medical centers. They assessed emphysema subtypes on CT scans by assigning a percentage of the lung volume affected by CLE, PLE and PSE respectively. Based on (Smith et al. 2014), $N = 205$ subjects do not exhibit emphysema, and are used here as the control set of no emphysema (NE) subjects. The remaining $N = 112$ subjects exhibit light ($N = 53$) or mild-to-severe ($N = 59$) emphysema. For these subjects, predominant emphysema subtype is defined as the subtype affecting the greatest proportion of the lungs. In the mild-to-severe cases, there are $N = 37$ CLE-predominant, $N = 12$ PLE-predominant, and $N = 10$ PSE-predominant subjects. Overall population prevalence of emphysema in the MESA COPD cohort is 27%, composed of 14% of CLE-subtype, 9% of PSE-subtype, and 4% PLE-subtype.

3.3.2 Population Evaluation of Emphysema Using PDCM

We first demonstrate the ability of our proposed PDCM lung shape mapping to study the spatial patterns of emphysema over a population of subjects in Fig. 3.2. For each scan in MESA COPD study, PDCMs of voxels inside individual lungs are generated, attributing to each voxel a coordinate (r, θ, ϕ). Voxel intensity values in PDCMs are then averaged and visualized along two types of projections:

1. *Angular projections*: intensity values averaged along r for each pair of angular directions (θ, ϕ);
2. *Radial projections*: intensity values averaged over all angular directions at a subset of $N_r = 60$ regular radial positions $r_1, ..., r_{N_r}$.

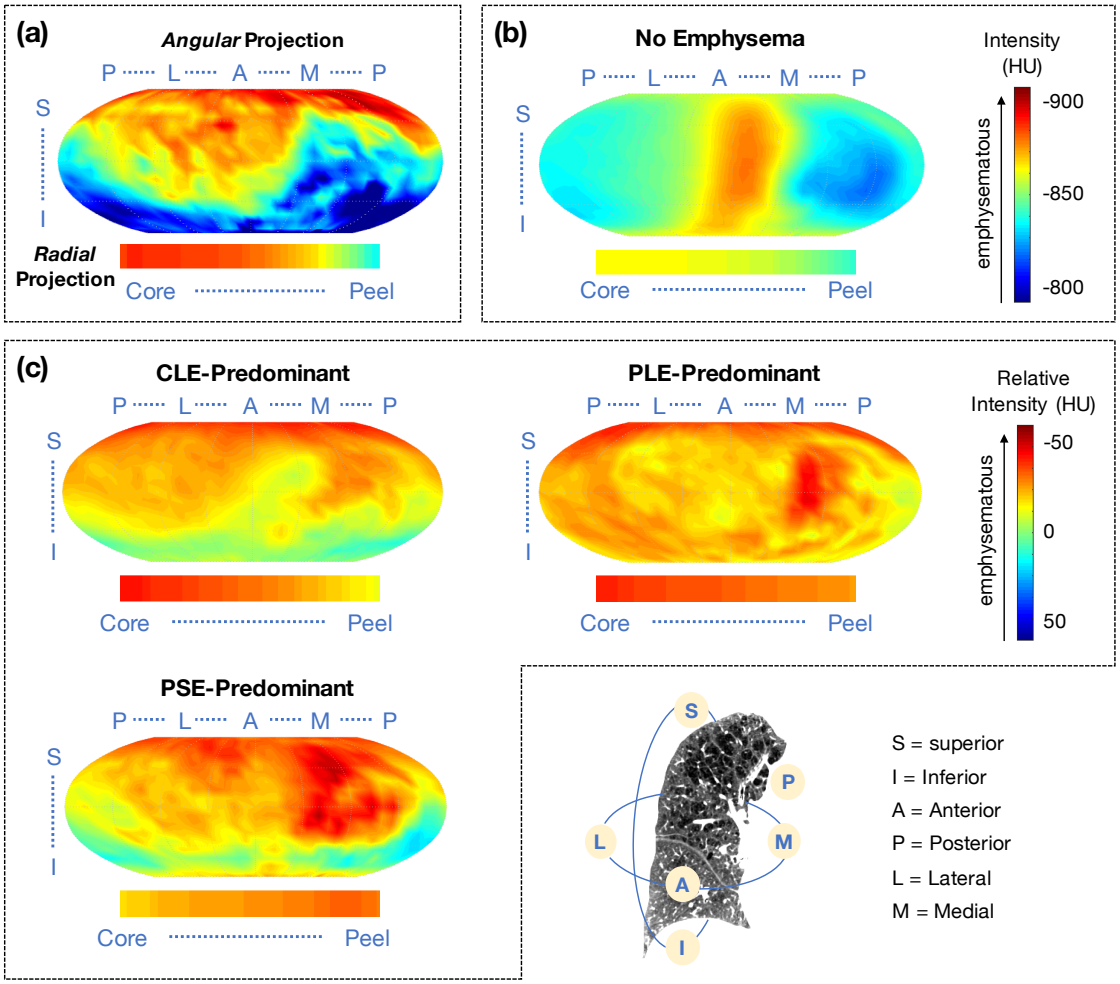

Figure 3.2: Population evaluation of emphysema spatial distribution in MESA COPD, using the proposed PDCM spatial mapping. (a) Illustration of PDCM-based intensity projections on a sample right lung. (b) Average intensity (in HU) on PDCM-based angular and radial projections for MESA-COPD subjects with no emphysema (N=205); (c) Average relative intensity differences, with respect to (b), on PDCM-based projections for MESA-COPD subjects with CLE-, PLE- and PSE-predominant emphysema (N= 37, 12 and 10 respectively).

An illustration of these two PDCM intensity projections on a sample lung are visualized in Fig. 3.2 (a).

Population-average PDCM angular and radial intensity projections over subjects without emphysema (NE) are displayed in Fig. 3.2 (b). The averaged angular projection shows a clear pattern of lower attenuations (i.e. intensity values) in the anterior versus poste-

rior region, which agrees with the intensity gradient due to gravity-dependent regional distribution of blood flow and air (Chabat et al. 2000; West 1963). The averaged radial projection shows a slight gradient from core to peel regions, which is likely due to the inclusion of voxels belonging to the mediastinal and costal pleura inside the lung mask.

Population-average PDCM intensity projections over subjects with CLE-, PLE-, and PSE-predominant emphysema subtypes are visualized in Fig. 3.2 (c). To highlight differences with respect to the control set, we display relative values after subtraction of the values from the corresponding NE average projection in Fig. 3.2 (b). Color coding represents relative intensity differences with more emphysema (more negative attenuation values) corresponding to the red color.

We can see on the relative *angular* PDCM intensity projections that regions of normal attenuation (green to blue) are absent for PLE-predominant subjects, whereas CLE- and PSE-predominant subjects appear to have emphysema regions (red) concentrated in the superior lung. The average relative *radial* PDCM intensity projections on emphysema subjects show systematic higher attenuation values, with more emphysema in the core regions for CLE-predominant subjects and more emphysema in the peel regions for PSE-predominant subjects.

3.3.3 Qualitative Evaluation of Discovered sLTPs

For the discovery of sLTPs, 3/4 of the total scans in MESA COPD study (N=238) were used for training, using random stratified sampling without replacement, while the other scans (N=79) were used for testing. We summarize the setting of pre-defined parameters for the sLTP learning in Table 3.1. In addition, spatial regularization weight λ is set via

Table 3.1: Parameter setting for sLTP learning.

Parameters	Setting
ROI size	= 25 mm^3, to approximate the size of secondary pulmonary lobules
β_1: random shift (for ROI sampling)	$\in [0, 25]$ mm
β_2: sample density (for ROI sampling)	= 3 samples per stack
# of textons: (for texture feature)	= 40, targeting 10 textons per standard emphysema subtype and normal tissue class, according to (Yang et al. 2017)
Texton size	3×3×3 pixels, according to (Yang et al. 2016b)
# of lung sub-regions (for spatial feature)	= 36, according to binning of (r, θ, ϕ) in Section 3.2.3.
N_{LTP}: # of LTPs in initial set	= 100, as suggested in (Yang et al. 2016b), for sufficient diversity of the patterns and being able to discover rare emphysema types

empirical tuning using Equation (3.9). Based on the relative texture homogeneity loss measure ΔSSW_T, we chose $L_T = 1\%$ which corresponds to $\lambda = 1.52$, above which ΔSSW_T increases drastically.

A total of 12 sLTPs were discovered using the full training set, and were used to label both the training and test scans in emphysema-like lung. Each sLTP was detected (i.e. $\%sLTP_k > 0$) in at least 5% of scans both in training and test sets. In Fig. 3.3, we illustrate in (a) the sLTP labeling of two sample CT scans; and in (b) the characteristics of each sLTP via visual illustrations of labeled patches, average occurrence in MESA COPD scans, and spatial distribution of their occurrence within the lungs. For the patch illustrations, 9 samples were randomly selected from all available labeled ROIs. For the average occurrence, we averaged $\%sLTP_k$ values over scans with $\%sLTP_k > 0$. For the spa-

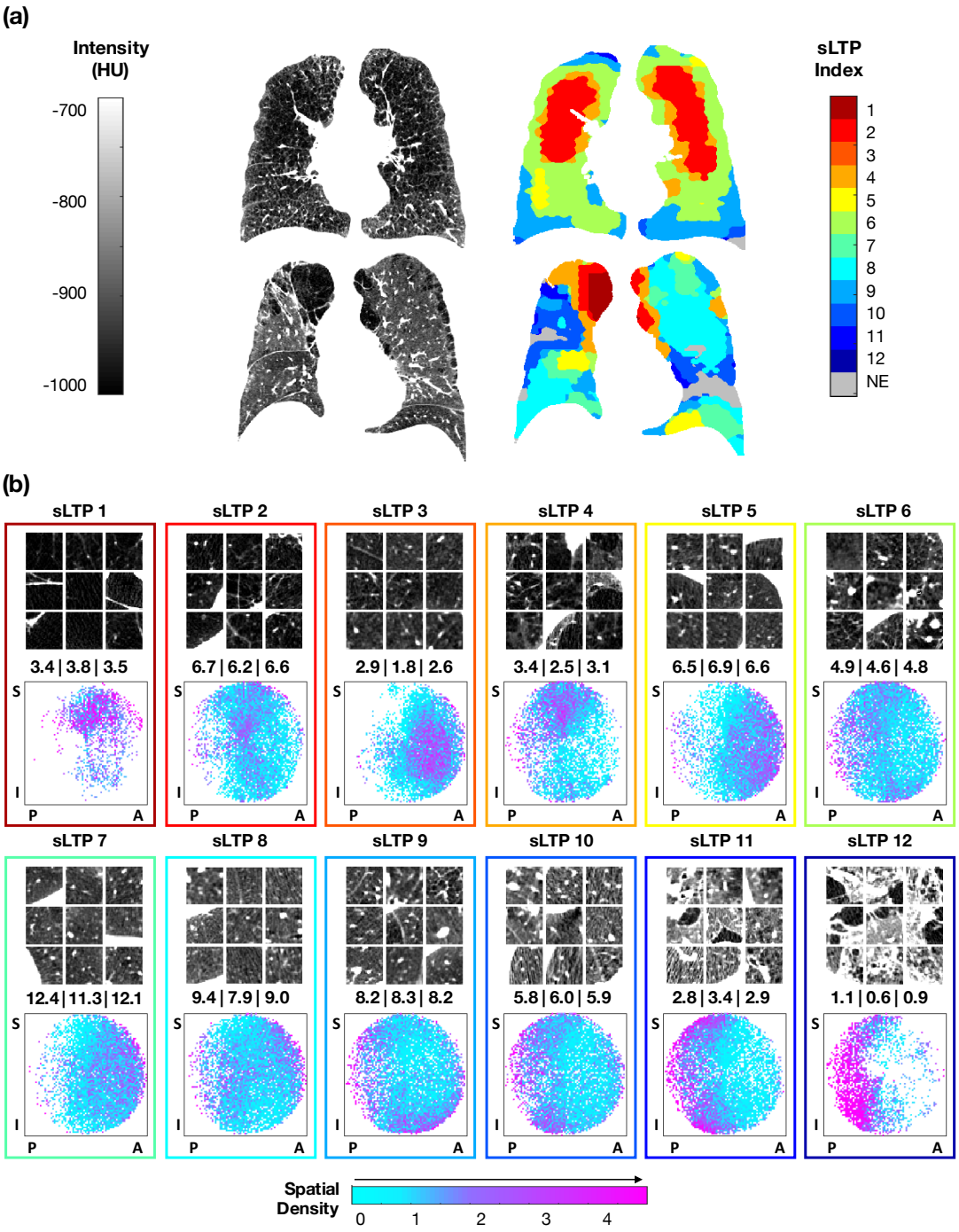

Figure 3.3: Qualitative illustrations of discovered sLTPs in MESA COPD. (a) Two examples of lung scans and their sLTP labeled masks; (b) Characteristics of $\{sLTP_k\}_{k=1,..,12}$, from top to bottom: texture appearance (visualized on axial cuts from 9 random ROIs); average $\%sLTP_k$ on MESA COPD scans presented within training | test | all cases; Spatial density plots of $sLTP_k$ using labeled ROIs (legend: S = superior; I = inferior; P = posterior; A = anterior positions).

tial distributions, we generated spatial scatter plots of sLTP locations from labeled ROIs, following the method described in 3.2.7, with $N_{\overline{\text{Iso}V}} = 5,000$, and $N_r = 60$.

We can observe that patches belonging to an individual sLTP appear to be textually homogeneous. sLTP 1 and 4 show clear spatial accumulation in superior (apical) regions, sLTP 3, 5 and 7 in anterior regions, and sLTP 10, 11 and 12 in posterior regions. All sLTPs returned similar occurrences in training and test sets. Some sLTPs are rare, such as sLTP 12 which covers $\sim$1% of the lungs when present, but is still found in 24 scans over the whole MESA COPD cohort.

3.3.4 Reproducibility of sLTPs

Reproducibility of sLTP Labeling versus Training Sets

To test the reproducibility of sLTPs learning, we first compare the set of $N_{\text{sLTP}} = 12$ sLTPs $\{sLTP_k\}$ generated with the full set of training scans, to $N_{\text{set}} = 4$ sLTPs sets $\{sLTP_k^c\}_{(c=1,2,3,4)}$ using subsets of training data by randomly eliminating 25% of the training scans. Reproducibility of sLTPs is evaluated on the ROI labeling task, by computing the average overlap of labeled test ROIs with the following metric:

$$R_{\text{ln}} = \frac{1}{N_{\text{set}} \cdot N_{\text{sLTP}}} \sum_{c=1}^{N_{\text{set}}} \sum_{k=1}^{N_{\text{sLTP}}} \frac{|\Lambda_{sLTP_k} \cap \Lambda_{\pi(sLTP_k^c)}|}{|\Lambda_{sLTP_k}|} \tag{3.14}$$

where Λ_{sLTP_k} denotes the set of ROIs labeled with $sLTP_k$, and $\pi()$ denotes the permutation operator on the $\{sLTP_k^c\}$ determined by the Hungarian method (Roth et al. 2002) for optimal matching between sets $\{sLTP_k\}$ and $\{sLTP_k^c\}$.

Compared with the $N_{\text{sLTP}} = 12$ sLTPs learned on the full training set, we discovered

$N^c_{\text{sLTP}} = 12$, 12, 13, and 13 sLTPs on training subsets. We obtain an overall labeling reproducibility measure of $R_{\text{ln}} = 0.91$ which corresponds to a high reproducibility level. We then further compute the reproducibility measure, denoted as R'_{ln}, among training subsets. The metric is similar to Equation 3.14, replacing $\{sLTP_k\}$ and $\{sLTP^c_k\}$ with sLTPs $\{sLTP^{c1}_k\}$ and $\{sLTP^{c2}_k\}$ ($c1 \neq c2$) learned on different training subsets. We obtain an overall labeling reproducibility measure of $R'_{\text{ln}} = 0.85$ (standard deviation = 0.07).

To evaluate the contribution of spatial features in sLTP learning, we further generate sets of lung texture patterns using only texture features (i.e. using initial LTPs without spatial augmentation in Section 3.2.4, and setting $\lambda = 0$ for replacement test in Section 3.2.5). We discovered 11 patterns using the full training set, and 11, 11, 12 and 12 patterns on training subsets. The reproducibility measures R_{ln} and R'_{ln} are 0.84 and 0.78 (standard deviation = 0.12) respectively, both are lower than the proposed sLTP learning, hence confirm the benefit of adding spatial features.

Reproducibility of sLTP Labeling versus ROI Sampling

As detailed in Section 3.2.6, sLTP labeling is based on a subset of voxels setting ROI positions, using SURS-based sampling strategy, which is controlled with the parameter β_2 (number of samples per stack). The selected ROIs have an influence on the final outline of the label map, which is expected to be minor if ROIs are sampled densely enough and if sLTPs are generic enough.

In this experiment, we test this hypothesis by generating two different sets of ROIs on test scans using two different random seedings, and measure the reproducibility of the generated label masks using the $\{sLTP_k\}$ discovered on the full training set, while

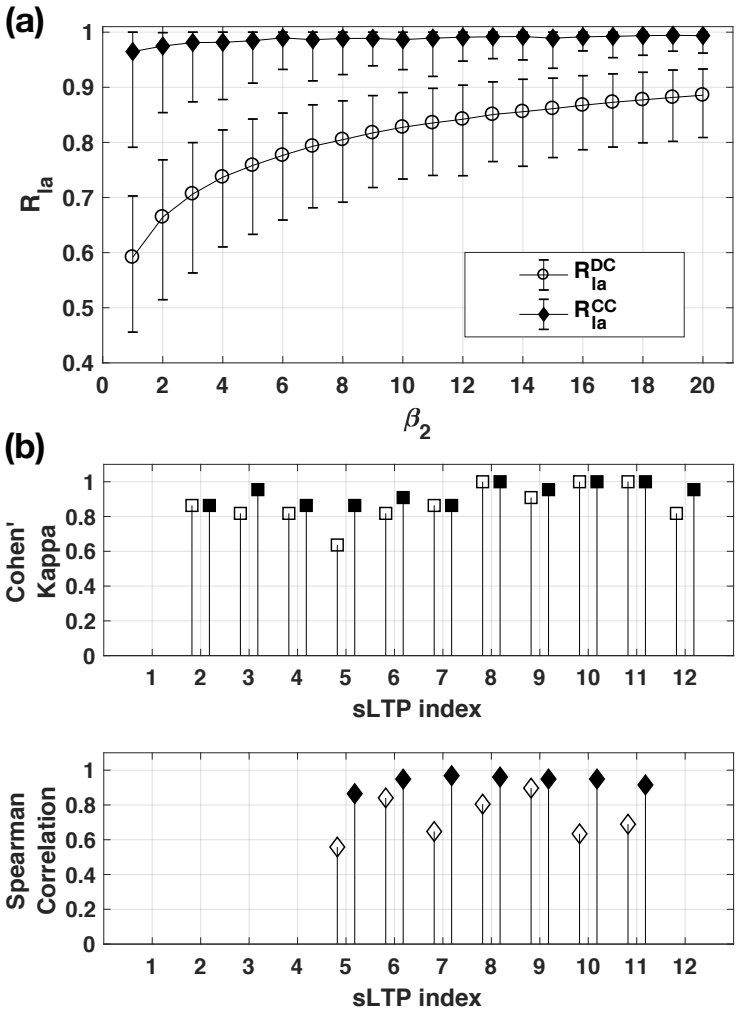

Figure 3.4: Results of sLTP reproducibility measures in MESA COPD. (a) Reproducibility measures R_{la} versus ROI sampling parameter β_2; (b) Reproducibility of sLTPs labeling across scanners (from EMCAP and MESA COPD studies) measured with Cohen's Kappa coefficients of $sLTP_k$ presence and Spearman correlation coefficients of $\%sLTP_k$ values (white = without and black = with intensity histogram mapping).

varying the β_2 parameter. We measure labeling reproducibility using the two sets of ROIs with the following metrics:

- $R_{\text{la}}^{DC}(sLTP_k, \beta_2)$ = average of Dice coefficients of sLTP masks over all test scans;
- $R_{\text{la}}^{CC}(sLTP_k, \beta_2)$ = Spearman correlation coefficients of $\%sLTP_k$ values within the lungs over all test scans.

We illustrate in Fig. 3.4 (a), the average, max and min values of R^*_{la} measures over all $\{sLTP_k\}$, for $\beta_2 \in [1,\ 20]$. Both reproducibility measures increase with β_2 in an exponential manner. We obtain an average $R^{DC}_{\text{la}} > 0.8$ when $\beta_2 > 10$, corresponding to sampling less than 0.05% points in each stack. We obtain an average $R^{CC}_{\text{la}} > 0.9$ when $\beta_2 > 5$. Minimum R_{la} values always occur for sLTP 12, which is the rarest sLTP, as reported in Section 3.3.3.

Reproducibility of sLTP Labeling versus Scanner Type

The 22 subjects from MESA COPD previously scanned within the EMCAP study, underwent different generations of CT scanners. This subset of population is relatively normal. The average time lapse between EMCAP and MESA COPD scans is 14-months. The mean of $\%emph_{-950}$, calibrated for outside air values, is 0.7% (min $<$ 0.1%, max = 3.9%) in EMCAP, and 2.6% (min = 0.3%, max = 9.5%) in MESA COPD, corresponding to an average increase of $\%emph_{-950}$ equal to 1.9%. Therefore, we use this subset of scans to evaluate the reproducibility of sLTP labeling versus scanner types.

We used the 12 sLTPs discovered on the full MESA COPD training set. Because of differences in scanner generations (axial CT in EMCAP versus spiral CT in MESA COPD) and radiation dose settings, intensity calibration was required, implemented in two steps: 1) equalizing the outside air mean intensity value (according to Häme et al. 2014); 2) histogram mapping of normal lung parenchyma identified with the HMMF-based emphysema masks. The sLTPs 2 to 12 were found to be present in both datasets, but sLTPs $\{2, 3, 4, 12\}$ occur in less than 6 pairs of scans. We report in Fig. 3.4 (b) the Cohen's Kappa coefficients of $sLTP_k$ presence for sLTPs 2-12, and the Spearman correlation coefficients

of $\%sLTP_k$ for the frequent sLTPs only (sLTPs 5 to 11). The Cohen's Kappa coefficients and Spearman correlations are all above 0.8, which confirms robust sLTP presence and percentage labeling on the 22 subjects scanned on different scanner types in two studies.

3.3.5 sLTPs' Ability to Encode Standard Emphysema Subtypes

When generating unsupervised lung texture patterns (either sLTPs in this work or earlier generations of LTPs in previous work), we expect them to be finer-grained than the three standard emphysema subtypes used in (Smith et al. 2014), while still capable to encode them, hence linking unsupervised image-based emphysema subtyping with clinical prior knowledge.

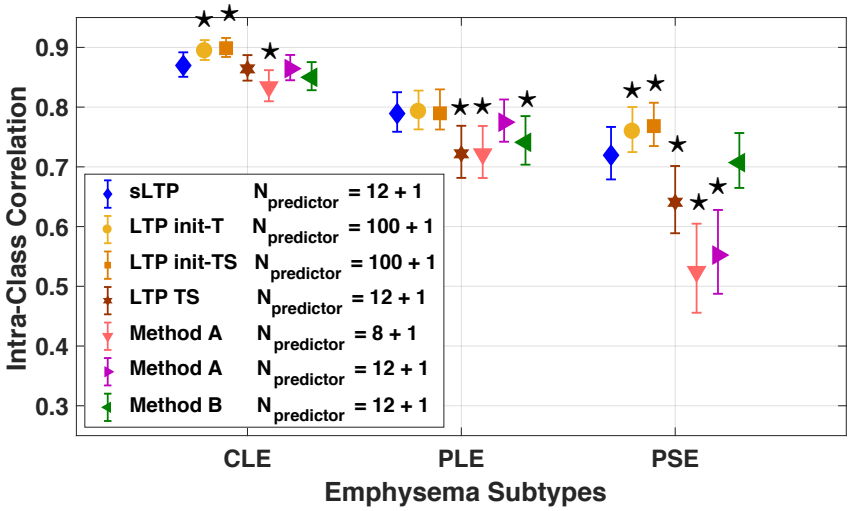

Figure 3.5: Intraclass correlation (ICC) and 95% confidence interval between standard emphysema subtype scores predicted from %sLTP versus ground-truth. Differences with sLTP-based values are marked as $\star$ when significant ($p < 0.05$).

The LTPs (or sLTPs) can be interpreted as either pure or a mixture of the three standard subtypes. We hereby evaluate the ability of the generated LTPs (sLTPs) to predict the overall extent of standard emphysema subtypes. To do this, we generate, for each scan and per lung, two signature vectors: 1) a LTP signature histogram composed of the

percentage of non-emphysema class (obtained as in Section 3.2.6) and the percentages of individual LTP (sLTP) in the emphysema-like lung. This normalized histogram is called the LTP predictor signature and is of size $N_{\text{predictor}} = N_{LTP} + 1$; 2) a ground-truth signature composed of the percentage of non-emphysema and the three standard emphysema subtypes, as visually evaluated in (Smith et al. 2014). A constrained multivariate regression model is used on labeled training scans to learn regression coefficients between the LTP and ground-truth signatures, using the following optimization:

$$\text{argmin}_A \|XA - Y\|_2^2 \text{ s.t. } 0 < A_{k,i} < 1 \text{ and } \sum\nolimits_i A_{k,i} = 1 \tag{3.15}$$

where $X_{N_{\text{scan}} \times N_{\text{predictor}}}$ is composed of all training LTP signatures in N_{scan} training scans, and $Y_{N_{\text{scan}} \times 4}$ contains the ground-truth signatures. $A_{N_{\text{predictor}} \times 4}$ is the matrix of regression coefficients $\{A_{k,i}\}$, which measure the probability of a voxel labeled as a certain predictor belonging to one of the ground-truth classes, and are therefore constrained to be in the range of $[0, 1]$. Optimization of regression was solved using the CVX toolbox [2].

Quality of prediction is measured with the intraclass correlation (ICC) between predicted and ground-truth exploiting the full MESA COPD dataset. We use a 4-fold cross validation (3/4 label masks used for training the regression and 1/4 used for testing and measuring prediction quality). Significance of differences in ICC values was assessed using Fisher's r-to-z transformation and a two-tailed test of the resulting z-scores.

In Fig. 3.5, we compare prediction quality with 7 sets of *emphysema-specific* LTPs (re)trained on the same set of emphysematous ROIs: 1) the 12 sLTPs learned in this study;

[2] http://cvxr.com/cvx

2-3) the initial set of 100 LTPs generated in this study before (denoted as LTP init-T) and after (denoted as LTP init-TS) spatial augmentation; 4) LTPs generated by one-stage clustering (denoted as LTP TS) of the proposed texture and spatial features, by setting $N_{LTP} = 12$ directly (this is to test the contribution of the proposed two-stage learning in Section 3.2.4); 5-6) LTPs re-generated using Method A (Häme et al. 2015b), discovered via graph partitioning of 100 candidates based on local spatial co-occurrence and with $N_{LTP} = 8$ as in the original work or 12; 7) LTPs re-generated using Method B (Yang et al. 2016b), discovered via merging 100 candidates based on texture similarity and local spatial co-occurrence, and setting $N_{LTP} = 12$ for the iterative merging.

Fig. 3.5 shows that the two sets of 100 LTP models achieve overall best prediction accuracy, and that the newly discovered 12 sLTPs have the best performance among the 5 small LTP sets. Difference of ICC values between the sLTPs and the 100 LTP models was not significant for PLE emphysema subtype.

3.4 Experimental Results in SPIROMICS and MESA Lung Study

3.4.1 Data

The SPIROMICS recruited 3,200 cases of COPD and controls (N = 200 non-smokers), 40-80 years of age with $\geq$ 20 pack-years of smoking, in 2010-2015 at 7 major sites and 5 smaller sites (Couper et al. 2013). Exclusion criteria included other chronic lung diseases except asthma, body mass index (BMI) > 40 kg/m2, prior lung resection, metal in the chest, and pregnancy. The MESA Study is a multicenter, prospective cohort study of whites, African-

Americans, Hispanics, and Chinese-Americans (Bild et al. 2002). MESA recruited 6,814 men and women 45-84 years of age in 2000-02 from the general population in 6 communities. Exclusion criteria at baseline were clinical cardiovascular disease, weight over 136 kg, pregnancy, and impediments to long-term follow-up. The MESA Lung Study enrolled participants sampled from MESA who underwent measurements of endothelial function, consented to genetic analyses, and completed an examination in 2004-06 (Tamimi, Serdarevic, and Hanania 2012).

All participants in SPIROMICS and MESA Lung Study underwent full-lung chest inspiratory CT on 64-slice or 128-slice helical scanners (120 kVp, 0.625-0.75 mm slice thickness, 0.5 sec. rotation time) in 2009-14 and 2010-12, respectively, following the same highly-standardized protocol in both studies (Sieren et al. 2016) and on the same CT scanners at 4 sites that were in both studies. In addition, all MESA participants underwent cardiac CT scans in 2000-02 (Bild et al. 2002), which provided complete imaging of the lower lung lobe segment (Hoffman et al. 2009).

