Title
Characterizing Pulmonary Nodules using Machine and Deep Learning Methods to Improve Lung Cancer Diagnosis

Author
Shen, Shiwen

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Characterizing Pulmonary Nodules using Machine and Deep Learning Methods to Improve Lung Cancer Diagnosis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Shiwen Shen

2018

ABSTRACT OF THE DISSERTATION

Characterizing Pulmonary Nodules using Machine and Deep Learning Methods to Improve Lung Cancer Diagnosis

by

Shiwen Shen
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2018
Professor Alex Anh-Tuan Bui, Co-Chair
Professor William Hsu, Co-Chair

Low-dose computed tomography (CT) screening has been widely used to detect and diagnose early stage lung cancer. Clinical trials have shown that low-dose CT reduced lung cancer mortality by 20% relative to plain chest radiography; however, challenges exist in current low-dose CT screening programs including high over-diagnosis rates, high cost and increased radiation exposure. This dissertation attempts to overcome these challenges by developing machine and deep learning models for automated lung cancer diagnosis and disease progression estimation. A novel lung segmentation approach was first developed using a bidirectional chain code method and machine learning framework. This method is designed to include the lung nodules attached to lung wall while minimizing over-segmentation error. Second, a hybrid ensemble convolutional neural network has been developed to classify lung nodule vs. non-nodule objects. The ensemble model combines the VGG, residual and densely connected module designs to improve the model classification robustness for external datasets collected with different acquisition parameters. Third, a hierarchical semantic convolutional neural network (HSCNN) has been described to classify lung nodule malignancy. Semantic characteristic features, predicted in parallel with the malignancy for each nodule, enable the interpretation of the model and improvement of malignancy prediction. Finally, a Bayesian framework combined with a continuous-time Markov model was developed to estimate the multi-state disease progression of lung cancer. The resulting model estimates individual lung

cancer state transition information, providing the basis for personalized screening recommendations. Extensive experiments and results have proved the effectiveness of these methods paving the way to optimize and improve the effectiveness of existing low-dose CT screening programs.

The dissertation of Shiwen Shen is approved.

Yingnian Wu

Ricky Kiyotaka Taira

Denise R Aberle

William Hsu, Committee Co-Chair

Alex Anh-Tuan Bui, Committee Co-Chair

University of California, Los Angeles

2018

Dedicated to my beloved family and friends!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my utmost gratitude and respect for my advisor, Dr. William Hsu, for his guidance and support throughout my Ph.D career. His dedication and expertise were the unyielding sources of inspiration and encouragement. I am thankful that he offered me the opportunity to first join the research project of the Medical Imaging Informatics (MII) group and the MII Ph.D program afterwards. I would also like to offer my most sincere thanks and respect to the co-chair of my committee, Dr. Alex Bui, for his insights and dedication to help me grow my knowledge and passion on medical informatics. He taught me to be meticulous in everything that I do and how to improve writing in all aspects. I would also like to thank all the other members of my thesis committee, Dr. Denise Aberle, Dr. Ricky Taira and Dr. Yingnian Wu, for offering guidance and feedback for all stages of this dissertation work.

To all current and past members of the MII group and center for domain specific computing (CDSC), thank you for providing me an intellectually stimulating environment for my study and research. I am extremely thankful to professor Jason Cong, Frank Meng, Corey Arnold, James Sayre, Suzie El-Saden, Craig Morioka, Qing Zhou, Songchun Zhu, Alan Yuille and Luminita Vese, for providing me with mentorship in various research topics and machine learning studies. I am also very thankful like to Isabel Rippy for her help as I made my way throughout my Ph.D study. Thank you to Shawn, Lew, Patrick, Denise, Bing, Weixia, Carlos, Audrey and Lily for your friendship and helping with administrative issues. I would like to thank the past the current MII graduate students, Bill, Anna, Kyle, Jean, Maurine, Simon, Johnny, Nova, Nick, Edgar, Panayiotis, Tianran, Jiayun, Karthik and Daniel for invaluable collaborations, enthusiasm for research and scientific debates resulting in exciting research ideas and reminding me for a lunch break.

I would like to thank UCLA for providing me the opportunity to pursue my graduate study and offering me the Graduate Division Fellowship and the Bioengineering Fellowship to fund my study and research. The research was also supported by the Center for Domain-

Specific Computing (CDSC) funded by the NSF Expedition in Computing Award CCF-0926127. Computing resources were funded by the NIH Data Commons Pilot and a donation of a Titan Xp graphics card by the NVIDIA Corporation.

As this dissertation includes contents of the following articles, I would like to thank, and re-thank, all co-authors for their contributions.

Shen S, Bui AAT, Cong J, Hsu W. An Automated Lung Segmentation Approach using Bidirectional Chain Codes to Improve Nodule Detection Accuracy. Computers in Biology and Medicine. 2015 Feb 1;57:139-49.

Shen S, Han SX, Petousis P, Meng F, Hsu W, Bui AAT. A Continuous Markov Model Approach Using Individual Patient Data to Estimate Mean Sojourn Time of Lung Cancer. American Medical Informatics Association (AMIA) Annual Symposium. 2015; San Francisco, USA.

Shen S, Han SX, Petousis P, Weiss RE, Meng F, Bui AAT, Hsu W. A Bayesian Model for Estimating Multi-state Disease Progression. Computers in Biology and Medicine. 2017 Feb 1;81:111-20.