Spirometry was performed following the American Thoracic Society recommendations (Miller et al. 2005) on a dry-rolling-seal spirometer in MESA Lung Study and a pneumotachograph in SPIROMICS. Predicted values were calculated using reference equations (Hankinson, Odencrantz, and Fedan 1999). COPD was defined as post-bronchodilator FEV1-to-FVC ratio less than 0.7 (Vogelmeier et al. 2017). Dyspnea was assessed using the modified Medical Research Council (mMRC) breathlessness scale (Norman, Sloan, and Wyrwich 2003), with scores above 0 corresponding to increasing levels of dyspnea-associated disability. Chronic bronchitis was defined by affirmative responses to questions about cough and phlegm production for $\geqslant$ 3 months each year for $\geqslant$ 2 consecutive

years (Kim et al. 2014b). In SPIROMICS, respiratory health status was assessed using the COPD Assessment Test (CAT) (Jones et al. 2009), and the COPD-specific St. George Respiratory Questionnaire (SGRQ-C), with higher scores indicating greater impairment (Hurst et al. 2010). The CAT test consists of 8 questions and yields a score from 0 (no impact) to 40 (very high impact). The SGRQ-C consists of 40 questions and yields a score from 0 (no impairment) and 100 (worst possible health). The minimum clinically important differences for the CAT and SGRQ-C scores are 2 and 4, respectively (Davey et al. 2015).

For full-lung CT scans in SPIROMICS and MESA Lung Study, we select local emphysema ROIs by first segmenting voxels with attenuation below −950 HU, and then using the upper limit of normal (ULN) (Hoffman et al. 2014) of $\%emph_{-950}$ to distinguish ROIs with ($\%emph_{-950} >$ ULN) from without ($\%emph_{-950} \leqslant$ ULN) emphysema regions, to account for differences in normal population variation in percent emphysema measure. For longitudinal cardiac CT scans in MESA, we further exploit the HMMF-based emphysema measure $\%emph_{\text{HMMF}}$ with adapted ULN values, given the better longitudinal performance of $\%emph_{\text{HMMF}}$ as demonstrated in (Yang et al. 2016a, and more details in Chapter 4).

3.4.2 sLTP Learning in SPIROMICS and Data Reduction

We first apply the proposed unsupervised learning algorithm on the full-lung scans from heavy smokers with COPD and controls in SPIROMICS. A random subset of 1,462 participants is initially used for learning the patterns. By applying the unsupervised learning of texture and spatial features from local ROIs, a set of 10 sLTPs are discovered.

To evaluate the inter-learner reproducibility, we re-learn in another non-overlapping subset of 1,460 participants in SPIROMICS. Again, 10 sLTPs are discovered. The learn-

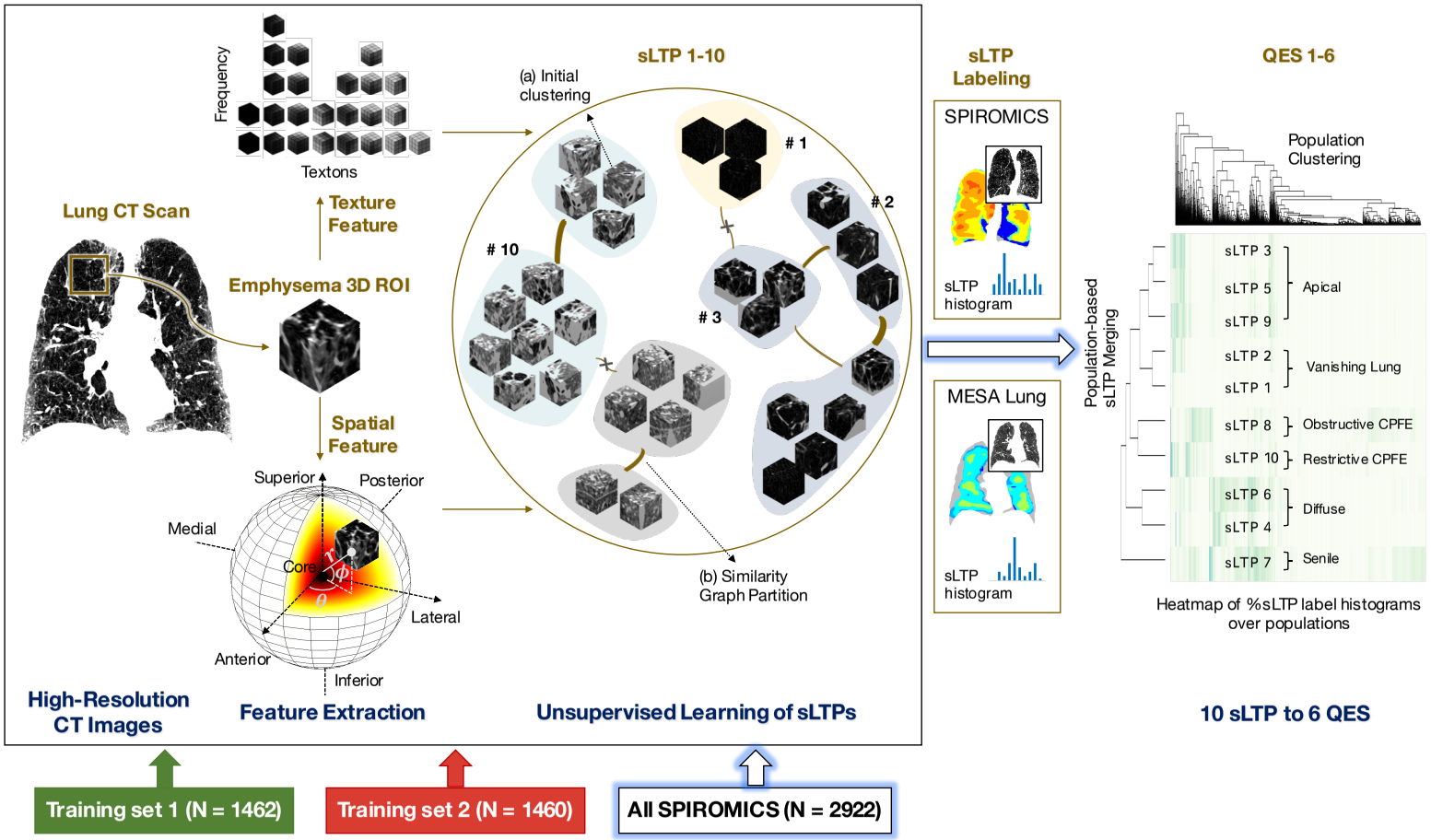

Figure 3.6: Pipeline for learning sLTPs in SPIROMICS, and data reduction to the six quantitative emphysema subtypes (QES) in SPIROMICS and MESA Lung Study. Local ROIs are extracted from emphysema regions in full-lung scans. Texture and spatial features from training ROIs are used for unsupervised learning of the spatial lung texture patterns (sLTPs). The unsupervised learning includes (a) a first stage of ROI clustering based on spatial and texture features and (b) a second stage of similarity graph partitioning of the learned patterns. The unsupervised learning is applied to two non-overlapping subsets of scans in SPIROMICS to evaluate inter-learner reproducibility, and the full set of SPIROMICS CT scans. Then, a final set of 10 sLTPs is generated, which is learned from the full set of SPIROMICS scans, and used to label all scans in SPIROMICS and MESA Lung Study. Data reduction is performed on all sLTP labeling histograms, and leads to six quantitative emphysema subtypes (QES).

ing reproducibility measure at regional level (computed similar to Equation 3.14) is 0.82, which indicates a high level of reproducibility. Then we label all full-lung CT scans in SPIROMICS with the two sets of sLTPs and evaluated the Spearman's correlation coefficients of the percentage of each sLTP within the lungs for each participant. The correlation coefficients are above 0.95 for all sLTPs, which confirms the reproducibility of the learning at the individual level.

Then we apply the unsupervised learning to scans from all 2,922 participants that we have both CT images and ULN values in SPIROMICS, which again yields 10 sLTP (as visualized in Fig. A.1). These 10 sLTPs are used to label all full-lung CT scans in SPIROMICS and MESA Lung Study.

There is evidence that some of the sLTPs overlapped by visual inspection of CT imaging. We collaborated with Dr. Yifei Sun[3] to investigate possible data reduction. Using the sLTP histograms from all scans, we examine the individual-level Spearman's correlations of %sLTP, and heatmaps of sLTP distributions in both SPIROMICS and MESA Lung (see details in Appendix A). We observe that some sLTPs have high population correlations (as shown in Fig. A.2 (a)), suggesting that they may represent the same emphysema subtype at different levels of severity. We therefore aggregate the following sLTPs: (1, 2), (3, 5, 9), (4, 6). This reduces the set of ten sLTPs into six final patterns, that we call the quantitative emphysema subtypes (QES).

Embedding the sLTP histograms in two-dimensions using t-SNE (Maaten and Hinton 2008) on the SPIROMICS and MESA Lung cohorts (as shown in Fig. A.2 (b)), and color-

[3]Dr. Yifei Sun is with the Department of Biostatistics, Mailman School of Public Health, Columbia University.

coding in the projection space the individual subjects by their predominant sLTP / QES, further confirms that the sLTPs can be aggregated into QES in a similar manner.

The whole pipeline to learn the 10 sLTPs in SPIROMICS and then generate the 6 QES with data from SPIROMICS and MESA Lung is illustrated in Fig. 3.6.

3.4.3 Quantitative Emphysema Subtypes (QES)

The six QES are shown in Fig. 3.7 in the order of frequency in SPIROMICS.

The first QES, labelled as *apical*, has a predominantly apical distribution with vascular changes. The second QES, labelled as *diffuse*, has a more diffuse distribution with less parenchymal destruction, apical sparing, and preserved or accentuated vasculature. The third QES, labelled as *senile*, is without visually distinct emphysema but has homogeneously reduced attenuation. The fourth QES, labelled as *restrictive combined pulmonary fibrosis/emphysema (CPFE)*, has distinct and discrete small holes at the level of the secondary pulmonary lobule in apical, posterior and inferior regions resembling centrilobular emphysema with local fibrosis visually suggestive of CPFE. The fifth QES, labelled as *obstructive CPFE*, has diffuse, patchy emphysema with intermingled regions of fibrosis suggestive of a different type of CPFE. The six QES, labelled as *vanishing lung*, has a predominantly apical distribution and visually demonstrates bullous emphysema resembling vanishing lung syndrome (Ladizinski and Sankey 2014) when severe and, when less severe, prominent lobular septal with reduced parenchyma with few vessels. 3

Histograms of QES per CT scan are generated in the SPIROMICS and MESA Lung Study. The apical, diffuse, restrictive CPFE and vanishing lung QES are much more prevalent among heavy smokers and in COPD. The senile and obstructive CPFE QES are equally

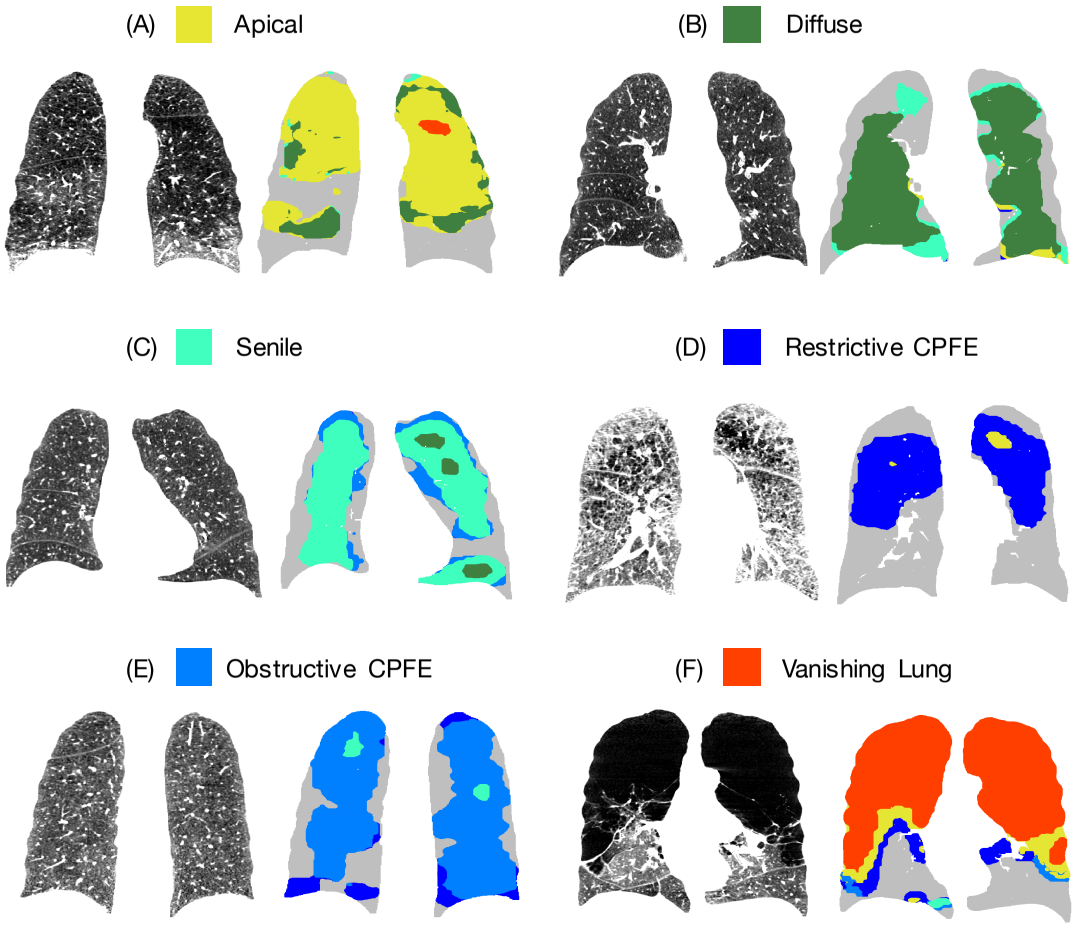

Figure 3.7: Coronal views of original CT scans (gray-scale images) and the corresponding labeled masks with the discovered quantitative emphysema subtypes (QES) on predominantly affected (the proportion of a certain QES larger than any other QES) sample cases in SPIROMICS. Color coding of QES is the same across examples; gray labeling denotes non-emphysematous regions.

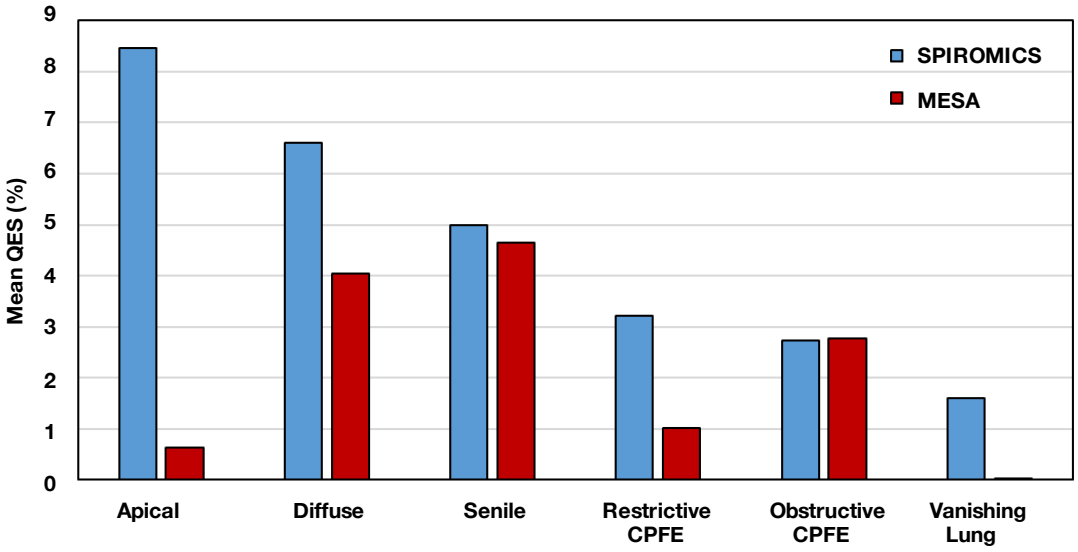

Figure 3.8: Mean values of %QES in heavy smokers in SPIROMICS (excluding the N=200 normal controls) and in the general population in the MESA Lung Study.

common in heavy smokers with COPD in SPIROMICS and in MESA Lung Study, approximately half of whom had never smoked (as shown in Fig. 3.8).

3.4.4 Association between QES and Symptoms

Collaborating with Dr. R. Graham Barr[4], we further investigated the clinical significance of the QES. Dr. Pallavi P. Balte[5] from Dr. Barr's lab helped us to run multivariable regression models to examine associations between the %QES and the respiratory symptoms.

Linear regression is used for MRC-Dyspnea, total SGRQ-C score, resting oxygen saturation, post six-minute walk test (6MWT) oxygen saturation, total 6MWT Distance and exacerbation count; logistic regression is used for the presence of MRC-Chronic bronchitis and wheeze in past 12 months. We adjust for continuous variables age, height and

[4]Dr. R. Graham Barr is with the Department of Medicine, College of Physicians and Surgeons, and Department of Epidemiology, Mailman School of Public Health, Columbia University.

[5]Dr. Pallavi P. Balte is with the Department of Medicine, College of Physicians and Surgeons, Columbia University.

weight as linear terms. Categorical variables include sex, race/ethnicity (white, African-American, Hispanic, and Asian-American), smoking status (never, former, current), pack-years (0, 0 to 10, 10 to 20, > 20), COPD status (yes or no) and CT scanner manufacturer (GE or Siemens). Moreover, we adjust for FEV1, percent emphysema, and other QES, to evaluate the complementarity of the information provided by individual QES[6].

All six QES have independent – but varying – associations with respiratory symptoms and function (as shown in Table 3.2).

In SPIROMICS, the apical QES is associated with greater dyspnea and symptom scores, desaturation on exertion only, reduced exercise capacity (shorter six-minute walk test distance) and greater exacerbation risk independent of demographics, body size, smoking history, lung function and other QES. Alone among QES, it is also associated with symptoms of a chronic productive cough. The diffuse QES, by contrast, is not associated with symptoms independent of lung function and other QES but is characterized by resting hypoxemia, which is not appreciably worsened by exercise, and with greater exacerbation risk. The restrictive CPFE QES is associated with greater dyspnea and symptom scores, desaturation at rest and exertion, and reduced exercise capacity. The senile QES and obstructive CPFE QES are not independently associated with symptoms. The vanishing lung QES is associated with increased dyspnea and desaturation on exertion only.

Findings for available measures in the MESA Lung Study were similar (Table 3.2).

[6]All multivariable regression models in this thesis, to study the clinical associations and prognostic significance of QES, was run by Dr. Pallavi P. Balte.

Table 3.2: Association of QES with respiratory symptoms, physiology and prognosis among smokers in SPIROMICS and the general population in MESA

	Apical	Diffuse	Senile	Restrictive CPFE	Obstructive CPFE	Vanishing Lung	Units
SPIROMICS †							
MRC-Dyspnea	**0.1**	-0.001	0.1	**0.2**	**-0.1**	**0.1**	Scale 0-4
SGRQ Score	**1.1**	0.8	-0.4	**4.2**	-1	0.4	Score 0-100
Resting O2 saturation (%)	0.004	**-0.4**	0.1	**-0.9**	0.2	-0.2	%
Post-exercise O2 saturation (%)	**-0.8**	**-0.4**	-0.3	**-1.8**	**0.4**	**-0.5**	%
6 Minute Walk Test Distance (m)	**-0.3**	-2.3	5.2	**-22**	5.2	-1.8	Meters
Exacerbations *	**0.1**	**0.2**	0.03	**0.3**	0.1	0.1	Count
MESA Lung †							
MRC-Dyspnea	**0.3**	-0.01	-0.0002	**0.2**	-0.02	0.02	Scale 0-4
Resting O2 saturation (%)	-0.5	-0.3	0.2	-0.8	0.3	-2.2	%
FEV1 (mL)	**-310**	18.8	14.6	**-126.5**	**-82.9**	**817.6**	Milliliters
FVC (mL)	**-150.2**	**153.7**	**102.8**	-100.3	**-56**	**1149.6**	Milliliters
FEV1/FVC (%)	**-6.9**	**-2.6**	**-2.6**	-0.9	**-2.0**	8.1	%
TLV (mL)	-64.2	**485.3**	**333.8**	**-378.4**	**145.9**	**1781.7**	Milliliters
MESA ‡							
CLRD Hospitalization	**2.9**	**1.5**	0.8	1	1.4	1.1	
CLRD Mortality	**2.2**	**1.5**	0.9	0.99	**1.8**	1.3	
All-cause Mortality	**1.6**	**0.9**	0.96	1.1	0.9	0.99	

CPFE=Combined Pulmonary Fibrosis/Emphysema; MRC= Medical Research Council; O2=Oxygen;
SGRQ=St. George's respiratory questionnaire; 6MW=Six minute walk;
FEV1= Forced expiratory volume in one second; FVC=Forced expiratory volume in one second;
HR=Hazards ratio; CLRD=Chronic lower respiratory disease
† β estimates compared to normal lung from multivariate linear regression models adjusted for age, sex, race, height, weight, smoking status, pack-years, COPD, scanner manufacturer, FEV1, other QES
* Additionally adjusted for income and education levels
‡ Hazards ratio from Cox proportional hazards models adjusted for age, sex, race, height, weight, smoking status, smoking pack-years, scanner manufacturer, other QES
All estimates are per 10% increase in QES.
Red=statistically significant worsening; **Blue**=statistically significant "improvement"; others=no significant association.

3.4.5 Association between QES and Physiology

We then examine physiologic alterations of QES in the MESA Lung Study. Four QES are associated with obstructive spirometry (reduced FEV1/FVC) consistent with COPD but the reason varied: the apical and obstructive CPFE QES have the expected larger reduction in FEV1 than FVC (Table 3.2) but the diffuse QES and senile QES have reduced FEV1/FVC due to increased FVC, with a consistent increase in total lung volume (TLV) on CT, presumably due to increased lung compliance. In contrast, restrictive CPFE QES are characterized by restrictive spirometry and reduced TLV despite being 'discovered' in SPIROMICS, an obstructive lung disease cohort. The vanishing lung QES is not associated with obstructive spirometry but demonstrates large increases in all lung volumes consistent with loss of elastic recoil from lung destruction.

3.4.6 Prognostic Significance of the QES

We first examine exacerbation risk in SPIROMICS, as previously defined (Woodruff et al. 2016). Apical, diffuse and restrictive CPFE QES are prospectively and independently associated with exacerbations (Table 3.2).

We then examine risk in a larger sample of MESA participants using cardiac CT scans acquired by MESA in 2000-02 (see more details in Chapter 4 and Chapter 5). In order to do this, we develop a deep learning method, based on unsupervised domain adaptation (Ganin et al. 2016; Kamnitsas et al. 2017) to handle the differences between full-lung and cardiac CTs. A convolutional neural network (CNN) classifier is trained with adversarial learning to learn domain-invariant features across imaging scanners and protocols (the

"domains"), while optimizing coherent labeling of QES on pairs of full-lung and cardiac CT scans. More details are provided in Chapter 5.

Among 6,660 participants in MESA Lung followed for a median of 13 years, there are 148 incident hospitalizations for CLRD and 74 deaths from CLRD adjudicated as previously described (Oelsner et al. 2016). All QES except senile and restrictive CPFE predict incident CLRD hospitalizations independent of demographics, body size, smoking history, and type of CT scanner. With additional adjustment for other QES (or percent emphysema), apical and diffuse QES independently predict hospitalizations (Table 3.2).

All QES except the senile QES predicted CLRD mortality independently. With additional adjustment for all other QES (or percent emphysema), the apical, diffuse and obstructive CPFE QES independently predicted CLRD deaths (Table 3.2).

3.4.7 Genome-wide Association Analysis

Collaborating with Dr. Ani Manichaikul[7], we performed a genome-wide association study (GWAS)[8] of the QES in addition to their component sLTPs in SPIROMICS (N = 2,538)[9]. Five significant ($p < 10^{-8}$) novel gene variants are found on different chromosomes for

[7]Dr. Ani Manichaikul is with the Center for Public Health Genomics, University of Virginia.

[8]All GWAS analysis used in this thesis was run by Dr. Ani Manichaikul.

[9]SPIROMICS participants who consented to genetic analysis were genotyped with the Illumina OmniExpress HumanExome BeadChip with SNP level quality control included filter on Hardy-Weinberg $p > 10^{-4}$ and removal of duplicated SNPs. Genome-wide imputation was performed using the Michigan Imputation Server with the Haplotype Reference Consortium (HRC) as the reference panel. Genetic analysis of sLTP and QES traits was performed through pooled analysis of SPIROMICS samples from all race/ethnic groups using a heterogeneous variance model (Sofer et al. 2018) to account for differences in trait distributions across race/ethnic groups, with covariate adjustment for age, sex, four PCs of ancestry, height, weight, CT scanner manufacturer, COPD stratum, current-smoking status and pack-years of smoking. Regression analyses were implemented using SNPTEST v2.5 (Marchini et al. 2007). GWAS results were filtered on 1) heterozygosity count (HC) > 30 and Hardy-Weinberg $p > 10^{-5}$ for genotyped SNPs, or 2) imputation R-squared > 0.5 and effective HC > 30 (where effective HC $=$ HC $\times$ imputation R-squared) for imputed SNPs.

the most severe sLTP of the apical QES, the restrictive CPFE QES, the obstructive CPFE QES (two gene variants), and the more severe sLTP of the vanishing lung QES.

Among the five gene variants, the variant identified for the most severe sLTP of the apical QES is near DRD1 ($p = 3.92 \times 10^{-8}$), which encodes dopamine receptor1 (DAR1). This variant is close to GWAS-significant for the two other sLTPs that constitute the apical QES ($p = 6.48 \times 10^{-6}$ and $p = 2.57 \times 10^{-5}$). DAR1 is present on the pulmonary vasculature and is implicated in vasoconstriction and intrapulmonary shunting (Bryan et al. 2012). Dopamine and other DAR1 agonists (including some anti-Parkinsonian drugs) increase pulmonary artery pressure (PAP) by enhancing hypoxic pulmonary vasoconstriction (Cheung and Barrington 2001, Hong et al. 2005), an effect that is blocked by haloperidol and numerous other DAR1 antagonists (Laurie et al. 2012).

3.5 Discussion and Conclusion

In this chapter, we proposed a novel unsupervised learning framework for discovering lung texture patterns for emphysema on full-lung CT scans, via incorporating spatial and texture features using an original cost metric, along with data-driven parameter tuning, and Infomap graph partitioning. Our methodological framework includes the introduction of a standardized spatial mapping of the lung shape utilizing Poisson distance map and conformal mapping to uniquely encode 3D voxel positions and enable comparison of CT scans without registration being required besides orientation alignment. Our lung shape spatial mapping PDCM enabled straightforward population-wide study of emphysema spatial patterns. By visualizing relative *angular* PDCM intensity projections on CLE-, PLE- and PSE-predominant subjects in MESA COPD, we observed that regions of

normal attenuation were absent for PLE-predominant subjects, which agrees with the definition of PLE (diffused emphysema subtype). CLE- and PSE-predominant subjects appeared to have emphysema regions concentrated in the superior part. This agrees with the observation made in (Smith et al. 2014) on the same dataset that CLE and PSE severity was greater in upper versus lower lung zones, whereas severity of PLE did not vary by lung zone. By visualizing relative *radial* PDCM intensity projections, we observed that emphysema subjects showed systematic higher attenuation values than subjects without emphysema, as expected. CLE-predominant subjects appeared to have more emphysema in the core part, whereas PSE-predominant subjects appeared to have more emphysema in the peel part. This agrees with the definitions of CLE and PSE. As a standardized tool, the proposed PDCM spatial mapping is not tied to emphysema pattern, and we will demonstrate its application to study spatial location of lung nodules, in Chapter 6.

With the proposed method, and using a prefixed percent emphysema threshold 1% to select emphysema-like lung, we discovered 12 spatially-informed lung texture patterns (sLTPs) on the MESA COPD cohort. Qualitative visualization showed that the discovered sLTPs appeared to be textually homogeneous with different spatial prevalence. Since we jointly enforce spatial prevalence and textural homogeneity, each sLTP can have spatial "outliers" that are texturally favored. Extensive evaluations showed that the discovered sLTPs were reproducible with respect to training sets, sampling of ROI for labeling, and certain scanner changes. The proposed incorporation of spatial and texture features obtained higher learning reproducibility compared to using texture features only, confirming the benefit of spatial regularization.

Moreover, the sLTPs discovered in MESA COPD study were able to encode the three

standard emphysema subtypes, and thus link unsupervised discovery with clinical prior knowledge. Prediction quality was better than previous methods, and close to the optimal level reached with 100 *emphysema-specific* LTPs. While intra-cluster LTP homogeneity increases with the number of LTPs, hence leading to higher prediction performance, working with 100 LTPs leads to redundancy between subtypes which is detrimental when studying associations of individual LTPs with clinical measures. One-stage clustering led to significantly lower prediction power for PLE and PSE subtypes, compared to sLTPs, which demonstrated the benefit of the proposed two-stage learning.