Shen S, Bui AAT, Hsu W. Robust Lung Nodule Classification using 2.5D Convolutional Neural Network. American Medical Informatics Association (AMIA) Annual Symposium. 2017; Washington, D.C, USA.

Shen S, Han SX, Aberle D, Bui AAT, Hsu W. An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification. In Preparation.

VITA

2013–2018	Graduate Student Researcher, University of California, Los Angeles
2017	Data Scientist Intern (summer), PayPal
2016	Data Scientist Intern (summer), Uber
2009–2012	Graduate Student Researcher, Shanghai Jiao Tong University M.S. Electrical Engineering
2012	Research Intern, Philips Research Aisa - Shanghai
2005–2009	B.S. Electrical Engineering, University of Electronic Science and Technology of China

PUBLICATIONS AND PRESENTATIONS

Shen S, Bui AAT, Cong J, Hsu W. An Automated Lung Segmentation Approach using Bidirectional Chain Codes to Improve Nodule Detection Accuracy. Computers in Biology and Medicine. 2015 Feb 1;57:139-49.

Duggan N, Bae E, **Shen S**, Hsu W, Bui AAT, Jones E, Glavin M, Vese L. A Technique for Lung Nodule Candidate Detection in CT using Global Minimization Methods. In International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition 2015 Jan 13 (pp. 478-491). Springer, Cham.

Shen S, Han SX, Petousis P, Meng F, Hsu W, Bui AAT. A Continuous Markov Model Approach Using Individual Patient Data to Estimate Mean Sojourn Time of Lung Cancer.

American Medical Informatics Association (AMIA) Annual Symposium. 2015; San Francisco, USA.

Shen S, Zhong X, Hsu W, Bui AAT, Wu H, Kuo M, Raman S, Margolis DJA, Sung KH. Quantitative MRI-Driven Deep Learning for Detection of Clinical Significant Prostate Cancer. 24th International Society of Magnetic Resonance in Medicine (ISMRM) Annual Meeting. 2016; Singapore, Singapore.

Shen S, Han SX, Petousis P, Weiss RE, Meng F, Bui AAT, Hsu W. A Bayesian Model for Estimating Multi-state Disease Progression. Computers in Biology and Medicine. 2017 Feb 1;81:111-20.

Shen S, Bui AAT, Hsu W. Robust Lung Nodule Classification using 2.5D Convolutional Neural Network. American Medical Informatics Association (AMIA) Annual Symposium. 2017; Washington, D.C, USA.

Shen S, Han SX, Aberle D, Bui AAT, Hsu W. An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification. expert systems with applications. 2018; in Submission.

Li M, **Shen S**, Chen Z, Gao W, Hsu W, Cong J. Computed Tomography Image Enhancement using 3D Convolutional Neural Network. 21st Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI). 2018; in submission.

CHAPTER 1

Introduction

1.1 Overview

Lung cancer is the leading cause of cancer death among both women and men [TSJ16]. The 5-year survival rate is only 17% for lung cancer [SJ15], but if detected early on, survival increases to 54% [SJ15]. Low-dose computed tomography (CT) is now the *de facto* imaging modality used to screen and identify nascent lung cancers, with the landmark National Lung Screening Trial (NLST) demonstrating a 20% mortality reduction for individuals undergoing low-dose CT (LDCT) relative to plain chest radiography [Tea11]. Compared to conventional chest radiography, CT generates high resolution, volumetric datasets that are able to resolve small and/or low-contrast nodules [LKH12]. However, several challenges exist in the use of LDCT in this setting, hindering accurate detection and effective screening. First, screening programs produce large volumetric datasets that are time-consuming and effort-intensive for radiologists to carefully review [Li07]. Second, less experienced radiologists have highly variable detection rates, particularly in subtle cases, as interpretation heavily relies on past experience [ZTB13]. In point of fact, it is often challenging – even for experts – to accurately differentiate malignant nodules from benign lesions, resulting in a high degree of false positives being the result [AAW03]. As found during the NLST, the positive predictive values for LDCT (i.e., the proportion of positive screens with a subsequent confirmed lung cancer diagnosis), were only 3.8, 2.4 and 5.2% in Screenings 1, 2 and 3, respectively [Pin14]. Lastly, concerns regarding radiation exposure, over-diagnosis, and over-treatment underscore the need for more individually-tailored screening that determines who should be screened and at what frequency [SHP17]. This dissertation focuses on improving lung cancer screening

and diagnosis through novel machine and deep learning models to overcome these challenges, including: 1) automated lung segmentation; 2) computer-aided nodule detection and lung cancer diagnosis in CT images; and 3) personalized periodic screening interval estimation.

1.2 Background and Motivation

Clinical trials show that existing low-dose CT screening programs have three challenges: 1) high over-diagnosis rates; 2) high cost; and 3) increased radiation exposure. The NLST study reported a 96.4% false positive rate for all positive screening results [Tea11]. The findings in NLST documented costs of $52,000 and $81,000 for one additional life-year and quality-adjusted life year (QALY) per person, respectively [BGS14]. In addition, it estimated approximately 1-3 lung cancer deaths are induced by radiation per 10,000 scanned subjects in the trial [Tea11]. Developing computer-aided detection/diagnosis (CAD/CADe) systems [CC12, SJG15, PZL08] and determining individualized, optimal screening intervals are perceived as keys to overcome these issues.

A variety of machine learning (e.g., support vector machines, decision trees) and statistical methods have been employed in CAD/CADe systems to improve the ability to accurately and consistently detect and diagnose lung cancer. This has been an active area of research in medical image analysis [RDF08, BKN05, CC13] over the past two decades. CAD/CADe systems have been explored to assist radiologists in the reading process, potentially increasing the positive predictive value and reduce the false positive rate in lung cancer screening for small nodules, as compared with human reading by thoracic radiologists [HPY17]. CAD/CADe systems have also been shown to make screening more cost-effective [DMM06, SCL16].

The CAD/CADe diagnostic workflow typically consists of four key components, as shown in Figure 1.1a: 1) lung segmentation [SBC15]; 2) nodule candidate generation [DBS15]; 3) nodule classification [FCS17]; and 4) lung cancer diagnosis [HKB14]. The first stage is an important preprocessing step to generate a region of interest (ROI) for subsequent analysis (i.e., the lung field) for most CAD/CADe systems. One commonly-missed type of pulmonary

nodule by the segmentation step is the juxtapleural nodule (as shown in Figure 1.2), which is attached to the wall of the lung. Although many works study lung segmentation, only a few explicitly handle the juxtapleural nodules; and evaluation is lacking. The second component aims to segment a large set of suspicious nodule candidates from these ROIs with high sensitivity, typically using thresholding and morphological operations [MHR10]. In the third stage, lung nodules and non-nodule objects (e.g., segments of airways, vessels, or other non-cancerous lesions) are classified using hand-crafted features and supervised classifiers [CC12, SJG15]. Lastly, a candidate lung nodule is classified as being either malignant or benign. Clinical information [Gur93], shape, texture, and other radiomic features [CZX12, HKB14] are employed in these classification models. Markedly, while many current machine learning methods are applied to the lung nodule classification/diagnosis task, most fail to have consistent performance when given external datasets [SCL16].

The growing collection of screening data opens up the possibility of modeling the natural history of lung cancer progression, subsequently determining optimal screening intervals per individual to make screening programs more effective and efficient [DCT95, Duf05b, CLC08]. As shown in Figure 1.1b, the natural progression of lung cancer can typically be modeled as transitioning through three states: a disease-free state (State 1), a preclinical state detectable via screening but asymptomatic (State 2), and a symptomatic state (State 3) [CLC08]. The mean sojourn time (MST) measures how fast a disease progresses from a preclinical state to a clinical state. Various statistical and temporal methods [CDT96, WRB05, CLC08, WER11, TCM17] have been developed to estimate MST, such as Markov models and differential-equations-based methods. Despite many efforts, it is still challenging to accurately estimate MST for various subject cohorts using conventional methods due to observation error and data sparsity issues.

This research focuses on three areas to address some of the aforementioned challenges: 1) lung segmentation; 2) nodule classification and diagnosis; and 3) lung cancer progression modeling. The ensuing three sections detail the motivation for each task.

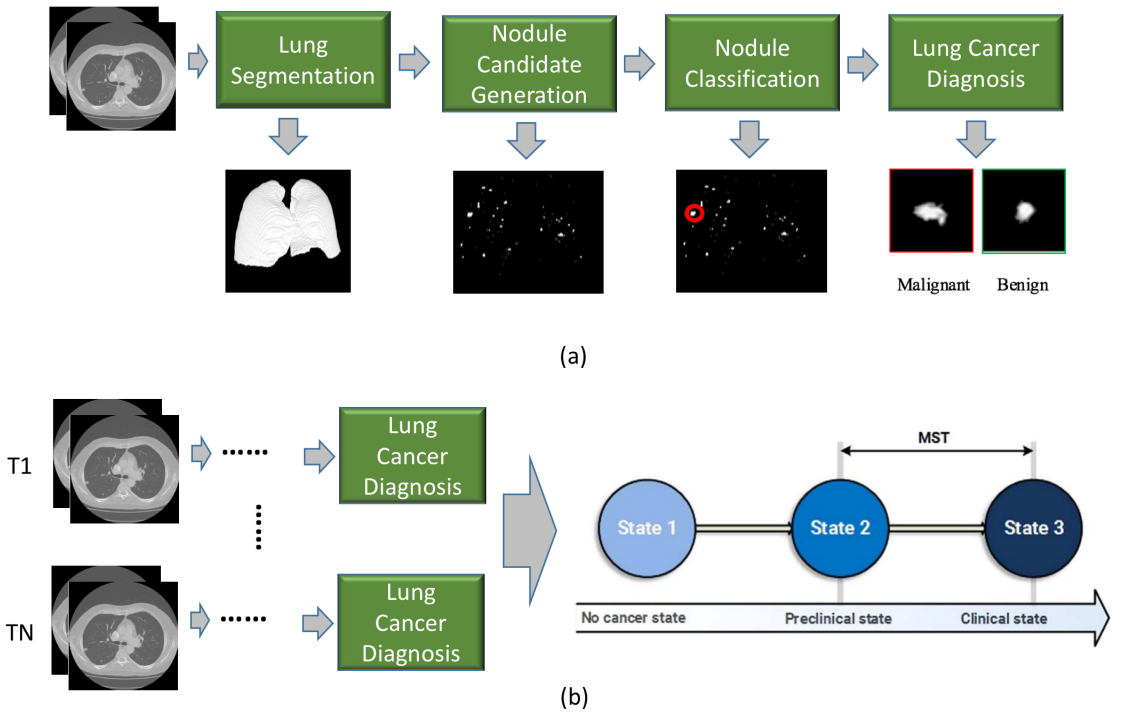


Figure 1.1: Computer-aided diagnosis of lung cancer and disease progression estimation.