Then we applied the unsupervised learning method to the larger cohort of SPIROMICS, using subject-specific threshold values to account for differences in normal population variation in percent emphysema. We discovered 10 sLTPs that were highly reproducible between independent training subsets in SPIROMICS. Population-based heatmaps and hierarchical clustering of sLTP histograms in SPIROMICS and the MESA Lung Study led to data reduction from 10 sLTPs to the final set of six quantitative emphysema subtypes (QES). The six QES were shown to have distinct CT representations and structures, are associated independently with unique patterns of respiratory symptoms and clinical events, have varying physiologic characteristics, and may have non-overlapping genetic associations, hence may facilitate personalized therapies.

Chapter 4

Robust Emphysema Quantification on Cardiac CT Scans Using Hidden Markov Measure Field Model

4.1 Introduction

Pulmonary emphysema is defined by a loss of lung tissue in the absence of fibrosis, and overlaps considerably with chronic obstructive pulmonary disease (COPD). Full-lung quantitative computed tomography (CT) imaging is commonly used to measure a continuous score of the extent of emphysema-like lung tissue, which has been shown to be reproducible (Mets et al. 2012), and correlates well with respiratory symptoms (Kirby et al. 2015). Cardiac CT scans, which are commonly used for the assessment of coronary artery calcium scores to predict cardiac events (Detrano et al. 2008), include about 70% of the lung volume, and can be obtained with low radiation exposure. Despite missing apical and caudal *individual* measurements, emphysema quantification on cardiac CT were shown to have high reproducibility and correlation with full-lung measures (Hoffman et al. 2009), and correlate well with risk factors of lung disease and mortality (Oelsner et al. 2014) at the *population-based* level.

With the availability of large scale well characterized cardiac CT databases such as the Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002), emphysema quantifi-

cation on cardiac scans has now been actively used in various population-based studies (Barr et al. 2010). However, currently used methods for emphysema quantification on cardiac scans rely on measuring the percentage of lung volume (referred to as $\%emph$) with intensity value below a fixed threshold. Although thresholding-based $\%emph$ is commonly used in research, it can be very sensitive to factors that lead to variation in image quality and voxel intensity distributions, including variations between scanner types, reconstruction kernel, radiation dose and slice thickness. Being able to segment emphysema robustly on cardiac CT scans will enable longitudinal study of emphysema progression, and is a prerequisite for applying our proposed lung texture learning in Chapter 3 to large scale cardiac CT datasets.

To study $\%emph$ on heterogeneous datasets of full-lung scans, density correction (Kim et al. 2014a), noise filtering (Schilham et al. 2006) and reconstruction-kernel adaptation (Bartel et al. 2011) have been proposed. These approaches consider only a part of the sources of variation, and their applicability to cardiac scans has not been demonstrated. The superiority of a segmentation method based on Hidden Markov Measure Field (HMMF) model was demonstrated in a previous study in our lab (Häme et al. 2014, 2015a) on full-lung scans. In this work, we propose to further adapt the parameterization of the HMMF segmentation model to cardiac CT scans from 6,814 subjects in the longitudinal MESA Lung Study. Our results compare HMMF-based and thresholding-based $\%emph$ measures for three metrics: 1) intra-cardiac scan reproducibility, 2) longitudinal correlation of $\%emph$ measures on "normal" subjects who are never-smokers without respiratory symptoms or disease (Hoffman et al. 2014), and 3) emphysema progression on "normal" and "disease" subjects.

4.2 Method

In sections below, we first overview the cardiac and full-lung CT data in MESA used in our evaluation, and then present the HMMF-based emphysema segmentation framework.

4.2.1 Data

The MESA Study consists of 6,814 subjects screened with cardiac CT scans at baseline (Exam 1, 2000-2002), and with follow-up scans in Exam 2 to 4 (2002-2008). Most subjects had two repeated cardiac scans per visit (same scanner). Among these subjects, 3,965 were enrolled in the MESA Lung Study and underwent full-lung scans in Exam 5 (2010-2012).

MESA cardiac scans were collected using either one type of EBT scanner from GE, or six types of MDCT scanners (cf. Figure 4.1 (c)) from GE or Siemens (Hoffman et al. 2009). The average slice thickness is 2.82 mm, and isotropic in-plane resolution is in the range [0.44, 0.78] mm.

Lung segmentation was performed with the APOLLO software (VIDA Diagnostics, Iowa). Longitudinal correlation of segmented lung volume in incremental cardiac exams is in the range [0.84, 0.95]. Cardiac CT scans were acquired at full inspiration with cardiac and respiration gating, while full-lung CT scans were acquired at full inspiration without cardiac gating.

For this study we selected a random subset of 10,000 pairs of repeated cardiac scans with one in each pair considered as the "better" scan in terms of inflation or scan quality (Barr et al. 2010). Out of these 10,000 pairs, 379 pairs were discarded due to corruption in one scan during image reconstruction or storage, detected via abnormally high values of

Table 4.1: Year and number of MESA cardiac and full-lung CT scans evaluated for HMMF-based emphysema segmentation.

Exam #	1	2	3	4	5
Year start-end	2000-02	2002-04	2004-05	2005-08	2010-12
# of subjects in MESA	6,814	2,955	2,929	1,406	3,965
# of normals evaluated	741	261	307	141	827
Scan type	cardiac	cardiac	cardiac	cardiac	full-lung
Total # of scans evaluated	6,088 ($\times$2)	1,164 ($\times$2)	1,645 ($\times$2)	724 ($\times$2)	2,984

mean and standard deviation of outside air voxel intensities (cf. Figure 4.1 (c) for ranges of normal values).

The selected subset involves 6,552 subjects, among which 2,984 subjects had a full-lung scan in Exam 5, and 827 are "normals", as detailed in Table 1. We processed a grand total of 9,621 pairs of repeated cardiac scans, 3,508 pairs of "better" longitudinal cardiac scans, and 5,134 pairs of "better" cardiac-full-lung scans.

4.2.2 HMMF-based Emphysema Segmentation

The HMMF-based method enforces spatial coherence of the segmentation, and relies on parametric models of intensity distributions within emphysematous and normal lung tissue that use:

1. A Gaussian distribution $N_E(\theta_E)$ for the emphysema class;
2. A skew-normal distribution $N_N(\theta_N)$ for normal lung tissue.

Using skew-normal distribution to fit the intensity of normal lung tissue on full-lung CT scans was originally proposed in (Häme et al. 2014). We found this model to be ap-

plicable to cardiac scans. Fig. 4.1 (a) gives examples of histogram fitting results for three cardiac scans from normal subjects.

For a given image $I : \Omega \rightarrow R$, the HMMF estimates on Ω the continuous-valued measure field $q \in [0, 1]$ by maximizing the posterior distribution P for q and the associated parameter vector $\theta = [\theta_E, \theta_N]$ expressed as:

$$P(q, \theta|I) = \frac{1}{R} P(I|q, \theta) P_q(q) P_\theta(\theta) \tag{4.1}$$

where R is a normalization constant. The Markov random field (MRF) variable q is a vector $q = [q_E, q_N]$, representing the intermediate labeling of both classes. Emphysema voxels are selected as $\{v \in \Omega | q_E(v) > q_N(v)\}$, from which $\%emph_{\text{HMMF}}$ is computed.

The distribution $P_q(q)$ enforces spatial regularity via Markovian regularization on neighborhood cliques C and involves a weight parameter λ in the potential of the Gibbs distribution. The likelihood $\text{P}(I|q, \theta)$ requires initialization of parameter values for both classes, which are tuned in this work to handle the heterogeneity of the dataset, as described below.

Parameter Tuning for Cardiac CT Scans

Likelihood Parameters

The parameters of intensity distributions are $\theta_E = [\mu_E, \sigma_E]$, $\theta_N = [\mu_N, \sigma_N, \alpha_N]$ where μ denotes the mean, σ the standard deviation and α the skewness of respective classes.

Normal tissue class: The standard deviation σ_N and the skew α_N are assumed to be sensitive to scanner-specific image differences. They are tuned separately for each scanner type by averaging on the subpopulation of normal subjects, after fitting their intensity

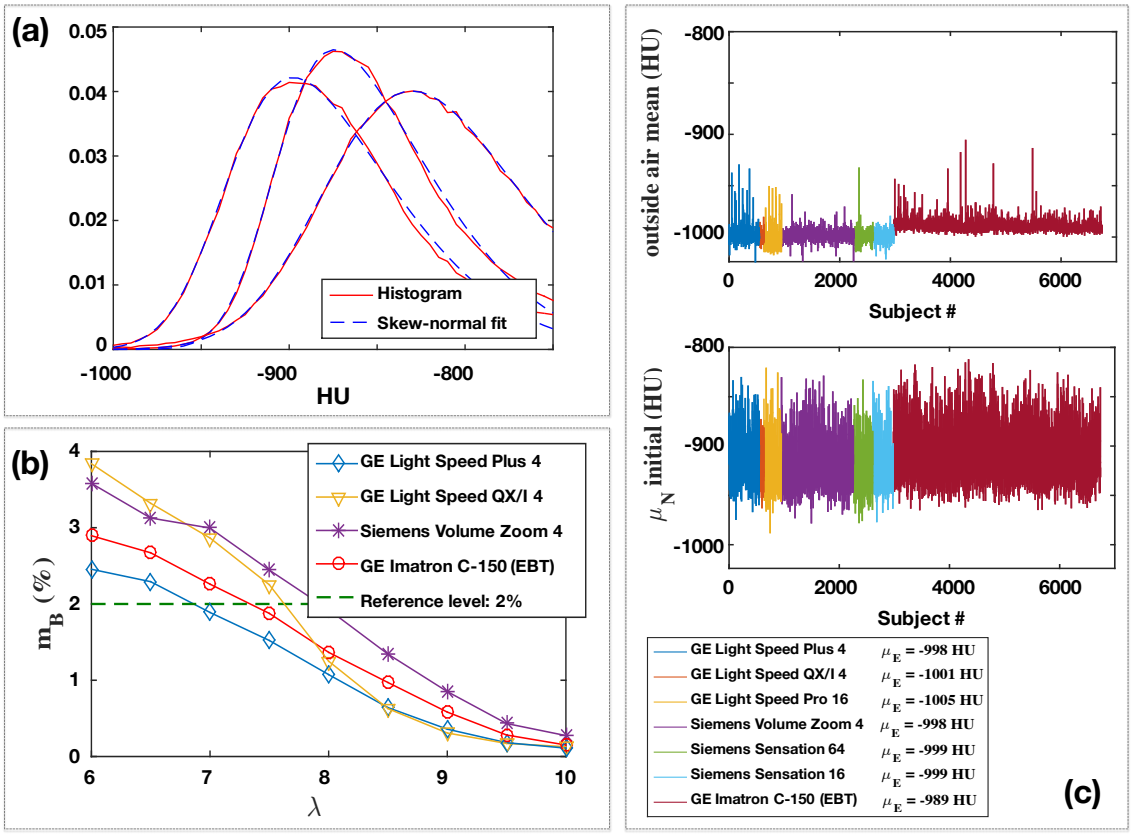

Figure 4.1: Illustration of the proposed HMMF-based framework for emphysema segmentation on cardiac CT scans. (a) Illustration of fitting lung-field intensity with skew-normal distribution on three cardiac scans. (b) Population average of $m_B(\lambda)$ for $\%emph_{\text{HMMF}}$ measured from normal subjects on four baseline cardiac scanners (S_B) versus λ values. The optimal λ_B value is chosen such that $m_B(\lambda_B) = 2\%$, for each scanner type. (c) From top to bottom: Outside air mean value (HU) per subject and per scanner used to tune μ_E; Initial μ_N value (HU) per subject and per scanner.

histograms. The initial value of mean μ_N is sensitive to inflation level and morphology and therefore made subject-specific via fitting individual intensity histograms with the pre-fixed σ_N and α_N. Measured initial μ_N values are plotted in Fig. 4.1 (c).

Emphysema class: The initial value of mean μ_E is set to the average scanner-specific outside air mean value, learned on a subpopulation of both normal and disease subjects from each scanner type, and illustrated in Fig. 4.1 (c). The standard deviation σ_E is set to be equal to σ_N since the value of σ is mainly affected by image quality. Both parameters

are therefore scanner-specific.

Spatial Regularization Parameters

Cliques: The spatial clique is set to 8-connected neighborhoods in 2-D planes instead of 26-connected 3-D cliques used in (Häme et al. 2014) to handle the slice thickness change from full-lung (mean 0.65 mm) to cardiac CT (mean 2.82 mm).

Regularization weight λ: The regularization weight λ is made scanner-specific to adapt to image quality and noise level. There are three scanner categories: scanners used only at baseline (S_B), scanners used at baseline and some follow up times (S_{BF}) and scanners used only at follow up (S_F). For scanners in S_B and S_{BF}, we chose, via Bootstrapping, the λ_B values (for each scanner type) that returns a population average m of the $\%emph_{\text{HMMF}}$ measure on the normal subpopulation equal to $m_B(\lambda_B) = 2\%$ (i.e. a small arbitrary value (Hoffman et al. 2014)). The selection process is illustrated in Fig. 4.1 (b). For scanners in S_{BF}, the same λ_B values are used at follow-up times, leading to population $\%emph_{\text{HMMF}}$ averages m_{BF} (λ_B). Finally, the λ_F are chosen such that $m_{BF}(\lambda_B) = m_F(\lambda_F)$.

Parameter Tuning for full-lung CT Scans

Parameters for the segmentation of full-lung scans with HMMF were tuned similarly to the previous work in (Häme et al. 2014), except for λ and the initial values of μ_N and μ_E. In the previous work, scans reconstructed with a smooth kernel were used as a reference to set λ for noisier reconstructions. In this work, having only one reconstruction per scan in MESA Exam 1-5, we propose to use the progression rate of $\%emph$ measured on longitudinal cardiac scans from the subpopulation of normal subjects. We set $m_{FL}(\lambda_{FL}) = m_{pr}$ with m_{pr} the predicted normal population average of $\%emph$ at the time of acquisition of the full-lung scans FL, based on linear interpolation of anterior

progression rates. This lead to λ in the range [3, 3.5] for different scanners, which is quite different from the range of λ values tuned on cardiac scans (Fig. 4.1 (b)).

4.2.3 Quantification via Thresholding

Standard thresholding-based measures $\%emph_{-950}$ were obtained for comparison, using a threshold of reference T_{ref}. Among standard values used by radiologists, T_{ref} =-950HU was found to generate higher intra-class correlation and lower extreme differences on a subpopulation of repeated cardiac scans.

For reproducibility testing on repeated cardiac scans (same scanner), an additional measure $\%emph_{-950G}$ was generated after Gaussian filtering, which was shown to reduce image noise-level effect in previous studies (Häme et al. 2014). The scale parameter of the Gaussian filter is tuned in the same manner as λ for the HMMF (i.e. matching reference average population values of $\%emph_{-950}$ on the subpopulation of normal subjects). This lead to scale in the range [0.075,0.175]; For longitudinal correlations, an additional measure $\%emph_{-950C}$ was computed correcting T_{ref} (HU) with respect to the scanner-dependent bias observed on mean outside air density values (μ_E), as:

$$T_{ref} = -950 + (\mu_E - (-1000)) \tag{4.2}$$

4.3 Experimental Results

4.3.1 Reproducibility within Cardiac Scans

Intraclass Correlation (ICC) on Repeated Cardiac Scans

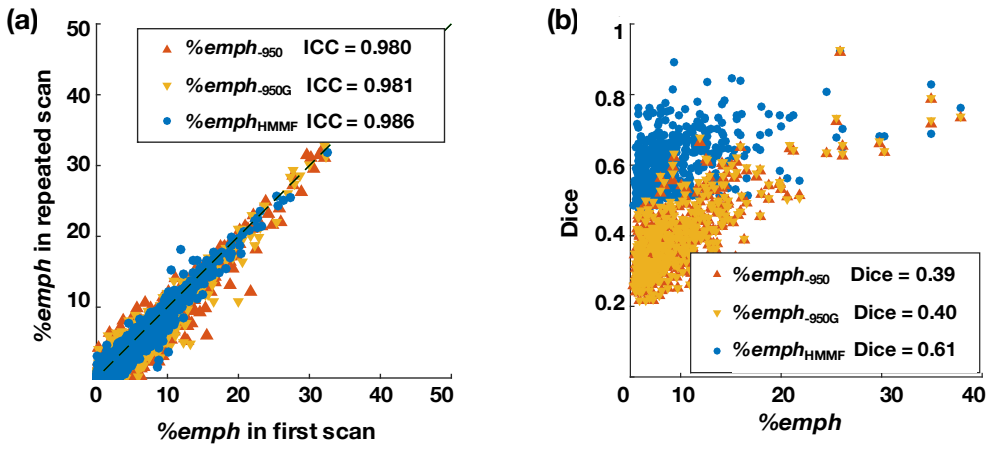

Figure 4.2: Reproducibility of thresholding-based versus HMMF-based $\%emph$ measurements on repeated cardiac scans in MESA Exam 1-4. (a) Intraclass correlation (ICC) (N = 9,621) on repeated cardiac scans; (b) Dice of emphysema mask overlap for disease subjects (N = 471) on repeated cardiac scans.

Scatter plots and ICC (average over Exams 1-4) of $\%emph$ in 9,621 pairs of repeat cardiac scans are reported in Fig.4.2 (a). All three measurements show high reproducibility (ICC $>$ 0.98). $\%emph_{-950G}$ provides minor improvement compared with $\%emph_{-950}$, which may be explained by the low noise level in MESA cardiac scans.

Spatial Overlap of Emphysema Masks on Repeated Cardiac Scans

Lung masks of repeated cardiac scans were registered with FSL (Smith et al. 2004), using a similarity transform (7 degrees of freedom). Spatial overlap of emphysema masks was measured with the Dice coefficient, on subjects with $\%emph_{-950} > 5\%$ (N = 471). Dice coefficient is defined as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{4.3}$$

where TP is the true positive segmentation, FP is the false positive segmentation and FN is the false negative segmentation.

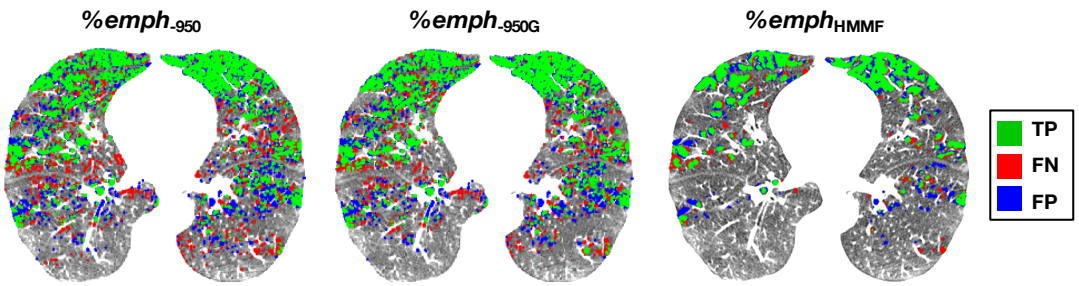

Figure 4.3: Example of emphysema spatial overlap on a baseline axial slice from a pair of repeated cardiac scans, using HMMF and thresholding-based segmentation. (TP = true positive, FN = false negative, FP = false positive).

Scatter plots and average values of Dice coefficient computed are reported in Fig.4.2 (b). Except for very few cases, HMMF returned higher overlap measures than thresholding, with an average Dice = 0.61, which is comparable to the value achieved on full-lung scans (0.62) in the previous work (Häme et al. 2015a). Fig.4.3 gives an example of spatial overlaps of emphysema segmented on a pair of repeated cardiac scans, where there is less disagreement with HMMF.

4.3.2 Longitudinal Correlation and Progression of $\%emph$

Pairwise Correlation on Longitudinal Cardiac Scans

For longitudinal cardiac scans, we correlated all baseline scans and follow-up scans acquired within a time interval of 48 months, in the population of normal subjects, who are expected to have little emphysema progression over time (due to aging). Fig.4.4 (a) shows that $\%emph_{\text{HMMF}}$ measures return the highest pair-wise correlations on longitudinal cardiac scans, followed by $\%emph_{-950C}$ measures.

Emphysema Progression

Differential $\%emph$ scores Δ were computed at follow up times t to evaluate emphy-

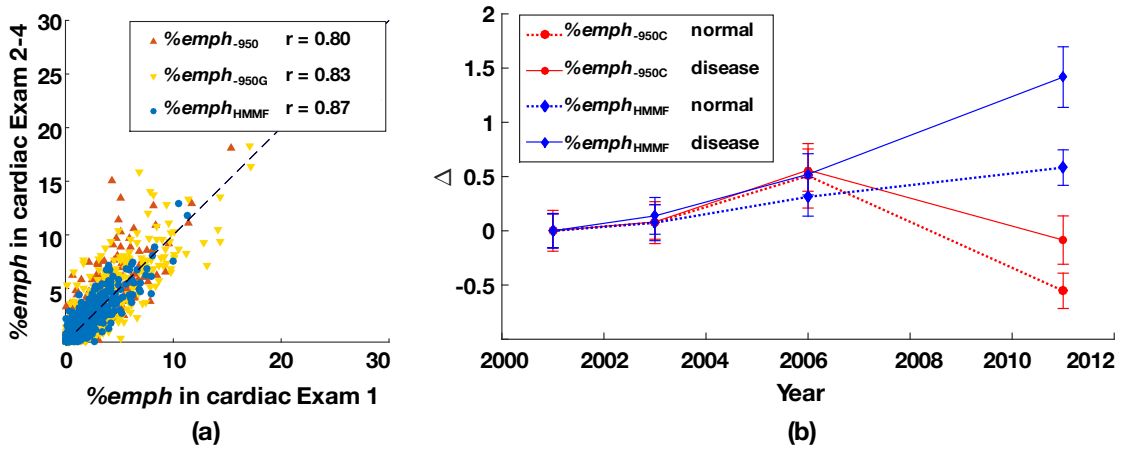

Figure 4.4: HMMF and thresholding-based $\%emph$ measures on longitudinal scans in MESA. (a) $\%emph$ measurements on longitudinal cardiac scans of normal subjects (N = 478); (b) Mean and standard error of the mean of emphysema progression measurement $\Delta(t)$ over time t (normal: N=87, disease: N = 238; r = pairwise Pearson correlation).

sema progression, as:

$$\Delta(t) = \%emph(t) - \%emph(baseline) \tag{4.4}$$

Mean values and standard errors of the mean of Δ for 87 normal subjects and 238 disease subjects who have three longitudinal cardiac scans and one full-lung scan are shown in Fig.4.4 (b). The $\%emph_{\text{HMMF}}$ measures progressed steadily along cardiac and full-lung (measuring on cardiac field of view) scans, and at different rates for normal and disease populations. The $\%emph_{-950C}$ measures progressed steadily across cardiac scans but decreased from cardiac to full-lung scans, which indicates that a single threshold is not able to provide consistency between cardiac and full-lung scans. Furthermore, thresholding based measurements on cardiac scans show similar progression rates in normal and disease populations, which is not what is expected.

Finally, we tested mixed linear regression models (Ahmed et al. 2014) on all longitudi-

nal scans to assess the progression of $\%emph$ over time after adjusting for demographic and scanner factors. The initial model (model 1) includes age at baseline, gender, race, height, weight, BMI, scanner type, voxel size, baseline smoking pack years and current cigarettes smoking per day. In the subsequent model (model 2), interaction terms between time (starting from the baseline) and age at baseline, gender, race, baseline smoking pack years and current cigarettes smoking per day were added to account for the variation in progression of $\%emph$ with time for demographic factors. In model 2 we observed that progression of $\%emph_{\text{HMMF}}$ was higher with higher baseline age (P $= 0.0001$), baseline smoking pack years (P < 0.0001) and current cigarettes smoking per day (P $= 0.03$). These findings were not significant for $\%emph_{-950C}$ except for baseline smoking pack years (P $= 0.0016$). Additionally, both models demonstrated that the effects of scanner types in cardiac scans were attenuated for $\%emph_{\text{HMMF}}$ when compared with $\%emph_{-950C}$.

4.4 Discussion and Conclusion

In this chapter, we introduced a dedicated parameter tuning framework to enable the use of an automated HMMF segmentation method to quantify emphysema in a robust and reproducible manner on a large dataset of cardiac CT scans from multiple scanners. While thresholding compared well with HMMF segmentation for intraclass correlation on repeated cardiac scans, only HMMF was able to provide high spatial overlaps of emphysema segmentations on repeated cardiac scans, consistent longitudinal measures between cardiac and full-lung scans, attenuated scanner effects on population-wide analysis of emphysema progression rates, and clear discrimination of emphysema progression rates between normal and disease subjects.

Exploiting HMMF segmentation to quantify emphysema on low-dose cardiac CT scans has great potentials given the very large incidence of cardiac CT scans. Being able to segment emphysema robustly across heterogeneous scanner types will enable longitudinal study of emphysema progression, and is a prerequisite for applying our proposed lung texture learning in Chapter 3 to the large cardiac CT dataset in MESA.

Chapter 5

Unsupervised Domain Adaption with Adversarial Learning for Emphysema Subtyping on Cardiac CT Scans

5.1 Introduction

Pulmonary emphysema can be characterized by specific texture patterns on CT images. Supervised and unsupervised learning of these texture patterns is an active field of research (Binder et al. 2016; Depeursinge et al. 2014; Yang et al. 2016b, 2017). As described in Chapter 3, in our previous study (Yang et al. 2017) we have established a set of robust emphysema-specific spatially-localized lung texture patterns (sLTPs) on full-lung high-resolution CT (HRCT) scans, using a dedicated parcellation of the lung shape to introduce location information in lung texture learning. So far, largely due the the limited availability of high-quality longitudinal full-lung CT data, lung texture patterns for emphysema have not been previously studied in longitudinal setting, while this is crucial for understanding disease progression.

The Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002) study consists of 6,814 subjects screened with cardiac CT scans at baseline (Exam 1, 2000-2002) and follow-ups (Exams 2-4, 2002-2008). Among these subjects, 3,965 were enrolled in the MESA Lung study, and underwent cardiac CT and gold-standard full-lung HRCT scanning in Exam 5

(2010-2012). This large dataset provides an invaluable opportunity for *population-level* longitudinal study of emphysema texture patterns. Emphysema quantification on cardiac CT scans was shown in (Hoffman et al. 2009) to have high reproducibility, high correlation with full-lung HRCT-based measures, and significant associations with risk factors of lung disease in population studies (Oelsner et al. 2014).

As described in Chapter 4, in our previous study (Yang et al. 2016a) we have established robust emphysema quantification on MESA cardiac scans. However, a straightforward application of either unsupervised learning or labeling of cardiac scans with the pre-defined sLTPs is still hindered by the following factors:

1. CT image quality in MESA is heterogeneous. Cardiac scans in Exam 1-4 are axial CT scans collected using one type of EBT scanner and six types of MDCT scanners, while scans in Exam 5 were collected with helical CT scanners;
2. Cardiac scanning protocols differ from full-lung HRCT scans. The average slice thickness is 2.82 mm for cardiac scans and is 0.65 mm for full-lung HRCT scans. This leads to downgraded lung texture details in cardiac scans;
3. The field of view (FOV) in cardiac scans only includes bottom 70% of the lungs, which prevents generating precise location information of lung textures.

To label sLTPs in cardiac scans, one option could be to register cardiac and full-lung scans, and learn a discriminative model, such as a convolutional neural network (CNN) model, to classify the cardiac ROIs using the sLTP labels in registered full-lung scans as ground truth. However, such registration is challenging given the differences in FOVs. Moreover, cardiac CT scans in earlier MESA exams do not have paired full-lung HRCT

scans, thus generalization to heterogeneous scanner types cannot be guaranteed with such approach.

To solve this problem, we propose the following original pipeline: we first synthesize cardiac CT scans from full-lung HRCT scans with ground truth sLTP labels and learn a CNN model for sLTP labeling on these synthetic cardiac scans. The synthetic scans inevitably present domain differences from real cardiac CT scans. For robust sLTP labeling on real cardiac scans, we further propose a CNN-based unsupervised domain adaptation with adversarial learning (UDAA). The UDAA module aims to fool an auxiliary domain discriminator at differentiating synthetic and real cardiac scans, thus enabling to learn domain-invariant feature representations. We apply our proposed UDAA framework to label 4,315 pairs of longitudinal cardiac CT scans, and test its robustness with respect to image domain differences.

5.2 Method

5.2.1 Data Cohort and Preprocessing

We exploit the ten sLTPs identified in the full-lung HRCT scans from SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS) (Couper et al. 2013), as described in Chapter 3. SPIROMICS is our reference dataset. It involves more than 55% of subjects (more than 1,800 subjects) that have COPD, including 616 severe COPD cases. Then we aim to label the MESA cohort, which has less disease cases.

sLTP labeling requires a pre-segmentation of emphysema regions. We use the hidden Markov measure field (HMMF) model for emphysema segmentation as described in Chap-

ter 4, which was shown to be able to handle scanner and subject variability in the HRCT and cardiac MESA scans (Yang et al. 2016a). Regions of interest (ROIs) are extracted from cardiac scans, with a size of $25\times25\times25$ mm^3 ($36\times36\times8$ voxels) and with percent emphysema larger than the upper limit of normal (ULN) of emphysema. The ULN values were defined by the reference equation derived from healthy never-smokers to account for known demographical differences in percent emphysema (Hoffman et al. 2014).