1.2.1 Lung segmentation

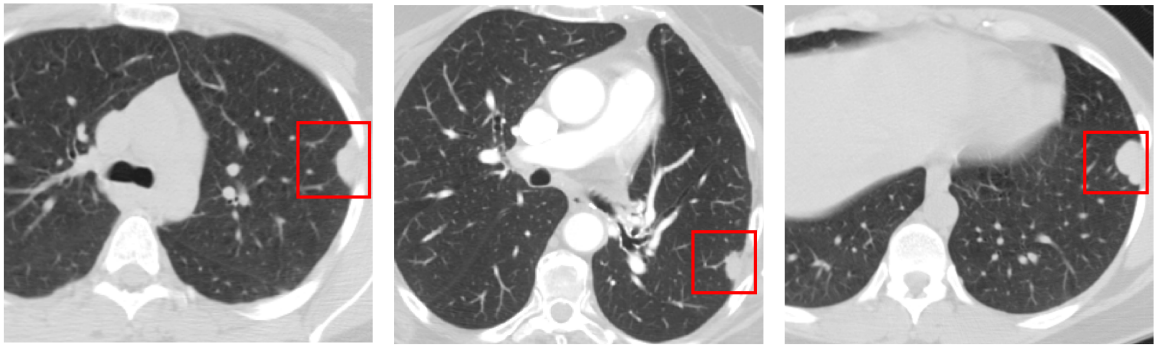


Figure 1.2: Illustrations of juxtapleural lung nodules.

Lung segmentation is a critical precursor in a pulmonary nodule CAD system, where the lung field is extracted and becomes the region of interest for detection and/or diagnostic tasks. This step sets the detection sensitivity upper bound for the whole system, as nodules in a "non-lung-field" region will not be found and analyzed. Juxtapleural nodules are one type

of pulmonary nodule that is commonly missed. While many works [AS04, AEF07, AF08] describe automatic lung segmentation of thoracic CT images, only a few explicitly handle the presence of juxtapleural nodules. Among the existing work that focus on juxtapleural nodules, most rely upon one (or more) predefined parameters, making algorithm performance sensitive to the variations in lung nodule shape and size. For instance, a "rolling ball" method has been employed to re-include the juxtapleural nodules for lung segmentation [AS04, BKN05, RDF08], comprising a morphological close operator with a round-shape structuring element. The effectiveness of the morphological operations is hence dependent on the predefined size of the selected "ball." As juxtapleural nodules vary in size and shape, selecting an optimal size that works well in all cases is difficult [BGM00, PRC08]. For example, a smaller sized structuring element will fail to capture larger-sized juxtapleural nodules; conversely, a large structuring element will cause over-segmentation and distortion of the local region. Moreover, evaluations in these works are often lacking [PRC08] or employ small test sets (e.g., dozens of cases) not fully representative of the range of characteristics of such nodules (e.g., variations in size, shape). Thus, it is highly desirable to have a novel parameter-free lung segmentation method, where no hand-picked parameter is required, to address the issues related to juxtapleural nodules.

1.2.2 Lung nodule classification and diagnosis

Dependent on the CT scanner and other real-world acquisition conditions, variability exists in image quality (e.g., signal to noise ratio, resolution) despite using published screening and diagnostic protocols. Although many methods have been developed for the lung nodule classification task, challenges exist with consistent performance when dealing with external datasets [SCL16]. One explanation for this failure is that the performance of machine learning methods heavily depend on the choice of representation for image content. Many traditional methods (e.g., support vector machines) rely heavily on feature engineering, which involves data preprocessing, transformation, and hand-crafted feature designs to identify discriminative features [BCV13]. Intensity, morphological, and texture features are often used to extract representations for lung nodules [AAW03, ZFF11]. One critical question raised is

how to define the optimal set of features that can encode the characteristics of a lung nodule [CHR15]. Another problem is that such low-level features may not fully represent the information of the image or capture the underlying statistical properties, given the complexity and heterogeneity of the data [SS13]. Present applications of machine learning methods further highlight a weakness of such approaches, as they are not able to adaptively learn the representation and discriminative information derived from raw data. This weakness limits their ability to be generalized to different tasks and using heterogeneous datasets. Opportunities thus exist to adapt and expand existing deep learning methods and technologies to learn complicated hierarchical abstractions and representations for lung nodules in a more data-driven fashion.

A barrier to adopting deep learning methods is their "black-box" approach, wherein interpreting and understanding how and why a model works remains a significant challenge [Lip16]. Being able to interpret the deep model is important for domain experts (e.g., radiologists) to understand the methods, integrate it into workflow, improve model performance, and ultimately enable clinical adoption. The problem lies, in part, in the fact that features generated by a deep learning method are unlike the conventional semantic features used by experts. For instance, lobulation and spiculation are widely used to describe nodules and diagnose lung cancer in CT images [HM16]. These semantic features represent domain knowledge long used in imaging interpretation, and may be helpful in building more robust prediction models. Thus, it is desirable to have a novel methods to provide interpretable deep learning methods for lung cancer diagnosis in conjunction with the domain knowledge captured by semantic features.

1.2.3 Cancer progression estimation

How fast lung caner progresses from a preclinical to an observable clinical state, the mean soujourn time (MST), defines how soon a lung cancer can be practically detected through imaging. Thus, MST is widely used [Duf05a] in the context of population screening, calculating the optimal interval between screens and estimating the extent of overdiagnosis.

MST varies given different imaging modalities and patient cohorts, and patients with higher MSTs (i.e., at lower risk of cancer) should have longer screening intervals. Individualized temporal models estimating MST can therefore help move screening recommendations from a traditional "one-size-fits-all" approach to more persoanlized policies. But several challenges exist in leveraging retrospective screening data to estimate MST. First, observations for disease states made in clinical practice are often subject to interpretation error, such as when radiologists incorrectly miss a cancerous nodule. Failure to model such observation error will bias any MST estimation [UHC10]. Second, missing or partial observations are common in clinical practice. For instance, some patients may miss a scheduled screening exam or undergo care at another facility where data is not shared. Third, the interval between screening exams is frequently irregular. Thus, the discretization of continuous time information results in the loss of valuable information [DCT95]. Fourth, the sample size of certain observed disease states may be very small (i.e., sparse), thus making estimation difficult. For example, patients will usually undergo an intervention if an early stage cancer is detected, thereby removing them from further observation. As a result, transitions to later states have fewer individuals with which probabilities can be estimated. Disease progression estimation models for periodic screening data are needed to overcome theses challenges.

1.3 Contributions

To address the issues described in Section 1.2, this work presents novel machine and deep learning methods to bridge the gap between screening and early detection of lung cancer. Three specific aims are defined:

1. *To develop a parameter-free lung segmentation method.* This novel bidirectional chain code method corrects the border of lung lobes to avoid excluding lung nodules attached to the boundary while minimizing over-segmentation error.

2. *To develop robust and transferable lung nodule detection models and interpretable lung cancer diagnosis models using deep convolutional neural networks (CNNs).* A well-known

limitation of current work is that most approaches do not generalize, with significant decreased performance with unseen datasets. A hybird ensemble convolutional neural network was first developed to detect lung nodules, combining the VGG, residual and densely connected module design. This work has been shown to achieve comparable and consistent results in independent datasets without fine-tuning. In addition, a new deep hierarchical semantic neural network (HSCNN) is described, the first such architecture for lung cancer diagnosis with interpretable radiology semantic features. This HSCNN framework predicts low-level radiologist-defined semantic features concurrent to predicting nodule malignancy, all in a single framework. Notably, predictions from this semantic neural network can be used to understand how the model works, in addition to serving as a semantic feature generator for unlabeled imaging datasets.

3. *To develop a novel statistical multi-state disease progression estimation model.* This model jointly models disease progression with observation error through the use of a continuous-time Markov model. A Bayesian approach is used to overcome problems caused by missing observations and data sparsity. This estimation model makes it possible to determine suitable screening periods more accurately. It also serves as the foundation to move towards individualized screening, where personalized screening paradigms.

Towards Aim 1, a method was developed for delineating the lung field ROI by automatically segmenting the lung lobe, correcting the border to avoid excluding nodules close to the lung boundary while minimizing possible over-segmentation. Notably, the algorithm addresses issues related to juxtapleural nodules. The developed method comprises three steps: 1) preprocessing to generate an initial lung lobe mask using adaptive thresholding; 2) detecting inflection points (both horizontally and vertically) to obtain all major concave and convex points along the lung lobe boundary using a bidirectional chain encoding method; and 3) correcting the lung boundary border using a support vector machine (SVM) to identify relevant pairwise connections based on extracted features, including position, concavity rate, and distance information.

For Aim 2, a 2.5D nodule detection model was created to classify pulmonary nodules versus non-nodule objects using a CNN. The method achieves robust and transferable results across different datasets by adapting an ensemble of state-of-the-art CNN architecture designs, including stacked convolution layer design, residual blocks, and densely connected modules. A hierarchical semantic deep convolution neural network was designed to differentiate benign versus malignant lung cancer. This network not only predicts lung cancer likelihood, but also outputs low-level semantic sub-tasks (e.g., spiculation and nodule diameter) simultaneously, which can be used to interpret and understand how the overall network's predictions. The information of each sub-task is also fed into the final malignancy prediction using a jump connection.

Lastly, for Aim 3, a 3-state progression model for lung cancer was developed (no-cancer, preclinical, clinical states). The transition between cancer states is represented in a continuous-time Markov model, which permits estimation of unobserved states and maintains the interval between screens unique per individual. A Bayesian framework is used to jointly estimate disease progression given observation error. The model also considers the inclusion of covariates for individualized disease progression rates estimation.

Collectively, the design and implementation of these Aims results in working imaging features and disease prediction models presenting a deeper understanding of lung cancer and improved diagnosis of disease.