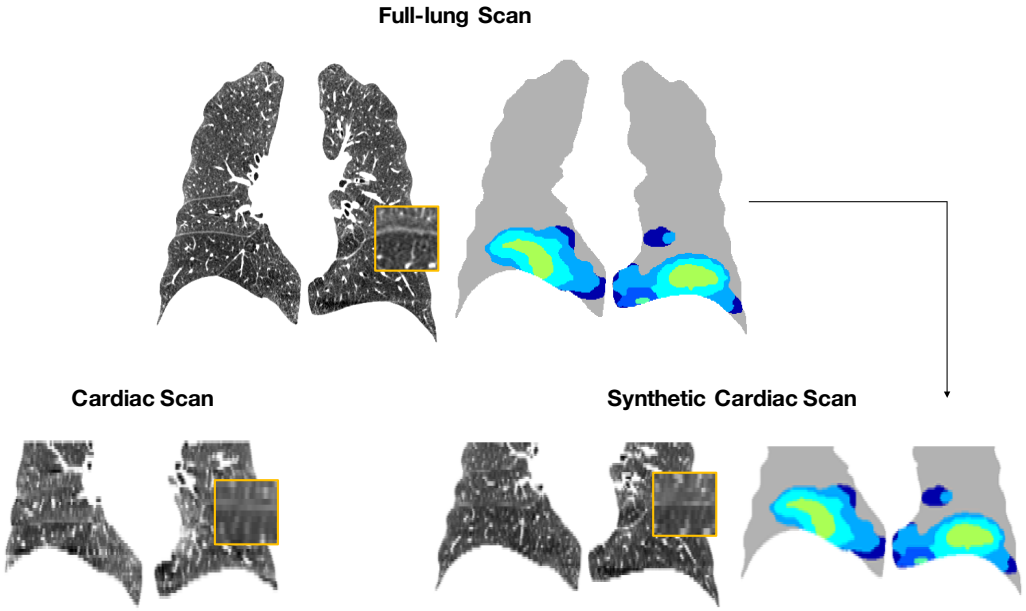

Figure 5.1: Illustration of the generation of synthetic cardiac CT scans and ground-truth sLTP labeling from full-lung CT scans in MESA Exam 5. Compared to full-lung CT scans, cardiac scans have lower spatial resolution and thus down-graded texture quality, as illustrated in the yellow-boxed patches that are zoomed in. Synthetic cardiac CT scans are generated by down-sampling full-lung images along the superior-inferior axis with a factor of 5, and keeping the bottom 2/3 of the lung.

From the full-lung HRCT scans equipped with ground-truth sLTP label maps, we generate synthetic cardiac CT data by down-sampling images along the superior-inferior axis with a factor of 5 (the ratio of average full-lung vs. cardiac scan slice thickness), and keeping the bottom 2/3 of the lung. The superior FOV cutting planes of intra-subject longitu-

Table 5.1: Number of MESA cardiac CT scans along with splits used to train and evaluate the proposed unsupervised domain adaptation with adversarial training (UDAA) framework.

Exam	# of Helical scans (train | val | test)	# of MDCT scans (train | val | test)	# of EBT scans (train | val | test)	Total # of scans
Exam 1	-	934 | 329 | 2,245	748 | 260 | 2,167	6,683
Exam 2	-	408 | 150 | 983	271 | 102 | 967	2,881
Exam 3	-	423 | 151 | 699	459 | 152 | 772	2,656
Exam 4	-	152 | 54 | 241	245 | 83 | 380	1,155
Exam 5	1,146 | 422 | 414	-	-	1,982

The initial training, validation and test split is based on full-lung HRCT scans in MESA Exam 5. Then in longitudinal cardiac exams, subjects that do not belong to the training or validation sets and do not have HRCT scans in Exam 5 (i.e. unseen during training and validation stages) are added into the longitudinal test sets.

dinal cardiac and synthetic cardiac scans are aligned, by segmenting the main bronchi in each scan as the physiological landmark, finding the superior axial slice levels that best match, and cutting all FOVs to the matched levels. Examples of a full-lung HRCT scan, a cardiac CT scan, and a synthetic cardiac CT scan are illustrated in Fig. 5.1 in coronal views. In our evaluation, we exclude scans without ULN values, or did not pass the FOV cropping algorithm. That leads to a final set of N = 15,357 longitudinal cardiac scans in MESA as detailed in Table 5.1.

The proposed UDAA framework to label LTPs in cardiac scans involves three main components. First, we train a CNN model to classify emphysema ROI from synthetic cardiac CT scans. The CNN model consists of a fully-convolutional component as image feature extractor, and a fully-connected component for LTP classification. Simultaneously,

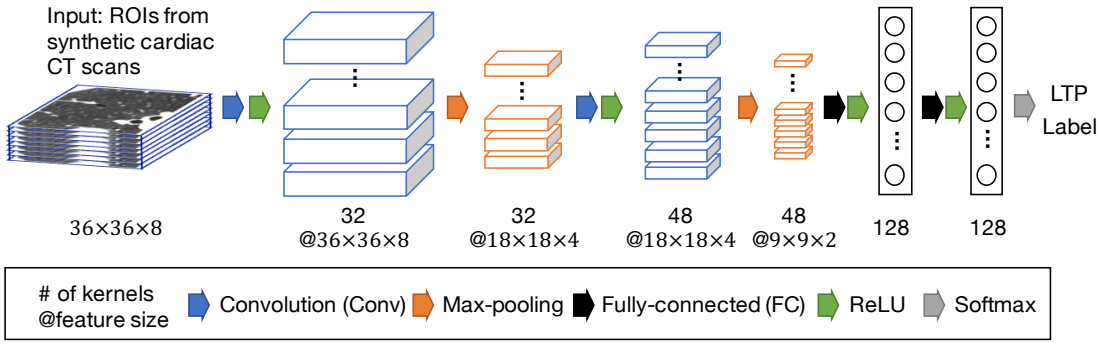

Figure 5.2: Illustration of the basic CNN architecture for sLTP classification on synthetic cardiac CT scans. The network contains two interleaved convolutional and max-pooling layers with 3D operations, and two fully-connected layers.

we extract features of real cardiac ROIs, and train a domain discrimination component. The UDAA framework is optimized such that the feature extractor can generate domain-invariant features, which are discriminative for the LTP classification task with synthetic cardiac ROIs, but able to fool the domain discriminator thus enabling robust LTP labeling in real cardiac ROIs. We now detail the UDAA model in the following sections.

5.2.2 CNN to Label Synthetic Cardiac ROIs

We propose to train the sLTP classifier on ROIs from synthetic cardiac scans, which are associated with ground-truth labels without requiring registration. Our CNN model used for sLTP labeling is illustrated in Figure 5.2.

Typically, a CNN alternatively stacks convolutional (Conv) layers and sub-sampling layers (e.g., max-pooling layers). In a Conv layer, small feature extractors (kernels) sweep over the topology and transform the input into feature maps (called activation maps). By denoting the i-th feature map of the l-th layer as h_i^l, and the k-th feature map of the

previous layer as h_k^{l-1}, a Conv layer is formulated as:

$$h_i^l = \sigma(\sum_k h_k^{l-1} * W_{ki}^l + b_i^l) \tag{5.1}$$

where W_{ki}^l and b_i^l are the filter and bias terms that connect the feature maps of adjacent layers, $*$ denotes the convolution operation, and σ is the element-wise non-linear activation function after the convolution operation.

In a max-pooling layer, activations within a neighborhood are summarized to acquire invariance to local translations, and down-sampled (controlled by the stride parameter) for dimension reduction. After several Conv and max-pooling layers, feature maps are flattened into a feature vector (thus dismissing the spatial relationship), followed by fully-connected (FC) layers, which is formulated as:

$$h^l = \sigma(W^l h^{l-1} + b^l) \tag{5.2}$$

where $h^{l-1} \in \mathbb{R}^P$ and $h^l \in \mathbb{R}^Q$ are the feature vectors in the *l-1* and l-th FC layers, $W^l \in \mathbb{R}^{Q \times P}$ is the weight matrix and b^l denotes the bias term. Finally, a softmax classification layer yields the prediction probability.

Our CNN model consists of interleaved Conv layers and max-pooling layers, and finally FC layers. In our implementation, we choose small-sized kernels (3×3×3) in Conv and max-pooling layers to integrate information in small 3D neighborhood, following the size of textons as in (Yang et al. 2017). A stride of 2 is used in max-pooling layers. Rectified linear units (ReLU) are used for the non-linear activation functions.

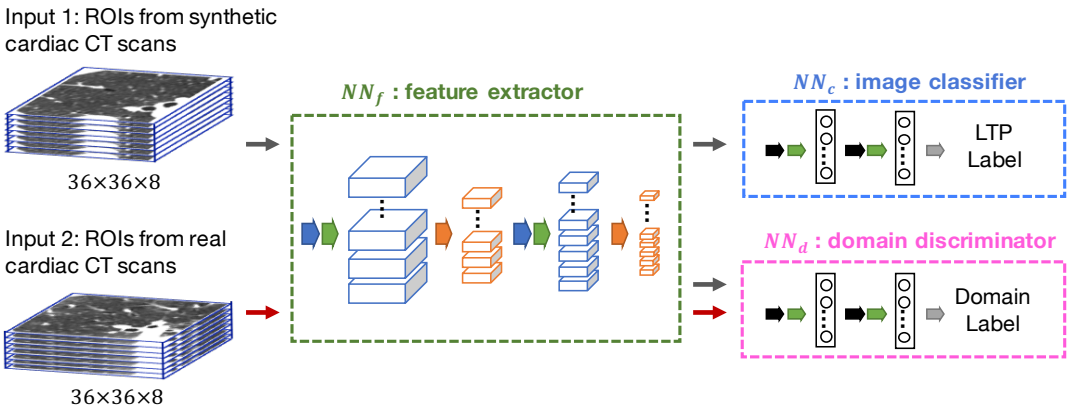

Figure 5.3: Illustration of the proposed unsupervised domain adaptation with adversarial training (UDAA) for sLTP labeling on cardiac CT scans in MESA. Compared to the basic CNN model, the UDAA model connects an domain adaptation component to the feature extractor via a gradient reversal layer to learn discriminative image features between synthetic and real cardiac scans.

To be noted that, we use a relatively shallow architecture, consisting of two 3D Conv layers and two FC layers, which proved sufficient in our experiments to achieve stable and high-accuracy sLTP labeling without introducing too much variance. The number of kernels in each layer is indicated in Figure 5.2, which were selected via grid search.

We carry over this CNN model to the UDAA module in Figure 5.3, and split it into the fully-convolutional component NN_f, as the feature extractor, and the FC component NN_c, as sLTP classifier.

5.2.3 UDAA Module

Adversarial Domain Discriminator

A common assumption in machine learning is that training and test data are drawn from the same probability distribution, which may not be true for the synthetic cardiac scans in training, and real cardiac scans in test data. Therefore, we propose to use unsupervised

and adversarial domain adaptation introduced in (Ganin et al. 2016) for the labeling of ROIs from real cardiac CT scans.

Unsupervised domain adaptation is a type of method that learns from samples and associated labels in a source domain D_S (in our case, they refer to the ROIs from synthetic cardiac scans):

$$S = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n_S} \sim D_S \tag{5.3}$$

and applies the learned model to unannotated samples from a different target domain D_T (here, ROIs from real cardiac cardiac scans), accounting for domain differences:

$$T = \{\boldsymbol{x_i}\}_{i=1}^{n_T} \sim D_T \tag{5.4}$$

The goal of the learning task is to build a classifier η with a low target risk (Kamnitsas et al. 2017):

$$R(\eta) = Pr_{(\boldsymbol{x},y)\sim D_T}(\eta(\boldsymbol{x}) \neq y) \tag{5.5}$$

The proposed adversarial training module is illustrated in Figure 5.3. The CNN feature extractor NN_f is used to generate feature representations $h(\cdot)$ for both synthetic and real cardiac ROIs. Only features from synthetic ROIs are used to train the sLTP classifier NN_c. The NN_f and NN_c form together a standard feed-forward network architecture.

Then, a third neural network component, called the domain discriminator (noted NN_d), is added into the framework, which takes all $h(\cdot)$ as input and tries to determine whether it comes from D_S or D_T. Such discrimination serves as an indicator of how much source-specific the representation $h(\cdot)$ is.

The domain discriminator NN_d is connected to the feature extractor NN_f via a gradient reversal layer (Ganin et al. 2016), which multiplies the gradient by a negative constant during the backpropagation to NN_f. Gradient reversal ensures that feature distributions over the two domains D_S and D_T are forced to be similar, thus resulting in domain-invariant features.

The sLTP classifier NN_c is optimized using a categorical cross-entropy loss function L_{class}, whereas the domain discriminator NN_d is optimized using a binary cross-entropy loss function L_{domain}. The adversarial training process minimizes the total loss function L_{total} which simultaneously maximizes domain discrimination loss and minimizes sLTP classification loss, as:

$$L_{total} = L_{class} - \alpha L_{domain} \tag{5.6}$$

where α is a positive weight that defines the relative importance of the domain-adaptation task for the sLTP classifier. This optimization is possible with regular stochastic gradient descent (SGD), given that NN_d is interconnected and gradients of L_{domain} can propagate back through the discriminator and into the feature extractor. During training, α is initiated at 0 and is gradually increased up to α_{max} using the following schedule suggested by (ibid.):

$$\alpha = \frac{2 \cdot \alpha_{max}}{1 + \exp(-\gamma \cdot p)} - 1 \tag{5.7}$$

where γ was set to 10, and p is the training progress, linearly increasing from 0 to 1. This strategy allows the NN_d to be less sensitive to noisy signal at the early training stages. The weight α_{max} is determined by maximizing the $ACC_{total} = ACC_{class} - ACC_{domain}$ metric in the validation set, where ACC_{class} is the sLTP classification accuracy and ACC_{domain}

is the domain classification accuracy.

Domain Discrimination in Longitudinal Setting

While the domain discriminator does not require registered inputs in D_S and D_T, there is a risk that the learning process is being driven by population-differences rather than domain-differences if the sampling is not regulated. We therefore enforce sampling of ROIs per training batch to come from the same subjects and similar locations, matching the relative distance vectors $\boldsymbol{d_i}$ between the center voxels of ROIs $\boldsymbol{x_i}$ to the lung mask bounding box (hence not using fine registration).

In our longitudinal setting, NN_d needs to discriminate synthetic cardiac ROIs in MESA Exam 5 from real cardiac ROIs in earlier exams. We further constrain the ROIs sampling for training NN_d, such that the percent emphysema difference is less than 5% for two ROIs $\boldsymbol{x_i} \in S$ and $\boldsymbol{x_j} \in T$, if $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ come from same subjects longitudinal scans with $|\boldsymbol{d_i} - \boldsymbol{d_j}| < 0.1$. This excludes pairing ROIs with drastic changes in inflation level or emphysema progression, which may introduce some bias when training NN_d.

5.3 Experimental Results

5.3.1 Experimental Setting

We use all full-lung HRCT scans and cardiac CT scans in MESA that have ULN values of percent emphysema (Hoffman et al. 2014). This resulted in N = 2,837 subjects in full-lung Exam 5, which we randomly divide into training S_{train}, validation S_{val} and test S_{test} sets with a ratio of 3:1:1.

In longitudinal cardiac exams, subjects belonging to S_{train} and S_{val} are used to op-

timize the UDAA framework, while subjects belonging to S_{test} and subjects not having HRCT full-lung Exam 5 (i.e. unseen during training and validation stages) are used as longitudinal test sets to evaluate the sLTP labeling performance. The final number of cardiac scans evaluated in this study is reported in Table 5.1.

To measure the influence of the domain-adaptation task in the UDAA module, we also test a source-only CNN training (i.e. training only NN_f and NN_c using the source synthetic scans, and applying the trained model to real cardiac scans). In validation steps, we measure ACC_{domain} for the source-only CNN model by adding a domain classifier similar to NN_d, but setting its gradient to zero when backpropagating to NN_f, so that the domain classifier does not impose any effect on the feature extractor.

5.3.2 Training and Validation Based on Local ROIs

We trained the UDAA domain adaptation module between synthetic cardiac scans in MESA Exam 5 and three target real cardiac scanning conditions (i.e. domains) that exist in MESA Exam 1-5:

1. Helical CT scans in Exam 5;
2. MDCT scans in Exam 1-4;
3. EBT scans in Exam 1-4.

We report in Table 5.2 the overall average values for ACC_{class}, ACC_{domain} and the associated α_{max} for our best model in S_{val}.

From Table 5.2, we can see that the UDAA module is able to generate CNN features that are less distinguishable by the domain discriminator (corresponding to lower ACC_{domain}

Table 5.2: Comparison of validation accuracy for sLTP labeling and domain classification, using the proposed CNN with and without domain adaptation module.

	ACC_{class}		ACC_{domain}	
Domain	CNN	UDAA	CNN	UDAA (α_{max})
Helical	0.898	0.888	0.655	0.560 (4.5)
MDCT		0.875	0.764	0.639 (7.0)
EBT		0.867	0.829	0.603 (7.5)

ACC_{class} = sLTP classification accuracy on synthetic ROIs;

ACC_{domain} = domain classification accuracy of real vs. synthetic ROIs.

values), while resulting in only minor decrease in the accuracy of sLTP classification task on the source synthetic domain.

5.3.3 sLTP Labeling on Longitudinal Cardiac Scans

We labeled all the test cardiac scans in MESA Exam 1-5. Longitudinal consistency of sLTP labeling is evaluated on intra-subject pairs of scans (cardiac vs. synthetic or cardiac vs. cardiac) acquired within a maximum time lapse of 48 months.

Reproducibility of sLTP Histograms

We compare sLTP histograms between the following pairs of scans:

1. Synthetic cardiac scans in Exam 5 (ground truth) vs. real cardiac scans in Exam 5;
2. Synthetic cardiac scans in Exam 5 vs. longitudinal cardiac scans in Exam 4;
3. Longitudinal pairs of real cardiac scans in baseline and follow-up visits from Exams 1-4 (denoted as ExB vs. ExF), with a time lapse shorter than 48 months.

In MESA Exams 1-4, we report separate measures for MDCT and EBT scanners, to evaluate the robustness of our proposed framework versus scanner types. Longitudinal

Table 5.3: Evaluation of reproducibility of sLTP labeling on longitudinal scan pairs in MESA acquired within a time lapse $\leqslant$ 48 months.

	N_p	N_k	χ^2 Distance		Correlation: Mean [Min, Max]	
			CNN	UDAA	CNN	UDAA
Ex5-Synthetic Ex5-Helical	369	8	2.81	**2.41**	0.73 [0.46, 0.90]	**0.79** **[0.52, 0.90]**
Ex5-Synthetic Ex4-MDCT	51	6	4.61	**3.33**	0.50 [0.19, 0.79]	0.60 [0.28, 0.81]
Ex5-Synthetic Ex4-EBT	73	6	5.10	4.60	0.57 [0.14, 0.81]	0.59 [0.17, 0.82]
ExB-MDCT ExF-MDCT	1,812	7	1.81	1.76	0.80 [0.67, 0.86]	0.82 [0.68, 0.90]
ExB-EBT ExF-EBT	1,839	10	2.15	2.10	0.76 [0.55, 0.90]	0.77 [0.59, 0.92]
ExB-EBT ExF-MDCT	171	8	4.69	**3.12**	0.53 [0.15, 0.84]	**0.67** **[0.42, 0.85]**

Bold = significantly better performance ($p < 0.05$).

scan pairs with overall non-emphysema changing by more than 20% were excluded from our evaluation.

The following metrics are used to compare the sLTP labeling histograms between pairs of CT scans:

1. Subject-level χ^2 distance between sLTP histograms;
2. sLTP-level intra-class correlations between synthetic cardiac scans and real cardiac scans in MESA Exam 5;
3. sLTP-level Pearson correlations for longitudinal pairs of scans.

Results are reported in Table 5.3, with correlation coefficients computed only on the N_k sLTPs that were present in at least twenty pairs of scans (out of N_p pairs of scans

being tested), for the sake of statistical power. Reported sLTP-level values consist of mean, minimum and maximum among the N_k sLTPs. Statistical differences between the source-only CNN model and the UDAA model were tested using t-test.

From Table 5.3, significant differences are observed between the two models generally when there is scanner-type change between longitudinal CT exams, and the proposed UDAA model always achieves better reproducibility. If there is no scanner-type change (same domain), reproducibility is similar for both models, which is expected.

Spatial Consistency of sLTP Labels

We further compare the spatial consistency of sLTP labeling in longitudinal cardiac scans. Given the potential inaccuracy of partial lung registration in cardiac scans, we report here a coarse measure, by dividing the bounding box of each left and right lung into 8 lung zones (superior/inferior, anterior/posterior, medial/lateral), thus construct a histogram of 8 bins for each sLTP location per lung.

Then we compute the χ^2 distances of local sLTP histograms between the longitudinal pairs of cardiac scans in Exam 1-4, with the same number of scans as studied in Table 5.3.

We again compare separately for ExB-MDCT vs. ExF-MDCT, ExB-EBT vs. ExF-EBT and ExB-EBT vs. ExF-MDCT scans. We found that χ^2 distances are significantly smaller (i.e. better) with the proposed UDAA model, compared to the original CNN model, for a number of 3, 4, 3 sLTPs respectively, while the other sLTPs do not show significant difference between UDAA and the source-only CNN model.

5.4 Discussion and Conclusion

In this chapter, we presented an unsupervised domain adaptation with adversarial learning (UDAA) framework to label MESA cardiac CT scans with emphysema-specific lung texture patterns (sLTPs) previously learned on full-lung HRCT scans, as described in Chapter 3. Translation to cardiac scans relies on *synthetic* cardiac scans generated from MESA HRCT scans previously labeled with sLTPs. These labels were used as our ground-truth for the supervised training component. Domain adaptation was exploited in an unsupervised context to transition from *synthetic* to real cardiac scans.

Comparison of sLTP labeling on intra-subject pairs of (cardiac, full-lung HRCT) scans from the same scanning session, and on longitudinal cardiac scans acquired within a maximum time lapse of 48 months, showed significantly better consistency with the UDAA framework than naively training a CNN model with synthetic-only data. To our knowledge, this is the first study on longitudinal subtyping of emphysema patterns on cardiac CT scans. Such tool can enable large-scale multi-sites longitudinal studies of emphysema subtypes over 10 years follow-up, and could potentially advance the understanding of emphysema progression and COPD.

Chapter 6

Discriminative Localization in CNNs for Weakly-Supervised Detection of Pulmonary Nodules

6.1 Introduction

Lung cancer is a major cause of cancer-related deaths worldwide. Pulmonary nodules refer to a range of lung abnormalities that are visible on lung computed tomography (CT) scans as roughly round opacities, and have been regarded as crucial indicators of primary lung cancers (MacMahon et al. 2005). The detection and segmentation of pulmonary nodules in lung CT scans can facilitate early lung cancer diagnosis, timely surgical intervention and thus increase survival rate (Henschke et al. 1999).

Automated detection systems that locate nodules of various sizes can assist radiologists in cancer malignancy diagnosis (Sluimer et al. 2006). Existing supervised approaches for automated nodule detection and segmentation require voxel-level annotations for training, which are labor-intensive and time-consuming to obtain. Alternatively, image-level labels, such as a binary label indicating the presence of nodules in a relatively larger field of view, can be obtained more efficiently.

Recent work (Anirudh et al. 2016; Messay, Hardie, and Tuinstra 2015) studied nodule segmentation using weakly labeled data without dense voxel-level annotations. Their

methods, however, still rely on user inputs for additional information such as exact nodule location and estimated nodule size during the segmentation.

Convolutional neural networks (CNNs) have been widely used for supervised image classification and segmentation tasks. It was recently discovered in a study (Zhou et al. 2016) on natural images that CNNs trained on semantic labels for image classification task ("what"), have remarkable capability in identifying the discriminative regions ("where") when combined with a global average pooling (GAP) operation. This method utilizes the up-sampled weighted activation maps from the last convolutional layer in a CNN. The localization capability of CNNs was demonstrated for detecting relatively large-sized targets within natural image, which is not the general scenario in medical imaging domain where pathological changes are more various in size and rather subtle to capture. However, this work sheds light on weakly-supervised disease detection.

In this chapter, we exploit CNN for accurate and fully-automated segmentation of nodules in a weakly-supervised manner with binary slice-level labels only. Specifically, we adapt classic image classification CNN models to detect slices with nodule, and simultaneously learn the discriminative regions from the activation maps of convolution units at different scales for coarse segmentation. We then introduce a candidate-screening framework utilizing the same network to generate accurate localization and segmentation. Experimental results on the public LIDC-IDRI dataset (Armato III et al. 2011; Clark et al. 2013) demonstrate that, despite the largely reduced amount of annotations required for training, our weakly-supervised nodule segmentation framework achieves competitive performance compared to a CNN-based fully-supervised segmentation method.

The nodule locations are informative for indicating malignancy, which is important in

lung cancer screening programs. Therefore, we further explore the spatial distribution of nodules with different malignancy scores in the LIDC-IDRI dataset, using the lung shape spatial mapping PDCM that we proposed in Chapter 3. Then we utilize the CNN-based activation maps proposed in this chapter to predict the malignancy risk of lung CT scans in the Kaggle DSB2017 dataset. We will explore and discuss more explicit characterizations of lung nodule malignancy risks in Appendix D.

6.2 Method

The proposed method for weakly-supervised lung nodule detection contains two stages: a CNN training stage to generate robust and discriminative activation maps, as illustrated in Fig. 6.1, and a nodule candidate screening stage to refine the detection results, as illustrated in Fig. 6.2.

In the first stage, we train a CNN model to classify CT slices as with or without nodule. The CNN is composed of a fully convolutional component, a convolutional layer + global average pooling layer (Conv+GAP) structure, and a final fully-connected (FC) layer. Besides providing a binary classification, the CNN generates a nodule activation map (NAM) showing potential nodule localizations, using a weighted average of the activation maps with the weights learnt in the FC layer.

In the second stage, coarse segmentation of nodule candidates is generated within a spatial scope defined by the NAM. For fine segmentation, each nodule candidate is masked out from the image alternately. By feeding the masked image into the same network, a residual NAM (called R-NAM) is generated and used to select the true nodule. Shallower layers in the CNN can be concatenated into the classification task through skip architec-

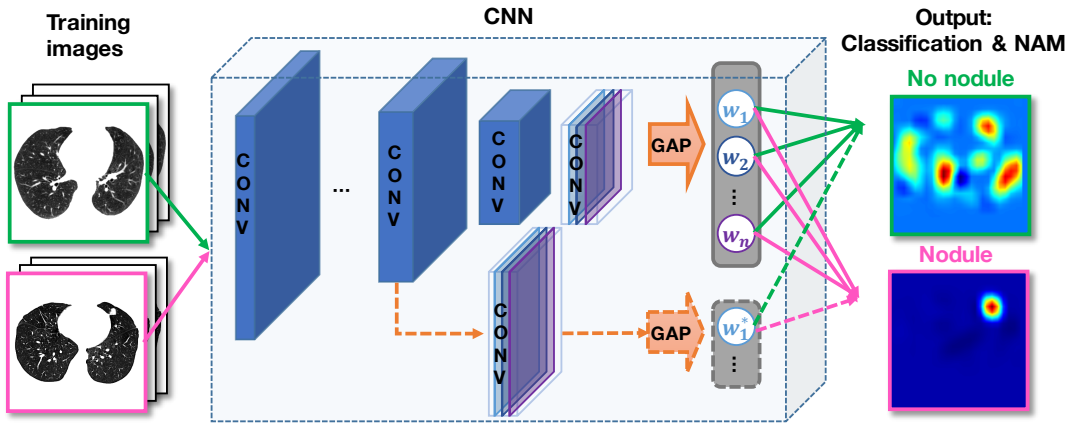

Figure 6.1: Illustration of the proposed framework to generate nodule activation maps (NAMs): a CNN model is trained to classify CT slices. A global average pooling (GAP) operation is used to summarize the activation maps of a convolutional (Conv) layer into a scalar. The final fully-connected (FC) layer will estimate weights to weigh the activation maps to generate the nodule activation maps (NAMs). Besides the last Conv layer, shallower Conv layers can also be connected to the final FC layer via GAP operations.

ture and Conv+GAP structure, extending the one-GAP CNN model to multi-GAP CNN that is able to generate NAMs with higher resolution.

The location information of lung nodule is informative for indicating malignancy (as described in Section 6.3.3). Therefore, the NAMs with discriminative regions of lung nodules can also be used to coarsely estimate the risk of lung cancer. We will discuss about this application in Section 6.2.4 and Section 6.3.4.

6.2.1 Nodule Activation Map

In a classification-oriented CNN, while the shallower layers represent general appearance information (color, edge, texture, etc.), the deep layers encode discriminative information that is specific to the classification task.