1.4 Organization of the Dissertation

The dissertation is organized as follows. Chapter 2 provides background on lung cancer modeling, computer-aided detection/diagnosis, existing works, and current limitations. Chapter 3 describes Aim 1's work, a novel lung segmentation method using bidirectional chain coding and SVM focused on handling juxtapleural nodules. Chapter 4 details Aim 2, with two deep learning models for lung nodule detection and diagnosis, highlighting transferability and model interpretability. A novel Bayesian model is introduced in Chapter 5, addressing Aim 3 and a model for lung cancer progression and estimating mean sojourn time. Chapter

6 concludes by summarizing the results of this study and discussing the limitations of this work and future directions.

CHAPTER 2

Background

2.1 Lung Cancer

Lung cancer is the leading cause of cancer-related mortality worldwide [SJ15]. The American Cancer Society estimates that lung cancer accounts for 27% of all cancer-related deaths in 2015; and the 5-year survival rate is only 17% on average. One major fact related to this low survival rate is that only 15% of lung cancers are diagnosed in an early stage [SJ15], when no obvious cancer symptoms are evident. In contrast, survival increases upwards of 54% if the lung cancer is detected early on. Early detection of lung cancer is a driving issue, with research ongoing to optimize identification of nascent disease.

2.1.1 Computed tomography for lung cancer screening

Computed tomography (CT) is now the most widely used screening modality to detect early stage lung cancer. In 2011, the landmark National Lung Screening Trial (NLST) showed a 20% mortality reduction for individuals with lung cancer who underwent screening using low-dose CT (LDCT) relative to plain chest radiography [Tea11]. Subsequently, based on this evidence the United States Preventative Services Task Force (USPSTF) gave a Grade B recommendation that annual screening for lung cancer with low-dose CT (LDCT) be performed in adults ages 55-80 who have a 30 pack-year (number of packs of cigarettes smoked per day multiplied by the number of years an individual has smoked) smoking history and either currently smoke or have quit within the past 15 years [For15]. This policy has spurred development and implementation of new lung cancer screening programs using LDCT.

Interpretation of CT scans is both labor-intensive and potentially challenging. Early

stage lung cancer manifests itself as pulmonary nodules, which appear as small round- or oval-shaped opacities on CT studies with diameters less than 30mm [HBM08]. With a growing number of CT scans to read, as well as the increasing resolution (e.g., typical thoracic CT scans presently have 200-500 slices), interpreting such large sets of data may lead to visual fatigue and/or strain, contributing to a decrease in diagnostic accuracy [KBC12]. In addition, less experienced radiologists have marked variability in detecting subtle lung cancers, as interpretation heavily relies on past experience. Substantial variability in the performance between radiologists has been reported for the detection of lung nodules [ARK09]. The rich airway and vessel structure further complicates the interpretation of CT scans. Even among experienced radiologists, CT screening yields a large number of false positives, leading to a severe over-diagnosis. For example, in the NLST, a total of 96.4% of the positive screening results in the low-dose CT group were found to be false positives [Tea11]. The benefits of CT screening for detection of early lung cancer is likely to be reduced by the high false positive rates due to benign nodules [MFF03]; reducing these false positives and identifying patients who need intervention could reduce costs and morbidities associated with unnecessary invasive interventions [MW14]. Thus, it is increasingly imperative to distinguish benign from malignant nodules [Doi05]. However, there are substantial challenges related to differentiating benign from malignant nodules, and indolent vs. aggressive cancers [MW14]. As such, computer-aided diagnosis (CAD) systems have been actively studied to assist physicians in solving this problem.

2.2 Computer-aided Diagnosis (CAD)

Computer-aided diagnosis (CAD) is a major research area in medical imaging [Doi05, MW14]. The basic idea of CAD is to use the quantitative output from an algorithm (computer software) as the "second opinion" to assist the radiologist in interpretation tasks related to disease detection [Doi07]. CAD has two broad goals: 1) to improve the radiologists' reading accuracy and consistency; and 2) to shorten the image interpretation time. The general design of CAD is to detect (locate) suspect disease regions (e.g., a lesion) in one or several

medical imaging modalities/images; and/or to generate a likelihood score for the presence of a certain disease. These two designs correspond to CAD for disease detection and CAD for differential diagnosis [Doi05]. The basic technologies involved in these imaging-based CAD schemes are [Doi05]: 1) image processing (segmentation) to extract and detect abnormalities; 2) image feature generation to quantify abnormal candidates; and 3) classification of candidates through image features to differentiate normal and abnormal (benign and malignant) tissue.

CAD has been developed in various imaging modalities such as magnetic resonance imaging (MRI), CT, projectional radiography, nuclear medicine, ultrasound, and digital pathology imaging; and is used for all parts of the body, including the head, thorax, heart, breast, liver, prostate, and bones. Examples of CAD schemes developed in the past include pathological brain disease detection in MRI [ZWW10, ZWW11, ZDJ15]; detection/diagnosis of breast lesions on mammograms [GAG08] and MRI [NNC10]; detection of colorectal polyps in CT [NY07] and colon capsule endoscopy [MFF14]; detection/diagnosis of lung nodules in chest x-ray and CT [RDF08, CC13]; detection of coronary artery disease in CT [AGS10]; and identification of subjects with Alzheimer's disease in MRI [ZWD14, ZDP15]. Typically, each CAD system is tailored for a specific disease and imaging modality. In general, these CADs depend on a conventional machine learning scheme, segmentation of region of interest, feature extraction, and classification to make the final decision.