Benefiting from the convolutional structure, spatial information can be retained in the activations of convolutional units. Activation maps of deep convolutional layers, there-

fore, enable discriminative spatial localization of the class/object of interest. In our case, we locate nodules with a specially generated weighted activation map called nodule activation map (NAM).

One-GAP CNN

For a given image I, we represent the activation of unit k at spatial location (x, y) in the last convolutional layer as $a_k(x, y)$. The activation of each unit k is summarized through a spatially global average pooling operation as $A_k = \sum_{(x,y)} a_k(x, y)$.

The feature vector constituted of A_k is followed by a FC layer, which generates the nodule classification score (i.e. input to the softmax function for nodule class) as:

$$S_{\text{nodule}} = \sum\nolimits_k w_{k,\text{nodule}} A_k = \sum\nolimits_k w_{k,\text{nodule}} \sum\nolimits_{(x,y)} a_k(x, y) \tag{6.1}$$

where the weights $w_{k,\text{nodule}}$ learnt in the FC layer essentially measure the importance of unit k in the classification task.

As spatial information is retained in the activation maps through $a_k(x, y)$, a weighted average of the activation maps results in a robust nodule activation map:

$$\text{NAM}(x, y) = \sum\nolimits_k w_{k,\text{nodule}} a_k(x, y) \tag{6.2}$$

The nodule classification score can be directly linked with the NAM by:

$$S_{\text{nodule}} = \sum\nolimits_{(x,y)} \sum\nolimits_k w_{k,\text{nodule}} a_k(x, y) = \sum\nolimits_{(x,y)} \text{NAM}(x, y) \tag{6.3}$$

By simply up-sampling the NAM to the size of the input image I, we can identify the discriminative image region that is most relevant to nodule.

Multi-GAP CNN

Although activation maps of the last convolutional layer carry most discriminative information, they are usually greatly down-sampled from the original image resolution due to pooling operations. We hereby introduce a multi-GAP CNN model that takes advantage of shallower layers with higher spatial resolution.

Similar to the idea of the skip architecture used in fully-convolutional network (FCN) (Shelhamer, Long, and Darrell 2016) for semantic segmentation, and multiple state-of-the-art image classification networks (He et al. 2016; Huang et al. 2017), shallower layers can be directed to the final classification task skipping the following layers. We also add a Conv+GAP structure following the shallow layers. The concatenation of feature vectors generated by each GAP layer is fed into the final FC layer. The NAM generated from the multi-GAP CNN model (multi-GAP NAM) is a weighted activation map involving activations at multiple scales.

6.2.2 Lung Nodule Detection and Segmentation

Detection Scope for Coarse Segmentation

For slices classified as "nodule slice", nodule candidates are screened within a spatial scope C defined by the most prominent blob in the NAM processed via watershed. They are then coarsely segmented based on the CT intensity values and using an iterated conditional mode (ICM) based multi-phase segmentation method (Israel-Jost et al. 2008), with the

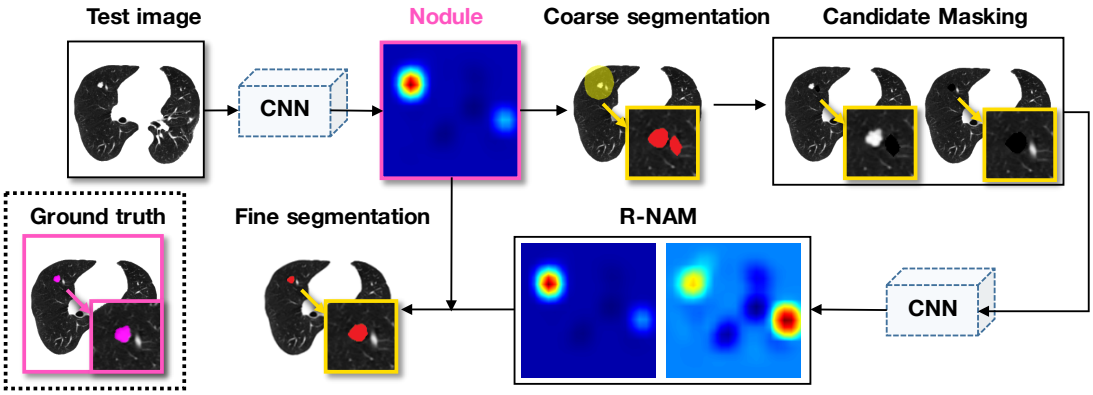

Figure 6.2: Illustration of the proposed lung nodule candidate screening framework: for test slices classified as "nodule slice", nodule candidates are screened using a spatial scope defined by the NAM for coarse segmentation. Residual NAMs (R-NAMs) are generated from images with masked nodule candidates for fine segmentation.

phase number equal to four as determined by global intensity distribution.

Candidate Screening for Fine Segmentation

The NAM indicates a potential but not exact nodule location. To identify the true nodule from the coarse segmentation results, i.e. which nodule candidate triggered the activation, we generate residual NAMs (R-NAMs) by masking each nodule candidate R_j alternately and feeding the masked image $I \backslash R_j$ into the same network. The most significant change of activations within C indicates the exclusion of the true nodule. Formally, we generate the fine segmentation by selecting the nodule candidate R_k following:

$$R_k = \text{argmax}_{R_j} \sum_{(x,y)\in C} \left[\text{NAM}_I(x,y) - \text{NAM}_{I\backslash R_j}(x,y)\right]^2 \tag{6.4}$$

where NAM_I is the original NAM, and $\text{NAM}_{I\backslash R_j}$ is the R-NAM generated by masking nodule candidate R_j. Our current implementation targets the segmentation of one nodule per NAM. Incidence of slices with two nodules is $\sim 1\%$ within slices with nodules in the

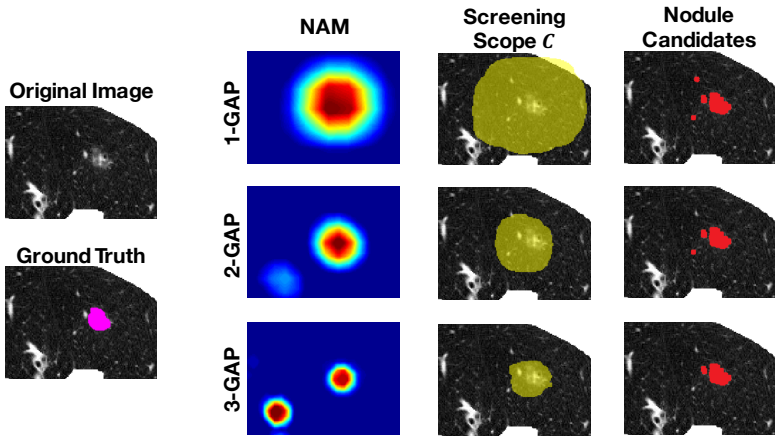

Figure 6.3: Illustration of 1-/2-/3-GAP NAMs, the screening scopes C and coarse segmentation results on a sample slice. The 1-GAP NAM is most discriminative showing only one high probability lung nodule region, while the 3-GAP NAM is least discriminative. Constraining the 3-GAP NAM with the screening scope defined on the 1-GAP NAM, the number of nodule candidates is reduced from four to two.

LIDC-IDRI dataset. No slices contain more than two nodules in this dataset.

Multi-GAP Models

For the multi-GAP CNN model, we observed a slight drop in classification accuracy compared with the one-GAP CNN model (see Section 6.3.2), which is expected since features from shallower layers are more general and less discriminative.

In light of this, we further propose a multi-GAP segmentation method by training both a one-GAP CNN model and a multi-GAP CNN model to combine the discriminative capability of the one-GAP system and finer localization of the multi-GAP system.

Specifically, we first detect "nodule slice" by the one-GAP CNN model for its higher classification accuracy. To define the screening scope for coarse segmentation, we first use the one-GAP NAM to generate a baseline scope C_1. If there is a prominent blob C_{multi} detected via watershed within C_1 in the multi-GAP NAM, we define the final scope C as C_{multi} to eliminate redundant nodule candidates with more localized spatial constraints.

When the multi-GAP NAM fails to identify any discriminative regions within C_1, the final screening scope C remains C_1. The R-NAM of the masked image is generated by the one-GAP CNN model and compared with one-GAP NAM within C_1.

Fig. 6.3 illustrates 1-/2-/3-GAP NAMs, the corresponding screening scopes C and coarse segmentation results on a sample slice. While multi-GAP NAM enables finer localization, one-GAP NAM has better discriminative power.

6.2.3 Network Architectures

Our CNN models for generating the NAMs are primarily based on the VGG-16 architecture, as illustrated in Fig. 6.4 (A). The VGG network is the winner of the localization task in ImageNet Large Scale Visual Recognition Competition (Krizhevsky, Sutskever, and Hinton 2012), or ILSVRC, in 2014. The last pooling layer pool5 and the FC layers fc6, fc7, fc8 in the original network are removed (Zhou et al. 2016). The Conv+GAP structure is added after conv5_3 layer for 1-GAP CNN, added after conv5_3 and conv4_3 layers for 2-GAP CNN, and added after conv5_3, conv4_3, and conv3_3 layers for 3-GAP CNN. In addition to the VGG-16 architecture, we further exploit deeper state-of-the-art CNN architectures such as the ResNet-50 (He et al. 2016, winner of ILSVRC 2015) and DenseNet-121 (Huang et al. 2017, winner of the best paper award in CVPR 2017), as discussed in Section 6.3.2.

We compare our model with a fully-supervised CNN method based on U-net architecture (Ronneberger, Fischer, and Brox 2015), as illustrated in Fig. 6.4 (B). U-net is a encoder-decoder type of network architecture that has been extensively used in biomedical applications to detect cancer (Dalmis et al. 2018), kidney pathologies (Thong et al. 2018) and tracking cells (Rad et al. 2018). In the encoder part, feature maps from Conv

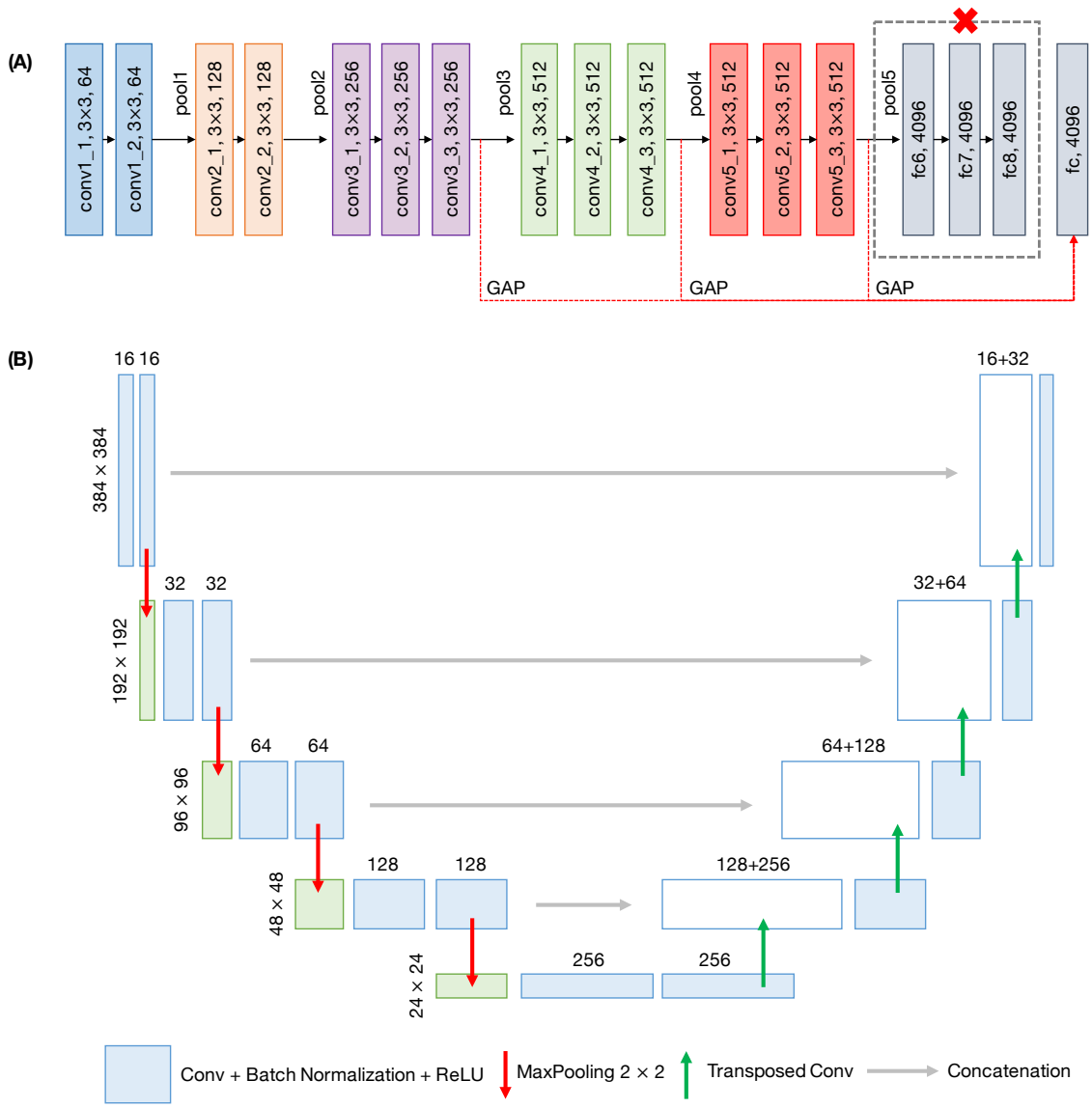

Figure 6.4: Illustration of the VGG-16 network architecture used for weakly-supervised lung nodule detection, and the U-net architecture used for fully-supervised detection. (A) The VGG-16 network, where the last max-pooling layer pool5 and the fully-connected layers fc6, fc7, fc8 in the original work (Simonyan and Zisserman 2014) are removed. Global average pooling (GAP) layers are added as indicated by the red dashed lines. (B) The U-net architecture. This figure is adapted from (Ronneberger, Fischer, and Brox 2015).

layers are down-sampled via max-pooling operations. In the decoder part, feature maps are up-sampled via transpose Conv operations to maintain high resolution in the final feature map. The skip connections concatenate the layers in the down-sampling path with corresponding layers in the up-sampling path. The U-net was proven to be a very

powerful segmentation tool in scenarios with limited data, and it has no restriction on the size of the input image since it does not involve any fully connected layer.

6.2.4 NAM-based Cancer Map

In addition to providing potential discriminative regions of lung nodules, the NAMs can be used as efficient nodule-specific feature representations of the lung CT images. Since nodule location is one risk factor of lung cancer (Horeweg et al. 2014), the potential nodule locations in NAM may advance the estimation of nodule malignancy risk.

Being able to predict lung cancer versus normal or benign cases is important for building up a cost-effective lung cancer screening program, and is gaining more and more attention recently. In the Kaggle Data Science Bowl 2017 (DSB2017) challenge[1], full-lung CT scans of 2,101 subjects that have high risk of lung cancer are provided. Each subject is associated with a binary label which is the pathological diagnosis result of with/without early lung cancer. The goal of this challenge is to accurately predict the malignancy risk of the test CT scans, given training and validation CT scans and their binary labels. In this application, each CT scan is treated as a data point for the final prediction task. Therefore, an efficient and low dimensional feature representation of the CT scan is required.

To efficiently summarize the lung nodule locations within the lungs, we propose to generate local NAMs (pre-trained in LIDC-IDRI) per CT image slice in Kaggle DSB2017, and then average and project them onto coronal, axial and sagittal planes for dimension reduction. Notice that the NAMs from the deep-most Conv layer is a natural low dimensional feature, and do not need to be upscaled to the initial image size in this application.

[1]https://www.kaggle.com/c/data- science-bowl-2017

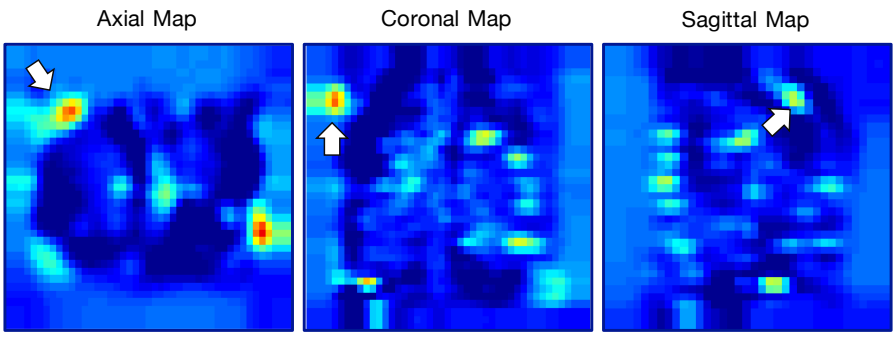

Figure 6.5: Illustration of the NAM-based cancer map on a sample lung CT scan. (Left) Axial projection of averaged NAMs; (Middle) Coronal projection of averaged NAMs; (Right) Sagittal projection of averaged NAMs.

Instead, we rescale the coronal, axial and sagittal NAMs to the same size (32×32 in our implementation) to be able to concatenate them into a single feature representation. We call the concatenation of the three projections the cancer map, which provides potential lung nodule location information from 3D perspectives.

Figure 6.5 is the illustration of a cancer map for one sample lung CT scan, which indicates the potential presence of a relatively large nodule in the superior, anterior and peel region of the right lung.

In the classification stage, we feed the cancer map per CT scan as the input to train a second CNN model. Considering the relatively small sample size in Kaggle DSB2017, this CNN model contains only two Conv layers (kernel size $= 3 \times 3$, followed by ReLU and max-pooling) with 32 and 48 kernels respectively, and then two FC layers (followed by ReLU) of size 128 and 256 respectively. Dropout (Srivastava et al. 2014) and batch normalization (Ioffe and Szegedy 2015) are used to prevent overfitting of the prediction.

6.3 Experimental Results

6.3.1 Data and Experimental Setup

Data used for evaluating the weakly-supervised lung nodule detection contains 1,010 thoracic CT scans from the public LIDC-IDRI database (see details in Chapter 2). Lungs were segmented and each axial slice was cropped to 384×384 pixels centering on the lung mask. Data used for evaluating the proposed NAM-based cancer map includes the 2,101 thoracic CT scans from the Kaggle DSB2017 dataset.

In the LIDC-IDRI dataset, nodules were delineated by up to four experts. Voxel-level annotations are used to generate slice-level labels, and are used as ground truth for nodule detection and segmentation evaluation. Nodules with diameter $<$ 3mm are excluded (Setio et al. 2016). Given the high false positive rate of nodule detection, we select slices with nodule if there were overlapped annotations by at least two experts, and select slices without nodule if no expert reported a nodule in the slice. We merge annotations from different experts using the STAPLE algorithm (Warfield, Zou, and Wells 2004). A total of $N_{\text{slice}} = 8,345$ slices with nodule are selected, and an equal number of slices without nodule are randomly extracted. Training, validation and test sets are generated by distributing the full set of subjects in a ratio of 4:1:1 through stratified sampling so that they have non-overlapping subjects and similar distribution of nodule occurrence. The total number of voxels belonging to nodule is $N_{\text{voxel}} = 1,658,981$. The number of labeling required for a fully-supervised method versus our method is $N_{\text{voxel}}/N_{\text{slice}} \sim 100$.

The training, validation and test split of Kaggle DSB2017 follows the official split,

which consists of 1,397 training scans, 198 validation scans (stage 1) and 506 test scans (stage2), to evaluate the proposed cancer maps.

Our weakly-supervised nodule detection and segmentation is primarily trained with slice-level weak labels in LIDC-IDRI dataset. Evaluations of nodule detection and segmentation performance are focused on slices with one nodule. Rare cases of slices with two nodules are discussed in Appendix C. Potential extension to detecting nodules in 3D regions is also preliminarily evaluated and discussed in Appendix C.

6.3.2 Nodule Detection and Segmentation Performance

Performance Compared to Fully-Supervised U-Net Method

Our weakly-supervised CNN model based on VGG-16 network is initialized with weights pre-trained on ImageNet. The learning rate of the newly added FC layers is 10 times the learning rate of the remaining VGG-16 layers. We trained using stochastic gradient descent with momentum (Qian 1999). The initial learning rate (10^{-2} for 1-GAP, 2×10^{-3} for 2-GAP, 10^{-3} for 3-GAP), learning decay (0.99), batch size (30) were set by grid search based on classification accuracy on the validation set. The best accuracy values are 88.4% for 1-GAP CNN, 86.6% for 2-GAP model, and 84.4% for 3-GAP model on the test set.

For U-net, the cost function is the negative mean Dice coefficient across mini-batch. The algorithm was optimized with Adam method (Kingma and Ba 2014). The initial learning rate (2×10^{-4}), learning decay (0.999), and batch size (20) were determined with grid search based on average Dice on the validation set. The U-net model is trained on voxel-level labels for nodule segmentation. For detection performance, a slice is labeled as "without nodule" if there is no segmented nodule in the slice.

Table 6.1: Comparison of nodule detection and segmentation performance between the proposed weakly-supervised models and fully-supervised U-net model

Method	TPR	FPR	FPR_{nodule}	Dice mean ± SD	TP Dice mean ± SD	TP DOA mean ± SD
1-GAP Coarse	**0.77***	**0.11***†	-	0.46 (±0.31)	0.61 (±0.20)	57.6 (±71.1)
2-GAP Coarse	0.76	-	-	0.50 (±0.34)	0.66 (±0.18)	41.6 (±53.6)
3-GAP Coarse	0.75	-	-	0.50 (±0.32)	0.67 (±0.18)	40.1 (±50.9)
1-GAP Fine	0.75	-	**0.14***	0.54 (±0.34)	0.73 (±0.15)	30.7 (±52.8)
2-GAP Fine	0.75	-	0.14	0.55* (±0.33)	0.74* (±0.14)	29.2* (±46.8)
3-GAP Fine	0.74	-	0.15	0.54 (±0.34)	0.74 (±0.14)	29.3 (±46.4)
U-net	0.74	0.29	0.26	**0.56** (±0.38)	**0.76** (±0.19)	**28.3** (±44.8)

* = best performance within our framework; **boldfaced** = overall best performance;

†= 1-GAP model is used for nodule slice-level detection within our framework.

True positive rate (TPR) of nodule detection, false positive rate (FPR) of "nodule" segmented on slices *without* nodule, false positive rate (FPR_{nodule}) of "nodule" segmented on slices *with* nodule, Dice overlap of nodule segmentation over all slices with nodule (Dice), Dice over truly detected nodules (TP Dice) and absolute difference of segmented areas over truly detected nodules (TP DOA, mm^2) are reported in Table 6.1. The proposed method achieves the best nodule detection performance over all models with regard to the TPR and false positive rates, and achieves competitive segmentation performance compared to the fully-supervised U-net model.

Since detection and segmentation performance are closely related to the nodule size, TP Dice and TP DOA versus nodule size are reported in Fig. 6.6.

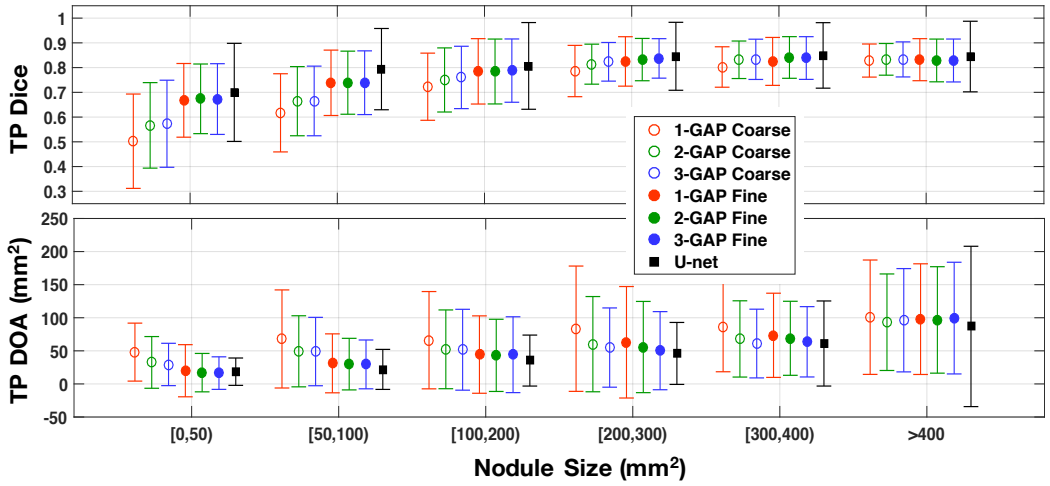

Figure 6.6: Comparison of nodule segmentation performance, measured by Dice over truly detected nodules (TP Dice) and absolute difference of segmented areas over truly detected nodules (TP DOA) (mean and standard deviation), versus nodule size between the proposed weakly-supervised method and fully-supervised U-net model.

Comparison of Performance over Different CNN Architectures

We further compare the weakly-supervised nodule detection performance, using CNN architectures based on VGG-16, ResNet-50 and DenseNet 121. The latter two networks were trained to generate NAMs in their deep-most Conv layers, using the same data split as the VGG-16, and again with training parameters determined via grid searching. The slice classification performance in validation and test sets and nodule detection performance in test set using the 1-GAP coarse segmentation is reported and compared in Table 6.2.

From Table 6.2, The best slice classification accuracy (ACC) in validation set is achieved by the DenseNet-121. During training, we also observe that the model convergence is fastest with the DenseNet-121 model, which is likely benefiting from its smaller number of trainable parameters (DenseNet is a type of deep CNN architecture with reusable features among layers in the same dense block), and hence fewer flexibility and smaller variance

Table 6.2: Comparison of nodule detection performance for the weakly-supervised NAM method versus CNN architectures

Method	Validation ACC	Test ACC	TPR	FPR	# of Parameters
VGG-16	0.887	0.884	0.77	0.11	15,242,050
ResNet-50	0.891	0.870	0.78	0.17	25,687,938
DenseNet-121	0.892	0.873	0.74	0.13	8,089,154

compared to the other two networks. While ResNet-50, which has the most parameters, is harder to train and needs careful initialization and hyper-parameter searching in our experiments. The model with the best validation ACC is applied to the test set, and we observe that VGG-16 has the best generalizability (highest test ACC) in our application. As for the nodule detection performance in test set, the ResNet-50 model achieves slightly higher TPR than the VGG-16 model, but with worse FPR, while the DenseNet-121 is least discriminative than the other two models.

We illustrate the NAMs generated with the three CNN architectures on sample CT slices, as shown in Fig. 6.7. The NAMs generated by the DenseNet-121 model has larger prominent blobs (hence larger nodule screening scopes), which is due to the lower resolution of activation maps in the last Conv layer in DenseNet-121 (down-sampled by 32 from the original image size), compared to the other two networks (down-sampled by 16).

Therefore, the VGG-16 architecture appears to be more suited (easier to train, and with overall higher classification accuracy and better detection performance) for our lung nodule detection task within this relatively small sample size.

To inspect the CNN features learned in the three CNN models, we visualize the kernels in Conv layers via activate maximization (Erhan et al. 2009). The idea behind activation

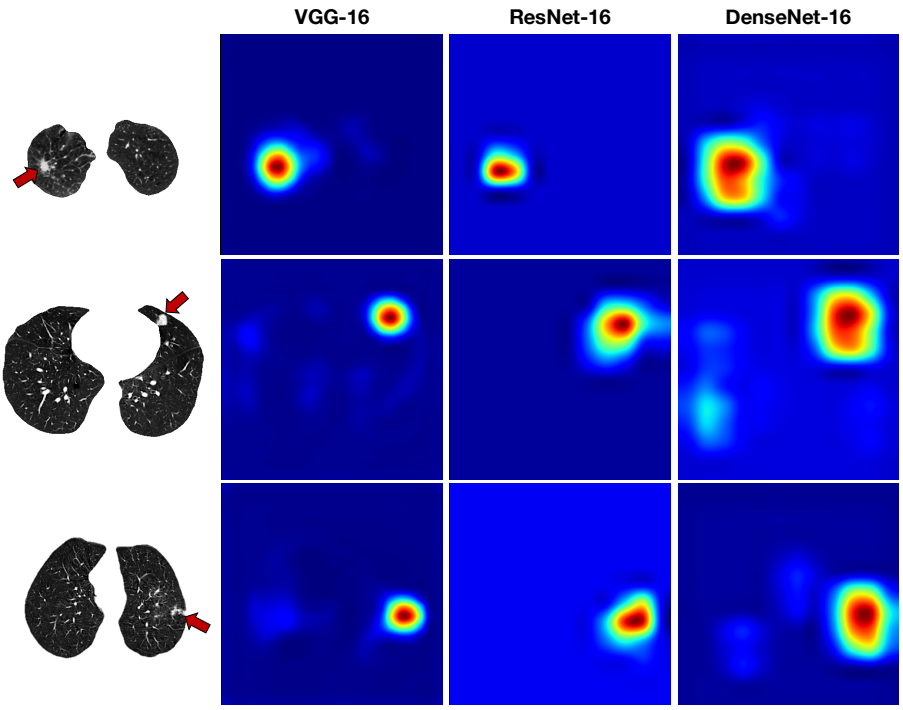

Figure 6.7: Visualization of NAMs generated with different CNN architectures on sample CT slices. From left to right: CT image slices, NAMs based on VGG-16, ResNet-50 and DenseNet-121.

maximization is to generate an input image that maximizes the filter output activations. This allows us to understand what sort of input patterns activate a particular filter.

For each CNN model, we select four Conv layers (see details in Appendix B), and visualize five random filters in each layer, as shown in Fig. B.1-B.3. From the visualizations, the shallower layers generally exhibit basic texture information, and deeper layers exhibit more complicated textures, and the very deep layers are less interpretable since they are operated on outputs from all previous layers. For that reason, we observe that the visualizations on the deeper ResNet-50 and DenseNet-121 models are less interpretable than that of the VGG-16 model.

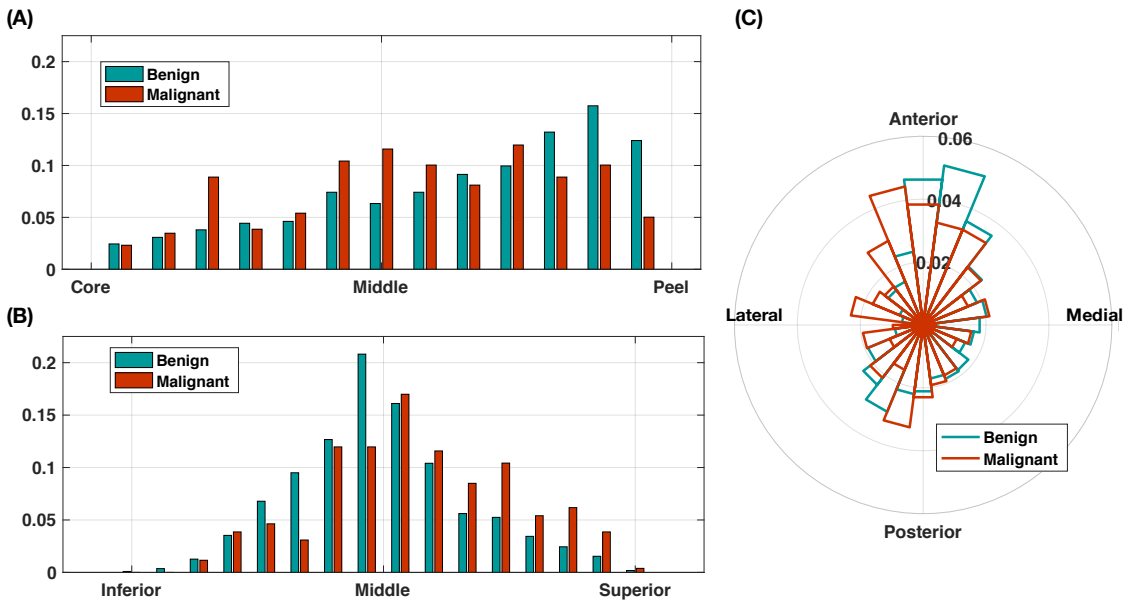

Figure 6.8: Spatial distribution of benign versus malignant nodules in the LIDC-IDRI dataset, measured by the PDCM spatial mapping.

6.3.3 Nodule Spatial Distribution in LIDC-IDRI

The nodule location information is a risk factor of lung cancer. In Chapter 3, we proposed a lung shape spatial mapping, the Poisson distance conformal mapping (PDCM), which can be used as a tool to study the lung nodule locations in a standardized coordinate system. Therefore, we apply the PDCM to all CT scan in the LIDC-IDRI dataset.

Each nodule in LIDC-IDRI is associated with one to four malignancy scores, ranging from 1 (very benign) to 5 (very malignant), estimated by four radiologists. We define the nodules with the majority score $\geqslant 4$ to be malignant and the rest to be benign. This results in N = 1,506 benign nodules and N = 285 malignant nodules. The distribution of the two nodule types from external lung surface (peel) to the lung core, superior lung to inferior lung, and anterior-lateral-posterior-medial directions are shown in Fig. 6.8.

From Fig. 6.8, benign nodules are more likely to locate at peel regions compared

to malignant nodules evaluated via *t*-test ($p < 10^{-6}$) as shown in Fig. 6.8 (A), while malignant nodules are more likely to locate at superior regions of the lungs ($p < 10^{-8}$) as shown Fig. 6.8 (B), which agrees with the conclusion that nodules in upper lung lobes have higher risk to be cancer (Horeweg et al. 2014; MacMahon et al. 2005). The spatial distribution in axial directions, as in Fig. 6.8 (C), does not show significant difference between benign and malignant nodules ($p > 0.05$).

6.3.4 Cancer Map in Kaggle DSB17

We applied the cancer map model proposed in Section 6.2.4 to predict the lung cancer risk, and participated the Kaggle DSB2017 challenge. For evaluation, we compute the log loss as used in this challenge:

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{6.5}$$

where n is the number of patients in the test set, $\hat{y}_i$ is the predicted probability of the CT image belonging to a patient with cancer, and y_i is 1 if the diagnosis is cancer, 0 otherwise.

Our model achieves a log loss of 0.45872, which ranks the 14th out of 1,972 teams in the Kaggle challenge.

6.4 Discussion and Conclusion

Automated systems that locate lung nodules of various sizes can assist radiologists in lung cancer diagnosis, and has gained more and more attentions. In this chapter, we first presented an original design for lung nodule detection and segmentation, extending a classification-trained CNN model with GAP operations, to learn discriminative regions at

different resolution scales utilizing only weakly labeled training data (present or not of a lung nodule). Coarse-to-fine segmentation extracted nodule candidates, and determined the true nodule exploiting a novel candidate-screening framework.

Compared with voxel-based labels, the number of labeling required for our method was reduced by $\sim$100 times. Detection performance of our weakly-supervised framework compared very favorably with a fully-supervised CNN model (higher TPR and lower FPR). Our average segmentation accuracy on detected nodules was also very high and got very close to the benchmark method for larger nodules. Fully-supervised CNN achieved, on average, more accurate segmentation when correctly detecting the nodule, which is expected since voxel-level annotation utilized during training provides more power to deal with various intensity patterns, especially at edges. On the other hand, standard deviations were smaller with the proposed method, hence indicates fewer large mistakes.

A machine learning model requiring only weakly-labeled data is key for a sustainable development of CAD systems, as expert time is scarce and expensive and as scanners continue to evolve significantly. Our work use transfer learning from a CNN trained on natural images; with more annotated data, it will be possible to train a fully dedicated network that is likely to be more effective.

The nodule locations are informative for indicating malignancy. In Chapter 3, we proposed a standardized lung shape spatial mapping PDCM. In this chapter, we studied the spatial location of nodule types (benign versus malignant) in the LIDC-IDRI dataset again using PDCM. Significant differences were observed in the spatial distributions between nodule types, from external lung to the core regions, and from superior lung to inferior lung regions. Lung nodule location can therefore be a useful radiological factor when

estimating the nodule malignancy risks.

In addition to providing potential discriminative regions of lung nodules, the NAMs proposed in this work can be used as efficient nodule-specific feature representations of the lung CT images. In the Kaggle DSB2017 challenge, we predicted early lung cancer risks using NAM-based cancer maps, and ranked 14th our of 1,972 teams. Lung cancer screening has been approved and is being implemented in the United States. Challenges present given the large number of high-risk individuals to screen, and the dominance of benign nodules among all nodules detected on CT. Automated methods that enable estimating early lung cancer risks based on CT images are important for a more cost-effective lung cancer screening program with less reading efforts and minimal number of biopsies. Our NAM-based cancer maps demonstrated the potential of detecting early lung cancers, trained with scan-level binary diagnostic labels, which are easy to acquire in clinical scenarios.

With detected nodules, we can perform more explicit characterizations to estimate nodule malignancy, which can provide richer information (than scan-level prediction) to assist radiologists in decision making and cancer management. In Appendix D, we will discuss our proposed method to explicitly classify given nodules as benign versus malignant, using a novel data augmentation framework based on generative adversarial network (GAN) to deal with the class imbalance over the two nodule types.

Chapter 7

Discussion and Conclusion

CT imaging continues to be the most important tool for assessing the parenchymal structure in the lungs. In this thesis, we presented novel methods for CT-based lung texture learning, targeting to significantly advance the understanding and diagnosis of two important lung diseases: 1) chronic obstructive pulmonary disease (COPD) and pulmonary emphysema; 2) lung cancer and pulmonary nodules.

Most existing CT-based lung texture learning methods to date have been limited to supervised approaches relying on manually annotated voxels or local regions of interest (ROIs) as ground truth, which are slow and labor-intensive to obtain. In this work, we exploited unsupervised and weakly-supervised learning requiring less or no annotations.

More specifically, we first developed an unsupervised machine learning method to discover novel CT image-based patterns for pulmonary emphysema, incorporating spatial and texture features. While there were three standard emphysema subtypes previously defined at autopsy, pathologists disagreed on the very existence of them. Clinical usages of the three standard subtypes have been limited, largely due to practical inter- and intra-rater variability when assessing them on CT. Our proposed unsupervised learning enabled us to discover a set of quantitative emphysema subtypes (QES) that were highly reproducible, and were associated independently with respiratory symptoms, clinical charac-

teristics and genetic variants, as demonstrated on large full-lung CT datasets including 2,922 subjects in the SPIROMICS and 3,128 subjects in the MESA Lung Study.

Then we extended the lung texture learning to the well-established MESA cardiac CT dataset, to enable large-scale longitudinal study of emphysema. Cardiac CT scans include approximately bottom 2/3 of the lung. Automated emphysema quantification methods have been available for decades, and the standard and adapted methods have achieved widespread acceptance for research purposes on full-lung high-resolution CT (HRCT) scans; while robust emphysema quantification on cardiac scans was not well exploited previously. In this thesis, we presented a framework which demonstrated the potential of robust emphysema segmentation across heterogeneous cardiac CT scans by dedicated parameterization to account for scanner and subject variability. Then we proposed a deep learning method based on unsupervised domain adaptation to handle the texture differences between full-lung HRCT and cardiac CT scans, which significantly increased the consistency of texture learning across imaging scanners and protocols (the "domains"). This enabled us to study the progression of QES on 17,039 longitudinal cardiac and full-lung CT scans over 10 years of follow-up in MESA.

Overall the discovered QES provide novel emphysema sub-phenotyping that may facilitate future study of emphysema development, understanding the stages of COPD and the design of personalized therapies.

To facilitate lung cancer screening, we presented a weakly-supervised deep learning method for lung nodule detection, as an alternative to the fully-supervised methods relying on voxel-level delineations that require high expense and reading efforts. The proposed method generated weighted nodule activation maps (NAMs) from convolu-

tional neural networks (CNNs) with skip-connections, and incorporated a novel candidate screening framework to reduce the number of false positives. Using the proposed weakly-supervised method, we achieved competitive performance compared to a fully-supervised method, while requiring $\sim$100 times less annotations. With the increasing scale of medical image data and the challenge of reading efforts, relying on massive ground truth labels to develop machine learning algorithms for computer-aided diagnosis (CAD) is becoming less practical. Alternatively, weakly-supervised learning is more suitable for a sustainable development of CAD systems, in the presence of heterogeneous data generated by constantly evolving scanner types.

The proposed NAMs are also efficient nodule-specific descriptions of lung CT images. Using NAM-based features, we were able to predict the risk of early lung cancer per subject when training on scan-level binary labels. This demonstrates the potential usage of deep learning methods for CT-based lung cancer prediction, in scenarios where training diagnostic information is available at individual-level. When annotations are available at nodule-level, we can perform more explicit characterizations to estimate nodule malignancy, which can provide richer information to assist radiologists in decision making and cancer management. The vast majority of lung nodules detected on lung CT scans are eventually benign, hence malignant nodules are generally underrepresented in existing datasets. Therefore, we proposed a novel data augmentation framework with class-aware nodule synthesis, to handle the issue of class imbalance when classifying benign versus malignant nodules. Our nodule synthesis enabled in-painting nodules with high fidelity in specified categories at any location of the lung, and was demonstrated to be beneficial for predicting nodule malignancy scores. Similar framework can also be applied to aug-

ment data for lung nodule detection, and can be extended to synthesizing other disease patterns and facilitate more fundamental problems such as lung and lobe segmentation, in the presence of rare diseases.

Overall, the proposed work enables the usage of a vast amount of unannotated and weakly annotated CT scans. Successful applications would potentially have a tremendous impact in the field, for diseases that affects millions around the world.

Unsupervised and weakly-supervised learning will continue to be the focus in future work, accounting for the practical challenge in annotation acquisition in medical domains. Investigating the variation of lung texture subtypes in normal and mild disease populations can be important to discover early disease signals, and will be considered in our future study, utilizing the MESA Lung Study, a dataset for general population. Further development is possible to refine the current emphysema subtyping and yield novel molecular quantitative emphysema subtypes (mQES), by incorporating other data and features, such as pulmonary vasculature on CT angiograms, expiratory lung CTs, longitudinal data points, and genetic factors.

In the future work, dedicated deep learning frameworks can be exploited for incorporating the multimodal data. Several progresses have been made with deep neural networks for multimodal learning in medical studies (Ramachandram and Taylor 2017). Recurrent layers can capture spatio-temporal relationships of input data, and can be utilized to characterize longitudinal data points. Shared representation layers are commonly used for feature fusion in deep learning to work with multi-modal data. Regularization of such multi-modal features to enforce inter- and intra-modality correlations is an active field of research. Deep generative models, such as deep belief network (DBNs, Lee et al. 2009),

GANs, and deep variational models, have been proposed to characterize joint statistical distributions of observed data and their associated classes, and are better suited for unsupervised and weakly-supervised settings. Layer-wise pre-training is useful for initializing unsupervised deep networks. Pre-defined layers based on wavelet transforms (Cheng, Chen, and Mallat 2016) were demonstrated to be powerful to help prevent overfitting when the training data set is relatively small. This will be exploited in future study.

Deep-learning associating imaging (radiomics) and genomics is still widely unexplored (Miotto et al. 2017). Bayesian deep learning has been introduced in computer vision (Kendall and Gal 2017) to model aleatoric uncertainty (i.e. uncertainty inherent to the data) by placing a distribution over the output of the model, and epistemic uncertainty (i.e. uncertainty in the model parameters) by putting prior distributions over weights. Exploiting recurrent networks and Bayesian deep learning for mixing radiomics (existing QES) with dynamic longitudinal data and static genetic data may reveal additional emphysema subtypes that suggest mechanistic pathways to treatment.

Emphysema and airway diseases jointly contribute to COPD. The genetic hits in our discovered Apical QES suggested the association between emphysema and pulmonary vasoconstriction, which is consistent with the previous suspicions that pulmonary vasculature plays a role in COPD development (Hueper et al. 2015). Therefore, it is important to exploit precision approaches of tree structures in lung CT, to study both airway trees and vasculature trees in high-resolution CT. Joint modeling the texture and spatial variability of emphysema and pulmonary trees over populations may help us to understand the contributions from emphysema, airway diseases, and possibly vascular abnormalities to COPD, and will be investigated in future studies.

Chapter 8

Related Publications and Competition Performance

8.1 Publications

Published / In Press:

- **Jie Yang** and et al., "Unsupervised Domain Adaption with Adversarial Learning (UDAA) for Emphysema Subtyping on Cardiac CT Scans: The MESA Study", *in IEEE International Symposium on Biomedical Imaging (**ISBI**)*, 2019.
- **Jie Yang** and et al., "Class-Aware Adversarial Lung Nodule Synthesis in CT Images", *in IEEE International Symposium on Biomedical Imaging (**ISBI**)*, 2019.
- Y. Gan, **Jie Yang** and et al., "Enhanced Generative Model for Unsupervised Discovery of Spatially-Informed Macroscopic Emphysema: The Mesa COPD Study", *in IEEE International Symposium on Biomedical Imaging (**ISBI**)*, 2019.
- X. Feng, **Jie Yang** and et al., "Alzheimer's Disease Diagnosis based on Anatomically Stratified Texture Analysis of the Hippocampus in Structural MRI", *in IEEE International Symposium on Biomedical Imaging (**ISBI**)*, 2018.
- C. Aaron, ..., **Jie Yang** and et al.: "A Longitudinal Cohort Study of Aspirin Use and Progression of Emphysema-like Lung Characteristics on CT imaging: the MESA Lung Study", *in **CHEST***, 2018.

- **Jie Yang** and et al., "Unsupervised Discovery of Spatially-Informed Lung Texture Patterns (sLTPs) for Pulmonary Emphysema: The MESA COPD Study", *in International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*, 2017.
- X. Feng*, **Jie Yang*** and et al., "Discriminative Localization in CNNs for Weakly-supervised Segmentation of Pulmonary Nodules", *in International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*, 2017.
  (* denotes equally contributed first author)
- J. Song, **Jie Yang** and et al., "Generative Method to Discover Emphysema Subtypes with Unsupervised Learning using Lung Macroscopic Patterns (LMPS): The MESA COPD study", *in IEEE International Symposium on Biomedical Imaging (**ISBI**)*, 2017.
- **Jie Yang** and et al., "Emphysema Quantification on Cardiac CT Scans using Hidden Markov Measure Field Model: the MESA Lung Study", *in International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*, 2016.
- **Jie Yang** and et al., "Explaining Radiological Emphysema Subtypes with Unsupervised Texture Prototypes: MESA COPD Study", *in International Conference on Medical Image Computing and Computer-Assisted Intervention Medical Computer Vision Workshop (**MICCAI MCVW**)*, 2016.
- **Jie Yang** and et al., "Texton and Sparse Representation based Texture Classification of Lung Parenchyma in CT Images", *in IEEE Annual International Conference of the Engineering in Medicine and Biology Society (**EMBC**)*, 2016.

Under Review:

- **Jie Yang** and et al., "Characterizing Alzheimer's Disease with Image and Genetic Biomarkers using Supervised Topic Models", *in Journal of Biomedical and Health Informatics (**JBHI**).*
- X. Feng, **Jie Yang** and et al.: "Deep Learning on MRI Affirms the Prominence of the Hippocampal Formation in Alzheimer's Disease Classification", *in **Brain** (bioRxiv:456277).*

To Submit:

- **Jie Yang** and et al., "Unsupervised Machine Learning to Define Quantitative Subtypes of Pulmonary Emphysema on CT", *in **Science** (to submit in January, 2019).*
- **Jie Yang** and et al., "Novel Subtypes of Pulmonary Emphysema Based on Spatially-Informed Lung Texture Learning", *in Transactions on Medical Imaging (**TMI**).*
- M. Wang, ..., **Jie Yang** and et al.: "Long-Term Exposure to Ambient Air Pollution and Longitudinal Change in Quantitatively Assessed Emphysema on Computed Tomography and Lung Function in the General Population: the MESA Air and Lung Studies", *in the Journal of the American Medical Association (**JAMA**) (under revision).*

8.2 Competition Performance

- Kaggle Data Science Bowl 2017: Can You Improve Lung Cancer Detection?

  14th out of 1972 teams

Bibliography

Ahmed, Firas S. et al. (2014). "Plasma sphingomyelin and longitudinal change in emphysema on CT. The MESA Lung study." In: *Biomarkers* 19.3, pp. 207–213. ISSN: 0954-6111.

Anderson, Augustus E et al. (1964). "Emphysema in lung macrosections correlated with smoking habits." In: *Science* 144.3621, pp. 1025–1026. ISSN: 0036-8075.

Anirudh, Rushil et al. (2016). "Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data." In: *SPIE Medical Imaging*. International Society for Optics and Photonics, pp. 978532–978532.

Anthimopoulos, Marios et al. (2016). "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network." In: *IEEE Transactions on Medical Imaging* 35.5, pp. 1207–1216.

Aoshiba, Kazutetsu, Naoko Yokohori, and Atsushi Nagai (2003). "Alveolar wall apoptosis causes lung destruction and emphysematous changes." In: *American Journal of Respiratory Cell and Molecular Biology* 28.5, pp. 555–562.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein gan." In: *arXiv preprint arXiv:1701.07875*.

Armato III, Samuel G et al. (2011). "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." In: *Medical Physics* 38.2, pp. 915–931. ISSN: 0094-2405.

Asherov, Marina, Idit Diamant, and Hayit Greenspan (2014). "Lung texture classification using bag of visual words." In: *SPIE Medical Imaging*. International Society for Optics and Photonics, 90352K–90352K–8.

Auerbach, Oscar et al. (1972). "Relation of smoking and age to emphysema: whole-lung section study." In: *New England Journal of Medicine* 286.16, pp. 853–857. ISSN: 0028-4793.

Barr, R Graham et al. (2007). "Impaired flow-mediated dilation is associated with low pulmonary function and emphysema in ex-smokers: the Emphysema and Cancer Action

Project (EMCAP) Study." In: *American Journal of Respiratory and Critical Care Medicine* 176.12, pp. 1200–1207.

Barr, R Graham et al. (2010). "Percent emphysema, airflow obstruction, and impaired left ventricular filling." In: *New England Journal of Medicine* 362.3, pp. 217–227. ISSN: 0028-4793.

Barr, R Graham et al. (2012). "A combined pulmonary-radiology workshop for visual evaluation of COPD: study design, chest CT findings and concordance with quantitative evaluation." In: *COPD* 9.2, pp. 151–159. ISSN: 1541-2555.

Bartel, Seth T et al. (2011). "Equating quantitative emphysema measurements on different CT image reconstructions." In: *Journal of Medical Physics* 38.8, pp. 4894–4902. ISSN: 0094-2405.

Bild, Diane E et al. (2002). "Multi-ethnic study of atherosclerosis: objectives and design." In: *American Journal of Epidemiology* 156.9, pp. 871–881. ISSN: 0002-9262.

Binder, Polina et al. (2016). "Unsupervised discovery of emphysema subtypes in a large clinical cohort." In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 180–187.

Bryan, Tracey L et al. (2012). "The effects of dobutamine and dopamine on intrapulmonary shunt and gas exchange in healthy humans." In: *Journal of Applied Physiology* 113.4, pp. 541–548.

Buzug, Thorsten M (2008). *Computed tomography: from photon statistics to modern cone-beam CT*. Springer Science & Business Media.

Carreira, Joao and Andrew Zisserman (2017). "Quo vadis, action recognition? a new model and the kinetics dataset." In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, pp. 4724–4733.

Ceresa, Mario et al. (2011). "Robust, standardized quantification of pulmonary emphysema in low dose CT exams." In: *Academic Radiology* 18.11, pp. 1382–1390.

Chabat, François et al. (2000). "Gradient correction and classification of CT lung images for the automated quantification of mosaic attenuation pattern." In: *Journal of Computer Assisted Tomography* 24.3, pp. 437–447. ISSN: 0363-8715.

Cheng, Xiuyuan, Xu Chen, and Stéphane Mallat (2016). "Deep Haar scattering networks." In: *Information and Inference: A Journal of the IMA* 5.2, pp. 105–133.

Cheung, Po-Yin and Keith J Barrington (2001). "The effects of dopamine and epinephrine on hemodynamics and oxygen metabolism in hypoxic anesthetized piglets." In: *Critical Care* 5.3, p. 158.

Choi, Yunjey et al. (2017). "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." In: *arXiv preprint* 1711.

Clark, Kenneth et al. (2013). "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository." In: *Journal of Digital Imaging* 26.6, pp. 1045–1057. ISSN: 0897-1889.

Couper, David et al. (2013). "Design of the subpopulations and intermediate outcomes in COPD study (SPIROMICS)." In: *Thorax*, thoraxjnl–2013.

Dahl, Morten et al. (2002). "Change in lung function and morbidity from chronic obstructive pulmonary disease in alpha1-antitrypsin MZ heterozygotes: a longitudinal study of the general population." In: *Annals of Internal Medicine* 136.4, pp. 270–279. ISSN: 0003-4819.

Dalmis, Mehmet Ufuk et al. (2018). "Fully automated detection of breast cancer in screening MRI using convolutional neural networks." In: *Journal of Medical Imaging* 5.1, p. 014502.

Davey, Claire et al. (2015). "Bronchoscopic lung volume reduction with endobronchial valves for patients with heterogeneous emphysema and intact interlobar fissures (the BeLieVeR-HIFi study): a randomised controlled trial." In: *The Lancet* 386.9998, pp. 1066–1073.

Depeursinge, Adrien et al. (2014). "Three-dimensional solid texture analysis in biomedical imaging: review and opportunities." In: *Medical Image Analysis* 18.1, pp. 176–196.

Detrano, Robert et al. (2008). "Coronary calcium as a predictor of coronary events in four racial or ethnic groups." In: *New England Journal of Medicine* 358.13, pp. 1336–1345.

Edge, John, George Simon, and Lynne Reid (1966). "Peri-acinar (paraseptal) emphysema: its clinical, radiological, and physiological features." In: *British journal of diseases of the chest* 60.1, pp. 10–16.

Ellis, Peter M and Rachel Vandermeer (2011). "Delays in the diagnosis of lung cancer." In: *Journal of Thoracic Disease* 3.3, p. 183.

Erhan, Dumitru et al. (2009). "Visualizing higher-layer features of a deep network." In: *University of Montreal* 1341.3, p. 1.

Galbán, Craig J et al. (2012). "Computed tomography–based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression." In: *Nature Medicine* 18.11, p. 1711.

Gangeh, Mehrdad J et al. (2010). "A texton-based approach for the classification of lung parenchyma in CT images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 595–602.

Ganin, Yaroslav et al. (2016). "Domain-adversarial training of neural networks." In: *Journal of Machine Learning Research* 17.59, pp. 1–35.

Gevenois, Pierre Alain et al. (1995). "Comparison of computed density and macroscopic morphometry in pulmonary emphysema." In: *American Journal of Respiratory and Critical Care Medicine* 152.2, pp. 653–657.

Ginsburg, Shoshana B et al. (2012). "Automated texture-based quantification of centrilobular nodularity and centrilobular emphysema in chest CT images." In: *Academic Radiology* 19.10, pp. 1241–1251.

Goodfellow, Ian et al. (2014). "Generative adversarial nets." In: *Advances in neural information processing systems*, pp. 2672–2680.

Gorelick, Lena et al. (2006). "Shape representation and classification using the poisson equation." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12, pp. 1991–2005.

Gulrajani, Ishaan et al. (2017). "Improved training of wasserstein gans." In: *Advances in Neural Information Processing Systems*, pp. 5767–5777.

Haidar, Haissam et al. (2006). "Characterizing the shape of anatomical structures with Poisson's equation." In: *IEEE Transactions on Medical Imaging* 25.10, pp. 1249–1257.

Häme, Yrjö et al. (2014). "Adaptive quantification and longitudinal analysis of pulmonary emphysema with a hidden Markov measure field model." In: *IEEE transactions on medical imaging* 33.7, pp. 1527–1540. ISSN: 0278-0062.

Häme, Yrjö et al. (2015a). "Equating emphysema scores and segmentations across CT reconstructions: A comparison study." In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 629–632.

Häme, Yrjö et al. (2015b). "Sparse sampling and unsupervised learning of lung texture patterns in pulmonary emphysema: MESA COPD study." In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 109–113.

Hankinson, John L, John R Odencrantz, and Kathleen B Fedan (1999). "Spirometric reference values from a sample of the general US population." In: *American journal of respiratory and critical care medicine* 159.1, pp. 179–187.

Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh (2018). "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pp. 18–22.

Hayhurst, MD et al. (1984). "Diagnosis of pulmonary emphysema by computerised tomography." In: *The Lancet* 324.8398, pp. 320–322.

He, Kaiming et al. (2016). "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Henschke, Claudia I et al. (1999). "Early Lung Cancer Action Project: overall design and findings from baseline screening." In: *The Lancet* 354.9173, pp. 99–105. ISSN: 0140-6736.

Hoffman, Eric A, Brett A Simon, and Geoffrey McLennan (2006). "State of the Art. A structural and functional assessment of the lung via multidetector-row computed tomography: phenotyping chronic obstructive pulmonary disease." In: *Proceedings of the American Thoracic Society* 3.6, pp. 519–532.

Hoffman, Eric A et al. (2003). "Characterization of the interstitial lung diseases via density-based and texture-based analysis of computed tomography images of lung structure and function 1." In: *Academic Radiology* 10.10, pp. 1104–1118.

Hoffman, Eric A et al. (2009). "Reproducibility and validity of lung density measures from cardiac CT scans - the Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study." In: *Academic Radiology* 16.6, pp. 689–699.

Hoffman, Eric A et al. (2014). "Variation in the percent of emphysema-like lung in a healthy, nonsmoking multiethnic sample. The MESA lung study." In: *Annals of the American Thoracic Society* 11.6, pp. 898–907.

Hogg, James C (2004). "Pathophysiology of airflow limitation in chronic obstructive pulmonary disease." In: *The Lancet* 364.9435, pp. 709–721.

Hong, Zhigang et al. (2005). "Pergolide is an inhibitor of voltage-gated potassium channels, including Kv1. 5, and causes pulmonary vasoconstriction." In: *Circulation* 112.10, pp. 1494–1499.