2.2.1 Automatic lung nodule detection and diagnosis

CAD [CC12, SJG15, PZL08] systems have been explored, establishing the potential to improve lung nodule detection/diagnostic accuracy in CT images. Previous studies have shown that CAD increases lung nodule detection rates [RLP05]; decreases false-positive rates [DMM06]; and compensates for deficient reader performance in the detection of the smallest lesions and of nodules without vascular attachment [ZTB13]. A computer-aided detection system may address elements of the following tasks: lung segmentation, nodule candidate generation, feature extraction, and classification.

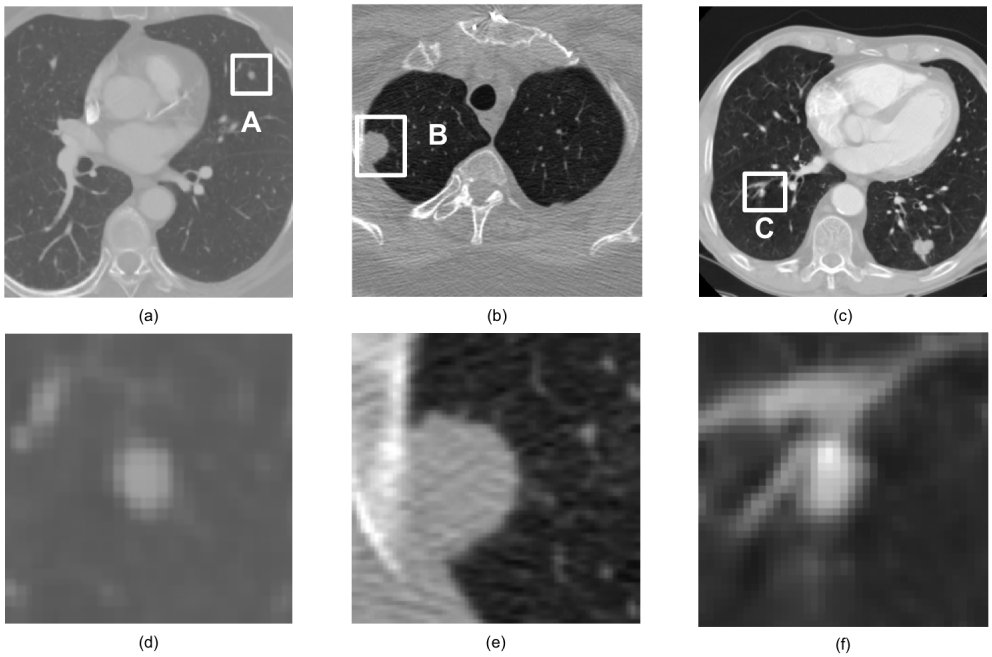


Figure 2.1: Three pulmonary nodule types: isolated, juxtapleural, and juxtavascular nodules: (a) CT slice with isolated nodule A; (b) CT slice with juxtapleural nodule B; (c) CT slice with juxtavascular nodule C; (d) magnified view of isolated nodule A; (e) magnified view of juxtapleural nodule B; and (f) magnified view of juxtavascular nodule C.

Pulmonary nodules can be grouped into three categories (Figure 2.1): isolated, juxtapleural, and juxtavascular. Isolated and juxtavascular nodules lay within the center of the ROI and are typically segmented without issue. But when lung segmentation fails to correctly define the lung boundaries, juxtapleural (or pleura-connected) nodules can be missed, and normal chest tissue outside of the lung can be included incorrectly as part of the ROI. In fact, evaluation of one CAD system found that 17% of all true nodules are missed due to poor lung segmentation [AS04]. Accurate lung segmentation is thus imperative to ensure accurate CAD system performance: the ability to identify true nodules ultimately sets the upper bound on CAD performance [RDF08]. Semi-automated segmentation methods [HAG82] have been previously employed, but the process of having an individual review each study is time consuming and arguably not scalable. Thresholding techniques [BGS03, KKC05] are widely used for initial lung field segmentation. Ross et al. [RED09] used adaptive

thresholding to generate the lung field. Active contours (snakes), level sets (LS), and other deformable boundary models [CV01] are also used to segment the lung lobe regions. A 2D geometric level set active contour method is employed in [SNM07]. The left and right lung lobes are automatically segmented with lobe boundaries initialized as the contour. Annangi et al. [ATR10] incorporated a prior shape term into a region-based active contour method and segment lung fields with a term describing edge feature points and a term representing region-based data statistics in x-ray images. 3D region growing and connected-component analysis is used in [YHS05] to extract the lung region. Ge et al. [GSC05] employed a k-means clustering method to perform pixel-wise lung filed extraction.