Horeweg, Nanda et al. (2014). "Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening." In: *The Lancet Oncology* 15.12, pp. 1332–1341.

Hu, Shiying, Eric A Hoffman, and Joseph M Reinhardt (2001). "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images." In: *IEEE Transactions on Medical Imaging* 20.6, pp. 490–498.

Huang, Gao et al. (2017). "Densely Connected Convolutional Networks." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2261–2269.

Hueper, Katja et al. (2015). "Pulmonary microvascular blood flow in mild chronic obstructive pulmonary disease and emphysema. The MESA COPD Study." In: *American journal of respiratory and critical care medicine* 192.5, pp. 570–580.

Hurst, John R et al. (2010). "Susceptibility to exacerbation in chronic obstructive pulmonary disease." In: *New England Journal of Medicine* 363.12, pp. 1128–1138.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In: *International Conference on Machine Learning*, pp. 448–456.

Israel-Jost, Vincent et al. (2008). "Vectorial multi-phase mouse brain tumor segmentation in T1-T2 MRI." In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 5–8.

Jin, Dakai et al. (2018). "CT-Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation." In: *arXiv:1806.04051*.

Jones, PW et al. (2009). "Development and first validation of the COPD Assessment Test." In: *European Respiratory Journal* 34.3, pp. 648–654.

Kamnitsas, Konstantinos et al. (2017). "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks." In: *International Conference on Information Processing in Medical Imaging*. Springer, pp. 597–609.

Kendall, Alex and Yarin Gal (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems*, pp. 5574–5584.

Kim, Song Soo et al. (2014a). "Improved correlation between CT emphysema quantification and pulmonary function test by density correction of volumetric CT data based on air and aortic density." In: *European Journal of Radiology* 83.1, pp. 57–63. ISSN: 0720-048X.

Kim, Victor et al. (2014b). "Clinical and computed tomographic predictors of chronic bronchitis in COPD: a cross sectional analysis of the COPDGene study." In: *Respiratory research* 15.1, p. 52.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980*.

Kirby, Miranda et al. (2015). "COPD: Do Imaging Measurements of Emphysema and Airway Disease Explain Symptoms and Exercise Capacity?" In: *Radiology*, p. 150037. ISSN: 0033-8419.

Klug, Harold P and Leroy E Alexander (1974). "X-ray diffraction procedures: for polycrystalline and amorphous materials." In: *X-Ray Diffraction Procedures: For Polycrystalline and Amorphous Materials, 2nd Edition, by Harold P. Klug, Leroy E. Alexander, pp. 992. ISBN 0-471-49369-4. Wiley-VCH, May 1974.* P. 992.

Korkinof, Dimitrios et al. (2018). "High-resolution mammogram synthesis using progressive generative adversarial networks." In: *arXiv:1807.03401*.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*, pp. 1097–1105.

Ladizinski, Barry and Christopher Sankey (2014). "Vanishing lung syndrome." In: *New England Journal of Medicine* 370.9, e14.

Laurie, Steven S et al. (2012). "Catecholamine-induced opening of intrapulmonary arteriovenous anastomoses in healthy humans at rest." In: *Journal of Applied Physiology* 113.8, pp. 1213–1222.

Lee, Honglak et al. (2009). "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations." In: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 609–616.

Leopold, JG and J Gough (1957). "The centrilobular form of hypertrophic emphysema and its relation to chronic bronchitis." In: *Thorax* 12.3, p. 219.

Liu, Ying et al. (2017). "Radiologic Features of Small Pulmonary Nodules and Lung Cancer Risk in the National Lung Screening Trial: A Nested Case-Control Study." In: *Radiology*, p. 161458.

Lynch, David A et al. (2015). "CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner Society." In: *Radiology* 277.1, pp. 192–205.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.Nov, pp. 2579–2605.

MacMahon, Heber et al. (2005). "Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society 1." In: *Radiology* 237.2, pp. 395–400. ISSN: 0033-8419.

Marchini, Jonathan et al. (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes." In: *Nature genetics* 39.7, p. 906.

Masutani, Yoshitaka, Ken Masamune, and Takeyoshi Dohi (1996). "Region-growing based feature extraction algorithm for tree-like objects." In: *Visualization in Biomedical Computing*. Springer, pp. 159–171.

McElvaney, Noel G et al. (2017). "Long-term efficacy and safety of $\alpha 1$ proteinase inhibitor treatment for emphysema caused by severe $\alpha 1$ antitrypsin deficiency: an open-label extension trial (RAPID-OLE)." In: *The Lancet Respiratory Medicine* 5.1, pp. 51–60.

Messay, Temesguen, Russell C Hardie, and Timothy R Tuinstra (2015). "Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset." In: *Medical Image Analysis* 22.1, pp. 48–62.

Mets, OM et al. (2012). "Quantitative computed tomography in COPD: possibilities and limitations." In: *Lung* 190.2, pp. 133–145. ISSN: 0341-2040.

Miller, Martin R et al. (2005). "Standardisation of spirometry." In: *European respiratory journal* 26.2, pp. 319–338.

Miotto, Riccardo et al. (2017). "Deep learning for healthcare: review, opportunities and challenges." In: *Briefings in bioinformatics*.

Murphy, Keelin et al. (2012). "Toward automatic regional analysis of pulmonary function using inspiration and expiration thoracic CT." In: *Medical Physics* 39.3, pp. 1650–1662. ISSN: 0094-2405.

National Lung Screening Trial Research Team (2011). "Reduced lung-cancer mortality with low-dose computed tomographic screening." In: *New England Journal of Medicine* 365.5, pp. 395–409.

Norman, Geoffrey R, Jeff A Sloan, and Kathleen W Wyrwich (2003). "Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation." In: *Medical care*, pp. 582–592.

Oelsner, Elizabeth C et al. (2014). "Association Between Emphysema-like Lung on Cardiac Computed Tomography and Mortality in Persons Without Airflow Obstruction A Cohort Study Emphysema-like Lung on CT and All-Cause Mortality." In: *Annals of Internal Medicine* 161.12, pp. 863–873.

Oelsner, Elizabeth C et al. (2016). "Classifying chronic lower respiratory disease events in epidemiologic cohort studies." In: *Annals of the American Thoracic Society* 13.7, pp. 1057–1066.

Pinsky, Paul F et al. (2013). "National lung screening trial: variability in nodule detection rates in chest CT studies." In: *Radiology* 268.3, pp. 865–873.

Puliyakote, Abhilash S Kizhakke et al. (2016). "Morphometric differences between central vs. surface acini in A/J mice using high-resolution micro-computed tomography." In: *Journal of Applied Physiology* 121.1, pp. 115–122.

Qian, Ning (1999). "On the momentum term in gradient descent learning algorithms." In: *Neural networks* 12.1, pp. 145–151.

Rad, Reza Moradi et al. (2018). "Multi-Resolutional Ensemble of Stacked Dilated U-Net for Inner Cell Mass Segmentation in Human Embryonic Images." In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 3518–3522.

Raghu, Ganesh et al. (2011). "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management." In: *American Journal of Respiratory and Critical Care Medicine* 183.6, pp. 788–824.

Ramachandram, Dhanesh and Graham W Taylor (2017). "Deep multimodal learning: A survey on recent advances and trends." In: *IEEE Signal Processing Magazine* 34.6, pp. 96–108.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.

Rosvall, Martin and Carl T Bergstrom (2008). "Maps of random walks on complex networks reveal community structure." In: *Proceedings of the National Academy of Sciences* 105.4, pp. 1118–1123.

Roth, Volker et al. (2002). "A resampling approach to cluster validation." In: *Compstat*, pp. 123–128.

Rubin, Geoffrey D et al. (2005). "Pulmonary nodules on multi–detector row CT scans: performance comparison of radiologists and computer-aided detection." In: *Radiology* 234.1, pp. 274–283.

Schilham, Arnold MR et al. (2006). "Local noise weighted filtering for emphysema scoring of low-dose CT images." In: *IEEE Transactions on Medical Imaging* 25.4, pp. 451–463. ISSN: 0278-0062.

Setio, Arnaud Arindra Adiyoso et al. (2016). "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge." In: *arXiv preprint arXiv:1612.08012*.

Shapiro, Steven D (2000). "Evolving concepts in the pathogenesis of chronic obstructive pulmonary disease." In: *Clinics in Chest Medicine* 21.4, pp. 621–632. ISSN: 0272-5231.

Shelhamer, Evan, Jonathon Long, and Trevor Darrell (2016). "Fully Convolutional Networks for Semantic Segmentation." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. ISSN: 0162-8828.

Shen, Wei et al. (2017). "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification." In: *Pattern Recognition* 61, pp. 663–673.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal (2016). "Cancer statistics, 2016." In: *CA: A Cancer Journal for Clinicians* 66.1, pp. 7–30.

Sieren, Jered P et al. (2016). "SPIROMICS Protocol for Multicenter Quantitative CT to Phenotype the Lungs." In: *American Journal of Respiratory and Critical Care Medicine* ja. ISSN: 1073-449X.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556*.

Sluimer, Ingrid et al. (2006). "Computer analysis of computed tomography scans of the lung: a survey." In: *IEEE Transactions on Medical Imaging* 25.4, pp. 385–405. ISSN: 0278-0062.

Smith, Benjamin M et al. (2014). "Pulmonary emphysema subtypes on computed tomography: the MESA COPD study." In: *The American Journal of Medicine* 127.1, 94.e7–23. ISSN: 0002-9343.

Smith, Stephen M et al. (2004). "Advances in functional and structural MR image analysis and implementation as FSL." In: *Neuroimage* 23, S208–S219. ISSN: 1053-8119.

Sofer, Tamar et al. (2018). "A Fully-Adjusted Two-Stage Procedure for Rank Normalization in Genetic Association Studies." In: *bioRxiv*, p. 344770.

Sørensen, Lauge, Saher B Shaker, and Marleen De Bruijne (2010). "Quantitative analysis of pulmonary emphysema using local binary patterns." In: *IEEE Transactions on Medical Imaging* 29.2, pp. 559–569. ISSN: 0278-0062.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *The Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Swensen, Stephen J et al. (1997). "The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules." In: *Archives of Internal Medicine* 157.8, pp. 849–855.

Tajbakhsh, Nima et al. (2016). "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE transactions on medical imaging* 35.5, pp. 1299–1312.

Tamimi, Asad, Dzelal Serdarevic, and Nicola A Hanania (2012). "The effects of cigarette smoke on airway inflammation in asthma and COPD: therapeutic implications." In: *Respiratory medicine* 106.3, pp. 319–328.

Thomashow, Michael A et al. (2013). "Endothelial microparticles in mild chronic obstructive pulmonary disease and emphysema. The Multi-Ethnic Study of Atherosclerosis Chronic Obstructive Pulmonary Disease study." In: *American Journal of Respiratory and Critical Care Medicine* 188.1, pp. 60–68.

Thong, William et al. (2018). "Convolutional networks for kidney segmentation in contrast-enhanced CT scans." In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.3, pp. 277–282.

Thurlbeck, WM (1963). "A clinico-pathological study of emphysema in an American hospital." In: *Thorax* 18.1, p. 59.

Vestbo, Jørgen et al. (2013). "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary." In: *American Journal of Respiratory and Critical Care Medicine* 187.4, pp. 347–365.

Vogelmeier, Claus F et al. (2017). "Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary." In: *American journal of respiratory and critical care medicine* 195.5, pp. 557–582.

Warfield, Simon K, Kelly H Zou, and William M Wells (2004). "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." In: *IEEE Transactions on Medical Imaging* 23.7, pp. 903–921.

West, JB (1963). "Distribution of gas and blood in the normal lungs." In: *British Medical Bulletin* 19.1, pp. 53–58.

Woodruff, Prescott G et al. (2016). "Clinical significance of symptoms in smokers with preserved pulmonary function." In: *New England Journal of Medicine* 374.19, pp. 1811–1821.

Xie, Yutong et al. (2016). "Lung nodule classification by jointly using visual descriptors and deep features." In: *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*. Springer, pp. 116–125.

Xu, Ye et al. (2006). "MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies." In: *IEEE Transactions on Medical Imaging* 25.4, pp. 464–475.

Yang, Jie et al. (2016a). "Emphysema Quantification on Cardiac CT Scans Using Hidden Markov Measure Field Model: The MESA Lung Study." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 624–631.

Yang, Jie et al. (2016b). "Explaining Radiological Emphysema Subtypes with Unsupervised Texture Prototypes: MESA COPD Study." In: *International MICCAI Workshop on Medical Computer Vision*. Springer.

Yang, Jie et al. (2017). "Unsupervised Discovery of Spatially-Informed Lung Texture Patterns for Pulmonary Emphysema: The MESA COPD Study." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 116–124.

Yu, Jiahui et al. (2018). "Generative Image Inpainting with Contextual Attention." In: *arXiv preprint arXiv:1801.07892*.

Zhou, Bolei et al. (2016). "Learning Deep Features for Discriminative Localization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Appendix A: Data Reduction From Ten sLTPs to Six QES in SPIROMICS and MESA Lung Study

The following figures present the visualizations of ten spatially-informed lung texture patterns (sLTPs) discovered in the SPIROMICS dataset, and the data reduction process from ten sLTPs to six quantitative emphysema subtypes (QES).

The data reduction is a collaborative work with Dr. Yifei Sun (Fig. A.2 (a)) from Department of Biostatistics at Columbia University, and Dr. Elsa D. Angelini (Fig. A.2 (b)) in our Heffner Biomedical Imaging Lab.

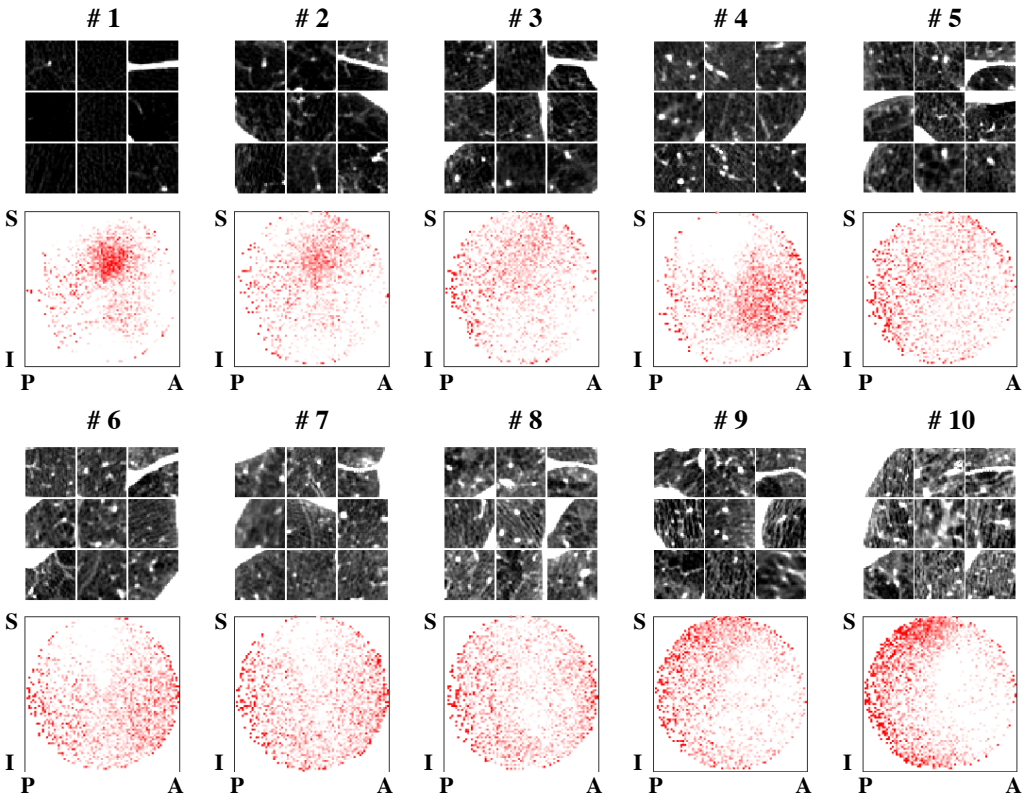

Figure A.1: Qualitative illustrations of the spatially-informed lung texture patterns (sLTPs, #1-10 ordered by mean Hounsfield Units) discovered in SPIROMICS. For each sLTP: (top) texture appearance on CT scans visualized on axial cuts from 9 random ROIs; (bottom) spatial density of labeled ROIs (red dots) on SPIROMICS (showing only spatial density larger than average); legend: S = superior; I = inferior; P = posterior; A = anterior positions.

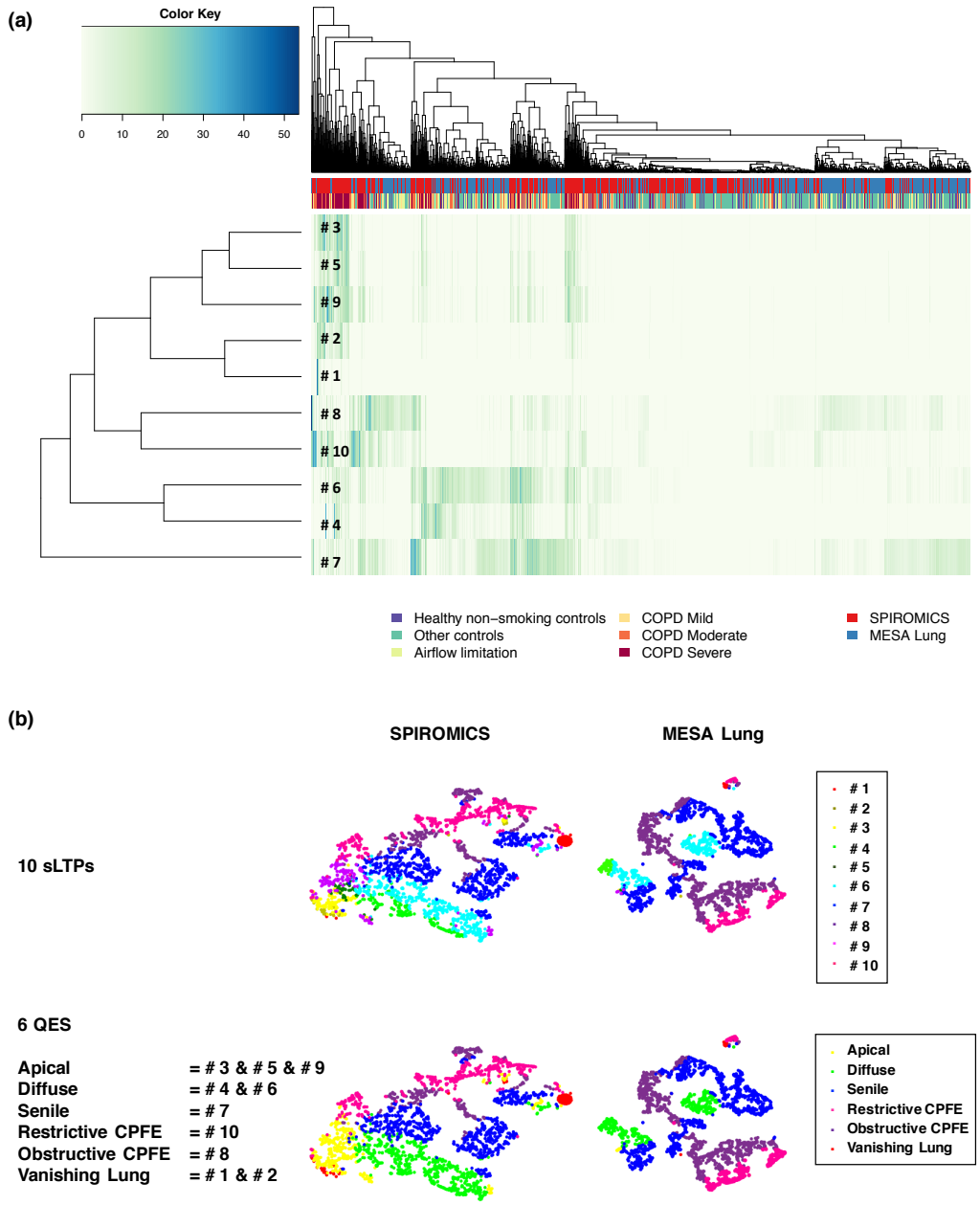

Figure A.2: Clustering of spatially-informed lung texture patterns (sLTPs, #1-10 ordered by mean Hounsfield Units) and quantitative emphysema subtypes (QES): (a) Heatmap and hierarchical clustering of all sLTP histograms in SPIROMICS and MESA Lung Study; (b) t-SNE two-dimensional projection of all sLTP histograms in SPIROMICS and MESA Lung Study, color-coded by the dominant sLTP or QES per CT scan.

Appendix B: Visualization of CNN Filters Learned for Weakly-Supervised Lung Module Detection

The following figures present the visualizations of CNN filters in CNN models (VGG-16, ResNet-50, DenseNet-121) that are trained for the weakly-supervised lung nodule detection in Chapter 6.

In the VGG-16, four representative convolutional layers from the shallow layers to the deep-most layer are selected, including Conv2_2, Conv3_3, Conv4_3, and Conv5_3 (Simonyan and Zisserman 2014). In each layer, we select five random filters, which are visualized in Fig. B.1.

In the ResNet-50, the last convolutional layers in the four bottleneck blocks (He et al. 2016) are selected, including res2c_branch2c, res3d_branch2c, res4f_branch2c, and res5c_branch2c. In each layer, we select five random filters, as visualized in Fig. B.2.

In DenseNet-121, the last convolutional layers in the four dense blocks (Huang et al. 2017) are selected, including conv2_block6_2, conv3_block12_2, conv4_block24_2, and conv5_block16_2. In each layer, we select five random filters, as visualized in Fig. B.3.

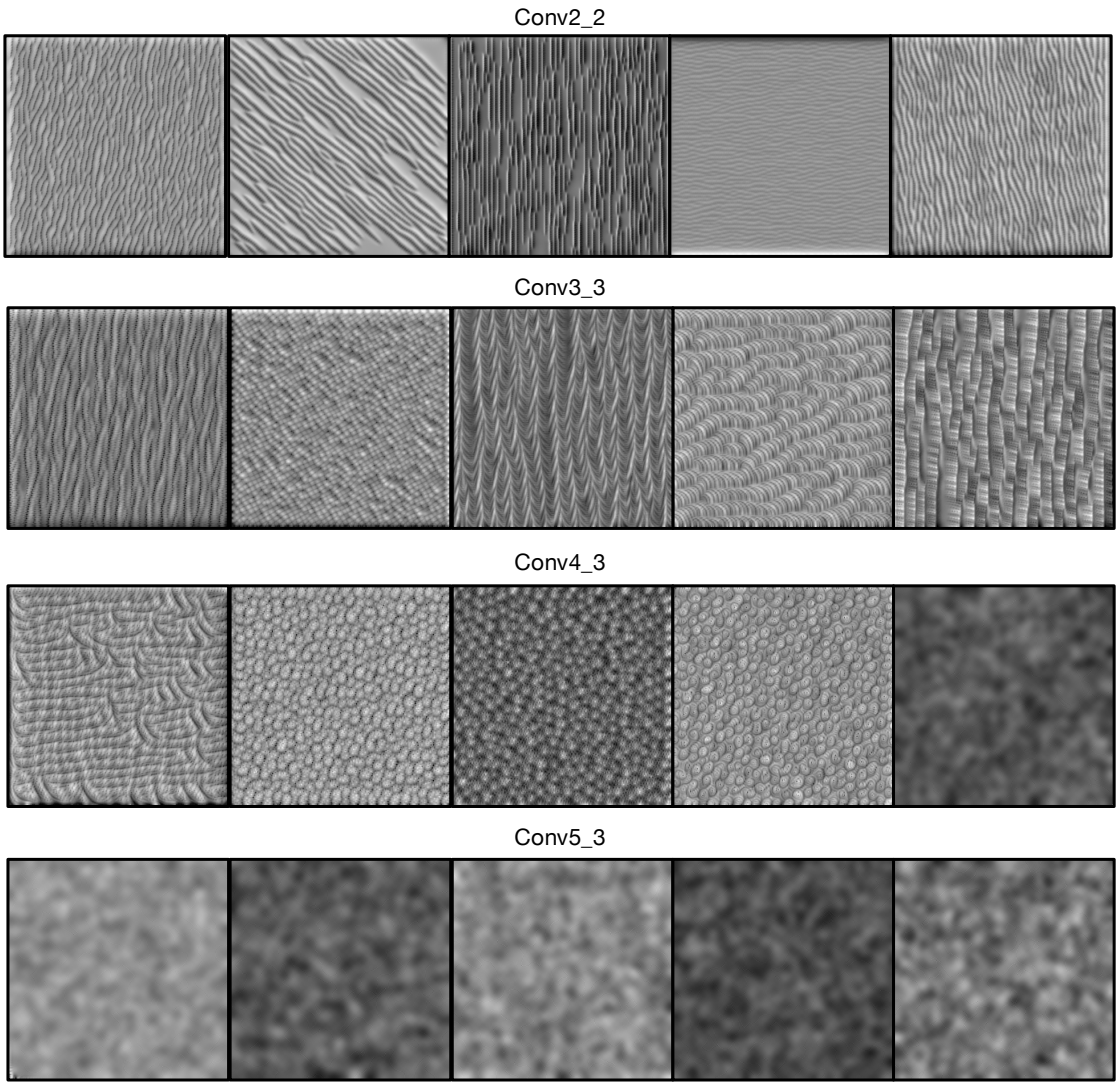

Figure B.1: Visualization of CNN filters in four convolutional layers (five random filters are visualized per layer) in the VGG-16 model trained for weakly-supervised lung nodule detection.

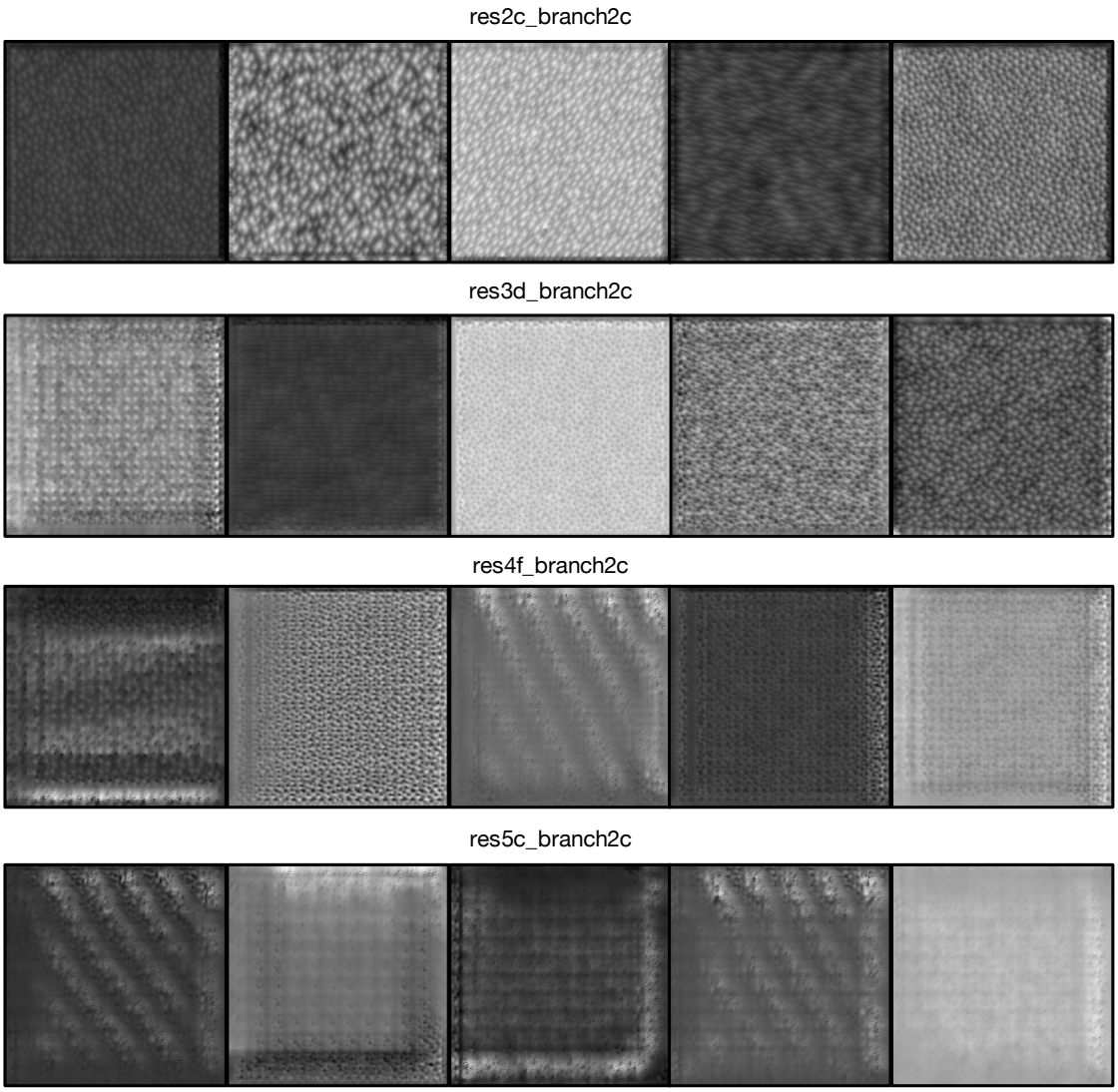

Figure B.2: Visualization of CNN filters in four convolutional layers (five random filters are visualized per layer) in the ResNet-50 model trained for weakly-supervised lung nodule detection.

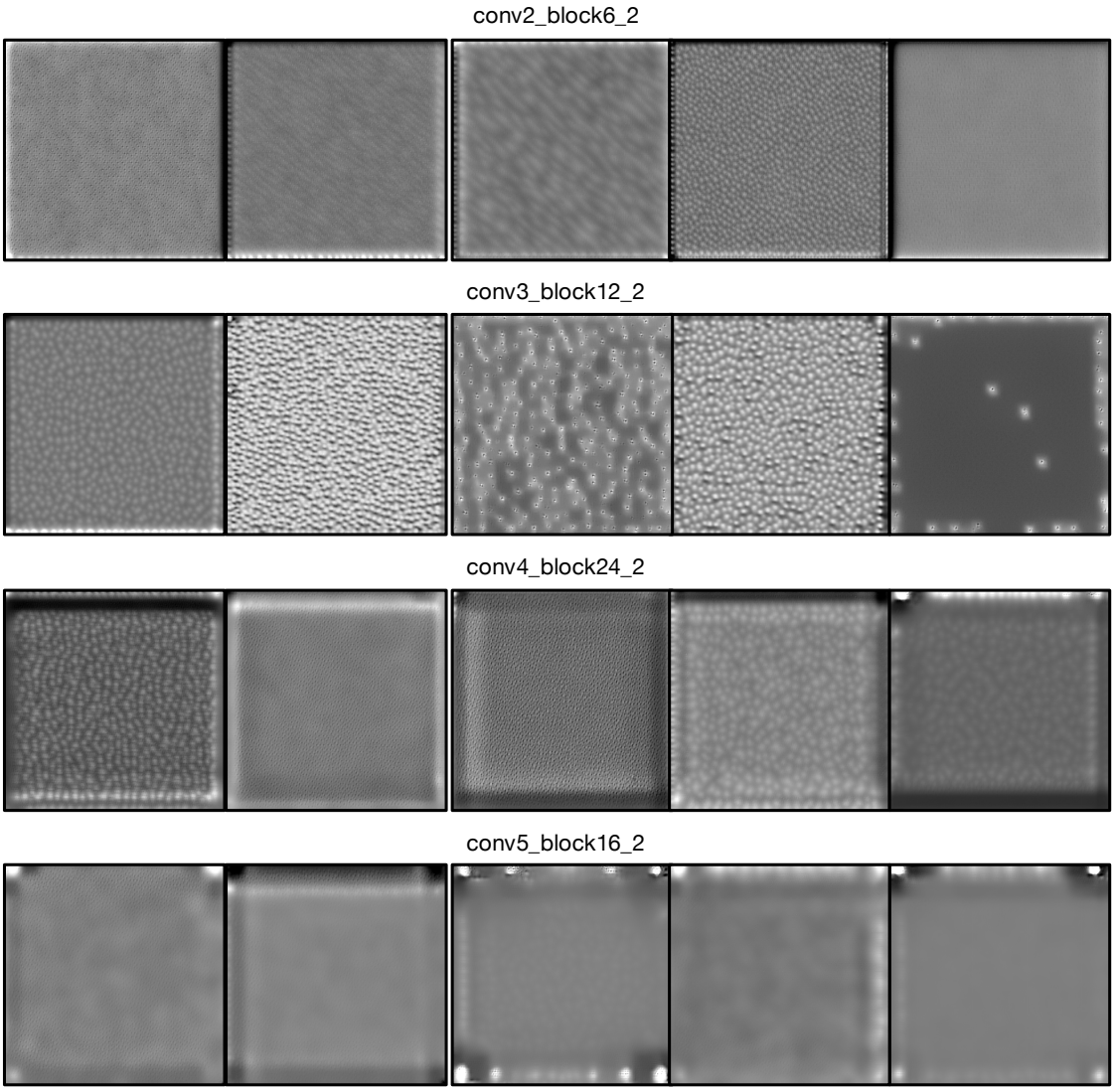

Figure B.3: Visualization of CNN filters in four convolutional layers (five random filters are visualized per layer) in the DenseNet-121 model trained for weakly-supervised lung nodule detection.

Appendix C: Weakly-Supervised Nodule Detection in Two-Nodule Slices and 3D Regions

In Chapter 6, we presented weakly-supervised lung nodule detection with slice-level labels in LIDC-IDRI dataset, using 2D CNN architectures. Evaluations focused on 2D slices with one nodule. In this appendix, we first discuss about nodule detection on rare cases of slices with two nodules (accounting for ∼1% nodule slices in the LIDC-IDRI dataset). Then we discuss about potential extension to detect nodules in 3D regions with some preliminary results.

Detection Results on Two-Nodule Slices

For slices with two nodules, our framework can detect nodules by segmenting the top two activation blobs in the NAM. We tested the detection on a total of 108 slices with two nodules, using the VGG-16 based model. The 2-GAP system achieves the best detection performance, where both nodules are correctly detected in 50 slices, and one of the two nodules is correctly detected in another 42 slices. With adequate training data, our framework can be extended to multi-class classification to automatically determine the number of nodules to segment in the slice.

Potential of Nodule Detection in 3D Regions

Extending the proposed framework to weak labels in 3D regions covered by continuous axial slices will enable the usage of contextual information in 3D neighborhoods. However, the definition of 3D regions with and without nodule can be tricky, when there is only part of the nodule presenting at the border of the field of view. In this work, we only explore the potential of classifying 3D regions (consist of 8 continuous axial slices) with either no nodule, or nodules covering at least the middle two slices. This results in 1,522 regions with nodule, and an equal number of regions without nodule from the LIDC-IDRI dataset, which are split into training, validation and test sets with a ratio of 4:1:1.

We tested the 3D version of ResNet-50 and DenseNet-121 for that purpose. When training from scratch, our training cannot converge due to the limited sample size. Then we tested transfer learning with same models that are pre-trained on a very large video dataset Kinetics (Carreira and Zisserman 2017). Using the ResNet-50, we can achieve a validation classification accuracy 0.921, and the test accuracy is 0.918. Using DenseNet-121, the validation accuracy is 0.924, and the test accuracy is 0.895.

Appendix D: Lung Nodule Malignancy Classification with Class-Aware Adversarial Nodule Synthesis in CT Images[1]

Introduction

Pulmonary nodules are crucial indicators of early-stage lung cancer. The majority of nodules detected in lung CT screening are eventually benign. A data-driven model that can accurately predict nodule malignancy risk from CT images may prevent unnecessary imaging or invasive follow-up procedures on benign nodules, thus increase the effectiveness of lung cancer screening programs (MacMahon et al. 2005).

Deep convolutional neural network (CNN) based approaches have demonstrated superior performance for image related tasks such as image classification and object detection, and now have been widely studied in CAD systems for automated detection of pulmonary nodules using lung CT. In Chapter 6, we presented a weakly-supervised approach based on CNNs to detect lung nodules with less annotations, which we believe is a more sustainable solution for CAD systems compared to fully-supervised methods.

However, to utilize CNNs for the classification of benign versus malignant nodules for lung cancer prediction, existing methods (Shen et al. 2017; Xie et al. 2016) are still

[1]This work was partially done while Jie Yang was interning at Siemens Corporate Research.

constrained by the quantity and the diversity of the training data available. Given the large imbalance of benign versus malignant nodules in real-world data, the malignant cases are generally largely underrepresented. In the the large National Lung Screening Trial (NLST) (National Lung Screening Trial Research Team 2011), a total of 96.4% of lung nodules examined did not result in a lung cancer diagnosis. Therefore, it is important to improve the efficiency of the medical machine learning systems by constructing a training dataset with better represented classes or improving the learning approaches.

Collecting more malignant cases from the general population is challenging, which can be very expensive and time consuming. In this chapter, to deal with the limited data availability of malignant nodules for lung cancer prediction, we propose a novel image classification system with class-aware data augmentation based on generative adversarial network (GAN) (Goodfellow et al. 2014).

Basic data augmentation techniques, such as random cropping, shifting, scaling, flipping and rotations, can be used to introduce a certain level of diversity during training stage, but cannot account for the diversity of nodule morphology and locations.

Some recent studies proposed to use GANs to synthesize lesions in medical image patches to augment the training data (Jin et al. 2018; Korkinof et al. 2018). Such methods train a generator network and a discriminator network to in-paint a missing (masked) area with the objects of interests (i.e. targeted lesions). The generator network is trained with a reconstruction loss between the synthetic patch and the real patch as well as an adversarial loss produced by the discriminator network. They concluded that the synthetic patches could improve the performance of the supervised learning tasks. However, such networks were designed to generate objects conditioning only on surrounding context

and random noises, lacking the capability of generating objects with manipulable properties which we believe to be important for many machine learning applications in medical imaging, such as balancing the classification datasets.

To synthesis lung nodule that are *class-aware*, we propose an adversarial learning framework conditioning on the target categories (benign vs. malignant). We formulate our approach as an image in-painting problem, so that the nodules can be synthesized at random locations within the lungs. Evaluating on the public The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset (Armato III et al. 2011), we show that with the proposed framework, we can synthesize nodules which have high fidelity and can improve the assessment of the nodule malignancy risk.

Method

To estimate the malignancy risk of pulmonary nodules in lung CT images, we first introduce a class-aware nodule synthesis framework to deal with the imbalance of benign versus malignant nodules in training data. Then we train a 3D deep CNN model for lung nodule malignancy classification.

The proposed nodule synthesis framework is formulated as an in-painting problem which fills a missing (masked) area in a 3D lung CT image patch with a lung nodule within the specified category. The framework contains the following three major components:

1. A coarse generator network and a refinement generator network to perform a two-step in-painting incorporating contextual information;
2. A local discriminator network and a global discriminator network to enforce the

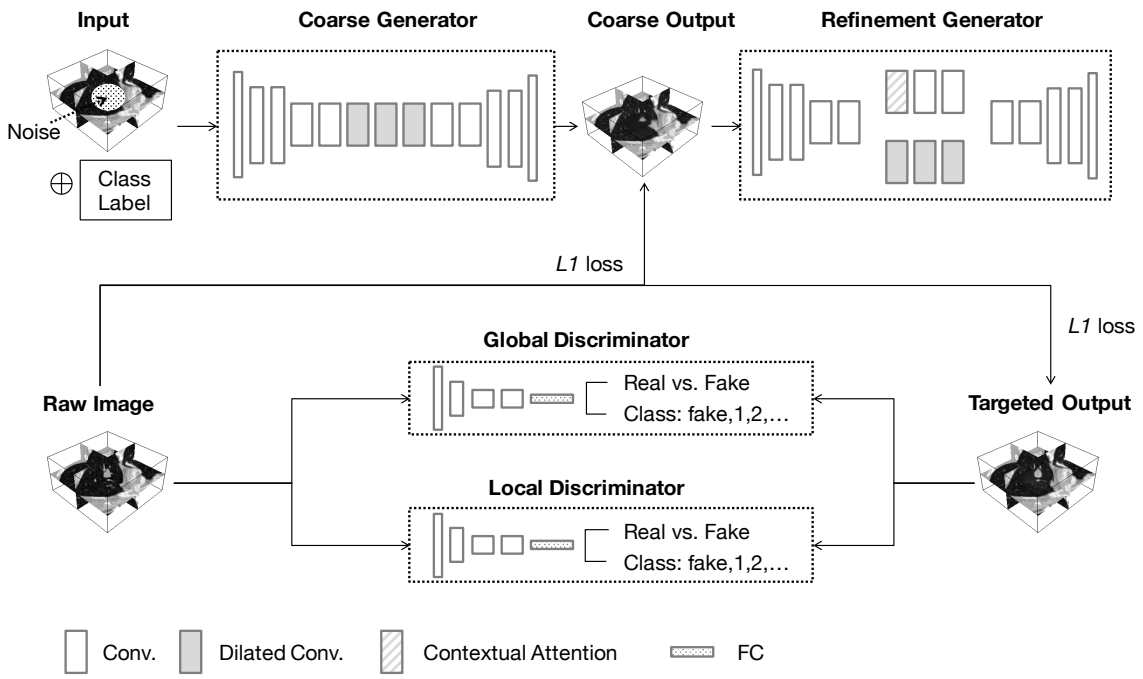

Figure D.1: Illustration of the proposed class-aware adversarial nodule synthesis framework. The noise masked 3D CT image patch is fed into two generators, a coarse generator and a refinement generator, sequentially. The same ground truth patch is used for computing the reconstruction $L1$ loss for both the coarse generator and the refinement generator. The refinement generator is trained with both $L1$ and the adversarial losses provided by both a local and a global discriminators. Each discriminator is responsible for predicting if a patch is fake as well as the nodule malignancy label.

local quality and the global consistency of the generated nodules;

3. Auxiliary domain classifiers in the discriminators to constrain the generated nodules with the specified category conditions.

The coarse generator is optimized with $L1$ reconstruction loss, and the refinement generator is optimized with both $L1$ reconstruction loss and adversarial loss. The nodule synthesis framework is illustrated in Fig. D.1.

After synthesizing sufficient malignant nodules for data augmentation, we train a 3D CNN to classify benign versus malignant patches. Transfer learning is used to enable the usage of very deep neural networks for better characterization.

We now detail these components in the below sections.

Coarse-to-Fine Generators

3D images patches are extracted from the lung CT volumes centering on annotated nodules. The nodule in each patch is replaced with a 3D spherical (to approximate the round shape of most nodules) noise mask generated according to the size determined the annotated nodule diameter.

The masked patch and a class-label map (padded to the same size as the image patch) are firstly fed into a 3D hour-glass CNN G_1, named coarse generator, to reconstruct the masked region with a coarsely synthesized nodule.

The output of G_1 is fed into another network G_2, named refinement generator, with a similar architecture as G_1 to refine the details. G_1 and G_2 together form the stacked image in-painting generators G.

Both G_1 and $G2$ are optimized with the reconstruction loss $L_{recon}^{(1,2)}$ between the reconstructed patch and the real nodule patch:

$$L_{recon}^{(1,2)} = L_{masked} + \lambda_1 L_{global} \tag{1}$$

where L_{mask} and L_{global} are respectively the normalized $L1$ loss across the masked area and the entire patch. While only mask areas are kept and are embedded into the original image patch context for the final output, the L_{global} is an important indicator here to stabilize the reconstruction. By optimizing $L_{recon}^{(1,2)}$, the stacked generators are trained to reconstruct the nodules in the original patch based on the lung tissue image context,

random noise mask, and also the class-label map as enforced in Section 8.2.

For the purpose of data augmentation, variability of the output is required for our nodule synthesis framework. Meanwhile, $L1$ loss is a classic loss function in GAN-based image synthesis to ensure that the synthesized output manifests information in the original input, and thus stabilize the training process (Korkinof et al. 2018).

Aside from $L^{(2)}_{recon}$, the output of the refinement generator G_2 is also optimized by an adversarial loss provided by two discriminator networks described in Section 8.2 and Section 8.2.

Contextual Attention

Features from surrounding regions can be helpful for in-painting the boundary of the nodules in the masked area. In a recent study (Yu et al. 2018), a contextual attention model is proposed to borrow the textures adaptively from the background patches to generate the foreground missing patches.

We use the contextual attention model to match between the foreground (missing regions) and background textures by measuring the normalized cosine similarity of their features:

$$s_{x,y,x',y'} = \left\langle \frac{f_{x,y}}{||f_{x,y}||}, \frac{b_{x',y'}}{||b_{x',y'}||} \right\rangle \tag{2}$$

where $s_{x,y,x',y'}$ represents similarity of patch centered in foreground (x, y) and background (x', y'), and f, b denote the features respectively.

Then we weigh the similarity to get attention score for each surrounding pixel, and finally, reconstruct foreground patches with background ones by performing deconvolution on attention score maps.

The contextual attention operation is embedded in the second generator in the refinement stage, as shown in Fig. D.1, and is differentiable and fully-convolutional.

Local and Global Discriminators

Two discriminator networks D_{local} and D_{global} are used to train $G1$ and $G2$ in an adversarial fashion. D_{local} is applied to the masked area only to refine the local nodule appearance; while D_{global} is applied to the entire patch for global consistency of the in-painting. We denote both discriminators as D_* for brevity.

We use the conditional Wasserstein GAN objective (Arjovsky, Chintala, and Bottou 2017) and enforcing the gradient penalty (Gulrajani et al. 2017) to train the discriminators D_* and the stacked generator G as

$$L_{adv} = E_x[D_*(x)] - E_{z,c}[D_*(G(z,c))] - \lambda_{gp} E_{\hat{x}}[(\|\nabla D_*(\hat{x})\|_2 - 1)^2] \tag{3}$$

where x is sampled from the real patch distribution, z is the image patch masked with random noise. $G(z,c)$ is the final output of the stacked generator networks, $\hat{x}$ is sampled uniformly between a pair of real and generated patches during training, according to the optimization with gradient penalty. And c is the class-label map.

The Wasserstein GAN, or commonly termed WGAN, improves the stability of model convergence and avoids model collapse compared to the original GAN model by introducing a new cost function using Wasserstein distance that has a smoother gradient everywhere. WGAN with gradient penalty (WGAN-GP) has now been widely used in natural image synthesis and image translation tasks (Choi et al. 2017).

Class-Aware Synthesis

To achieve "class-aware" nodule synthesis for data augmentation in training nodule malignancy classification, we further add an auxiliary domain classifier D_{cls} on top of each discriminator network to ensure G to generate nodules in the targeted class c. D_{cls} tries to classify the image patch x into the class label c (0 = fake, 1 = benign and 2 = malignant). The label 0 is used to prevent the generator from in-painting near-identical nodules that are easy to classify but less diversified. D_{cls} is optimized with the class-aware loss L_{cls} as:

$$L_{cls} = E_{x,c}[-logD_{cls}(c|x)] \tag{4}$$

where $D_{cls}(c|x)$ represents a probability distribution over class labels c for an image patch x. When optimizing D_{cls}, x represents either a real image patch or a generated image patch, and c represents the real class label; While in the adversarial training to optimize G, c represents the target class label for a generated image patch x, so that G minimizes this objective to generate images that can be classified by D_{cls} as the target class.

The objective function for our whole class-aware nodule synthesis learning can then be summarized as:

$$\begin{aligned} L_{D_*} &= L_{adv} + \lambda_{cls}^{(D_*)} L_{cls} \\ L_G &= -L_{adv} + \lambda_{cls}^{(G)} L_{cls} + \lambda_{recon} L_{recon} \end{aligned} \tag{5}$$

3D Deep CNNs with Transfer Learning

To classify the CT nodule patches into benign and malignant classes, we propose to use deep CNNs pre-trained for video classification tasks, with transfer learning.

Fine-tuning of deep CNN weights pre-trained on very large-scale nature image datasets such as the ImageNet (Krizhevsky, Sutskever, and Hinton 2012) was demonstrated to have superior performance compared to training from scratch in medical image domain (Tajbakhsh et al. 2016).

Video data resembles 3D images volumes where the temporal dimension is analogous to the third spatial dimension manifesting the consistency of nearby frames/slices. It is much easier to scale up the collection and annotation of natural video datasets than medical image datasets.

It was very recently demonstrated that, for human action recognition tasks in videos, spatiotemporal three dimensional convolutional kernels (3D CNNs) are more effective than CNNs with two-dimensional (2D) kernels, when large-scale frame-wise annotated datasets such as the Kinetics dataset are readily available (Carreira and Zisserman 2017). And state-of-the-art CNN architectures pre-trained on Kinetics were demonstrated to have superior performance when generalizing to smaller video datasets (Hara, Kataoka, and Satoh 2018).

Therefore, using video pre-trained networks enables the usage of very deep 3D CNNs for our task, and is helpful for stabilizing the network training and preventing overfitting.

Experimental Results

Data and Experimental Setup

We evaluate our proposed methods on the LIDC-IDRI dataset (Armato III et al. 2011) consisting of diagnostic and lung cancer screening thoracic computed tomography (CT)

Subset	Benign	Malignant Real	Malignant Synthesized	Total
Train	1,004	191	-	1,195
Train+Syn	1,004	191	463	1,658
Validation	251	47	-	298
Test	251	47	-	298

Table D.1: The split of training, validation and test data in LIDC-IDRI for the proposed class-aware nodule synthesis.

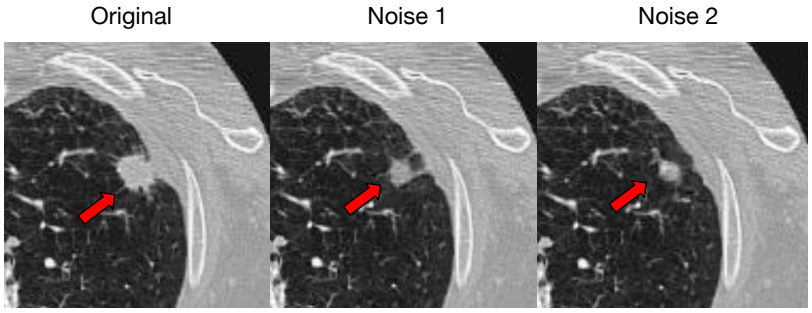

Figure D.2: Examples of the nodule synthesis results with the same input patch and different the initial noise masks. Left) Original image patch with a malignant nodule; Middle-Right) In-painted nodule patches that are generated with two sets of random noise seeds.

scans with marked-up annotated lesions.

The LIDC-IDRI dataset consists of 1,010 patients and 1,018 chest CT imaging studies in total. The nodules were annotated by four radiologists as benign and malignant by giving a score ranging from 0 (confident benign) to 5 (confident malignant). We define the nodules with the majority score ≥ 4 to be malignant and the rest to be benign.

Visual Evaluation of the Nodule Synthesis Results

In our experiments, we extract the nodule patches from the LIDC-IDRI dataset with the resolution $1 \times 1 \times 2$mm and the size $64 \times 64 \times 32$. The patches are randomly split into the training set, validation set and testing set according to the patients as shown in Table D.1.

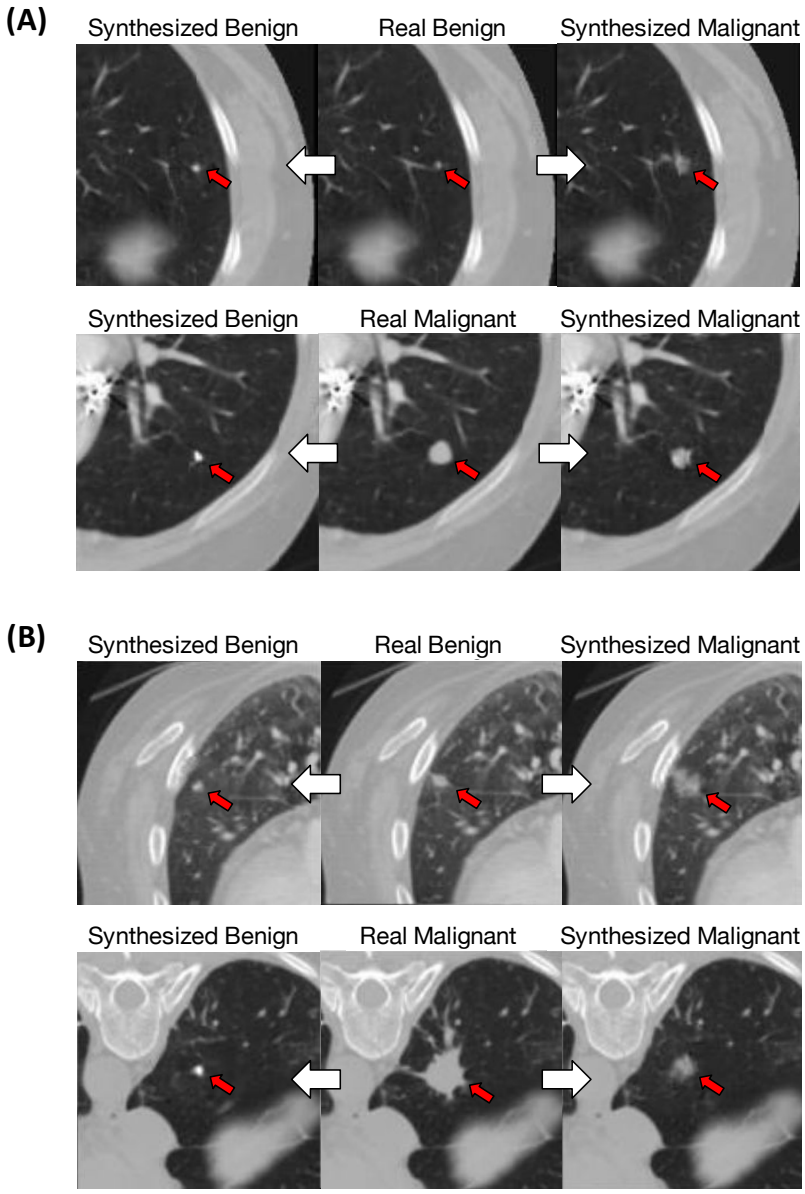

Figure D.3: Examples of the nodule synthesis results by altering the nodule malignancy labeling condition. (A) Nodule synthesis results on image patches originally with benign nodules; (B) Nodule synthesis results on image patches originally with malignant nodules (Left = in-painted image patch to synthesize a benign nodule; Middle = original image patch; right = in-painted image patch to synthesize a malignant nodule).

We train the proposed nodule synthesis framework on the training patches only. We first show in Fig.D.2 that the network could generate nodules with different morphology and textures using different noise masks, which is a desirable property for our purpose of

data augmentation. The two noise masks used in Fig.D.2 are Gaussian noises with zero mean and standard deviation = 0.2, that are generated with two random seeds.

Then in Fig.D.3, we demonstrate the nodule synthesis results by conditioning on different target class labels (benign or malignant). We can observe that, given the same background patch, the framework is capable of generating nodules with different specified malignancy labels.

Quantitative Evaluation of Nodule Malignancy Prediction

The trained generator is used for synthesizing 463 patches containing malignant patches since malignant nodules are relatively rare in the original LIDC-IDRI dataset. The synthetic nodule patches are combined with the original training patches to train the 3D classification CNNs in our proposed setting.

To evaluate the effectiveness of the synthetic patches on estimating the lung nodules malignancy, we trained three 3D ResNet-based CNN architectures with different network depth, respectively ResNet-50, ResNet-101, and ResNet-152 (He et al. 2016). All the networks were initialized with the weights pre-trained on the Kinetic video dataset (Carreira and Zisserman 2017; Hara, Kataoka, and Satoh 2018). The cross-entropy loss was used for training the CNN classifiers.

We also evaluated the differences between the unweighted (Raw) and weighted cross entropy loss (Raw + Weighted Loss). Traditional data augmentation methods including random cropping and scaling were used for training all the networks. The testing accuracy (ACC), sensitivity (SEN), specificity and the area under the ROC curve (AUC) presented in Table.D.2 were selected based on the highest AUCs on the validation set.

Network	ACC	SEN	SPE	AUC
Raw Training				
ResNet-50	0.859	0.660	0.896	0.862
ResNet-101	0.861	0.653	0.901	0.847
ResNet-152	0.873	0.596	0.924	0.860
Raw Training + Weighted Loss				
ResNet-50	0.842	0.681	0.873	0.836
ResNet-101	0.826	**0.723**	0.847	0.810
ResNet-152	0.829	0.702	0.853	0.818
Raw Training + Synthesis				
ResNet-50	0.883	0.702	0.916	0.867
ResNet-101	0.893	0.702	0.928	0.881
ResNet-152	**0.903**	0.660	**0.948**	**0.883**

Table D.2: The nodule malignancy classification results with different network architectures and different data balancing strategies.

From Table.D.2, our method with the synthetic patches (Raw + Synthesis), the 3D ResNet152 achieved the highest accuracy, specificity and AUC score across all the trials. The overall mean AUC scores also indicate that the synthetic nodule patches are helpful for improving the nodule malignancy classification performance. Using weighted loss, the sensitivity for predicting malignant class can increase, but with a large sacrifice in specificity and AUC.

Discussion and Conclusion

There is a large data imbalance of benign nodules versus malignant nodules presenting in general population as shown in the large-scale study (National Lung Screening Trial Research Team 2011), or even in populations that are high risk of lung cancer as shown in

the public dataset (Armato III et al. 2011). In this appendix, to deal with the issue of limited data availability of malignant nodules for lung cancer prediction, we propose a novel image classification framework with an adversarial in-painting based nodule synthesis to generate training samples in targeted category for data augmentation.

The qualitative results show that the proposed framework is capable of generating lung nodules in the specified malignancy class, with high visual fidelity. By evaluating on the nodule patches obtained from CT scans in the LIDC-IDRI study, we show that the generated nodules can be helpful for improving the classification performance on an imbalanced lung nodule dataset. The proposed work demonstrated the potential of reliable nodule synthesis, and can be applied to other tasks such as nodule detection, and lung lobe segmentation with nodule presence.