While many algorithms perform automatic lung segmentation of thoracic CT images, only a few explicitly handle juxtapleural nodules; however, evaluation is often lacking [PRC08] or a small test set not fully representative of the range of appearance of such nodules (e.g., variations in size, shape) is used. Hu et al. [HHR01] presented a fully automated lung segmentation method using combinations of morphological operations to ensure the inclusion of juxtapleural nodules. The effectiveness of the morphological operations was dependent on the shape and the size of the selected structuring element. As juxtapleural nodules vary in size and shape, selecting an optimal size and shape that works well in all cases is difficult. For instance, a smaller sized structuring element will fail to capture larger-sized juxtapleural nodules; conversely, a large structuring element will cause over-segmentation and distortion of the local region. Similarly, a "rolling ball" method has also been used [AS04, BKN05, RDF08], comprising a morphological close operator with a round-shape structuring element. Likewise, the size of the ball is hard to optimize across the variation observed in juxtapleural nodules, as noted in [BGM00, PRC08]. Pu et al. [PRC08] propose a point-wise lung segmentation algorithm, called adaptive border marching (ABM), designed to address juxtapleural nodules. An inclusion criterion was defined based on the ratio between the Euclidean distance of two points on the boundary and the maximum height perpendicular to their connecting line segment. This ratio was used to adaptively adjust the size of a search step for choosing point pairs by comparing itself to a fixed threshold. For all point pair candidates inside one concave region, only the outermost one (which forms the biggest convex

hull) was connected (this process is equivalent to the gift-wrapping algorithm [NSN12]). Selecting a threshold parameter to control over-segmentation across all cases is problematic. Varshini et al. [VBA12] extended ABM by merging two lung segmentations obtained using a small threshold and a large threshold separately, but provided no evaluation. Kim et al. [KKN03] also presented a contour-marching method to avoid peripheral nodule exclusion near the lung boundary. This method tracked the lobe boundary to detect suspicious areas with texture features similar to a true nodule. A region growing method was applied to each identified area to re-include it as part of the lung region. Markedly, texture features alone are unable to detect all juxtapleural nodule regions along a boundary. Defining the search window and threshold for region growing method are also inherent challenges to this approach. Ye et al. [YLD09] used a Freeman chain code to correct the contour of a lung lobe. For all pixels along the boundary, chain codes are used to detect critical points by examining the transition between concave and convex points, determined by a predefined threshold value. All critical point pairs were then connected to form a revised border. Using a preset threshold to define concave/convex regions may not be effective across the natural variation seen in imaging studies and anatomy; and connecting all detected critical point pairs may lead to over-segmentation. Choi et al. [CC12] detailed a similar chain code method, but rather than detecting transitions to find critical points, gradient information was extracted from the chain code and only connecting point pairs whose change in gradient value is below a given threshold value are considered. Both methods detect convexity changes in only the horizontal (or vertical) direction, which is in general not sufficiently robust to correct under-segmentation (as illustrated in Section 3.2). Ko et al. [KB01] use a curvature-based method to correct the initial lung mask. The curvature for each point on the boundary is calculated to detect a rapid change and a segment is inserted to correct such regions. As can be seen from the above methods, all rely upon one (or more) predefined parameters (which may vary between CT scans), making algorithm performance sensitive to the normally observed distribution in lung nodule shape and size. Moreover, little validation is given on the effectiveness of the proposed border correction methods on clinical data. This highlights the needs for the proposed method, which will be presented in Chapter 3 and assessed over

a large dataset. This proposed approach eliminates the need for predefined parameters in order to operate on the full spectrum of nodule sizes/shapes.

After segmenting the lung lobes, nodule candidates were identified inside the lung fields. Multiple gray-level thresholding techniques [MHR10, CC12] combined with morphological opening operations were applied to segment suspicious subjects. Pixel intensities ere transformed to Hounsfield units, with values for soft tissues falling into known ranges. Opening operations were used to separate nodules with attached vessels or airways. After multiple thresholding, Choi et al. [CC12] adopted rule-based pruning to remove obvious non-nodule objects based on maximum and minimum volumes, radii, and areas. The rules were defined corresponding to the purpose of detecting lung nodules between 3 to 30 millimeters. McCulloch et al. [MKM04] proposed a method employing hybrid multi-stage models. A three-level modeling architecture consisting of anatomy (top level model), shape (middle level model), and signal (bottom level model), identified and classified different regions such as lung nodules, blood vessels, and lung parenchyma. Domain knowledge was built into the mathematical models to quantify different regions. A Bayesian model selection architecture determined the most probable model for each region inside the lung lobes comparing against existing representations. Regions were considered to be suspicious nodule candidates if the nodule model provided the highest probability among all models. Nodule enhancement filters were also designed to apply as a preprocessing step prior to enhancing the low-contrast nodules or to separate the nodule candidates with other attached structures. Ge et al. [GSC05] identified nodule candidates using a weighted k-means clustering segmentation with two output clusters, nodule and background clusters. Image features were calculated pixel-wise from both the original image and a median-filtered image, and were used as the input for the cluster. The classification criterion for each pixel was the ratio of distances from the feature vectors to the center of these two clusters. Paik et al. [PBR04] developed a surface normal overlap score-based method as an enhancement filter for nodule candidate detection. This method was based on a Hoffman transform and each voxel generated a score proportional to the number of surface normal lines that pass through a neighborhood of the voxel. Li et al. [LS03] proposed a selective enhancement filter to enhance lung nodules and suppress airways,

Table 2.1: Review of current CT features used for lung nodule classification.

Type	Feature Name	Type	Feature Name
Morphological feature	Area	Morphological feature	Eccentricity
Morphological feature	Diameter	Morphological feature	Perimeter
Morphological feature	Circularity	Morphological feature	Distance to center of current lung
Morphological feature	Compactness	Morphological feature	Bounding box
Morphological feature	Principle axis length	Morphological feature	Projection elongation
Morphological feature	Projection compactness	Morphological feature	Principle axis ratio
Morphological feature	Mean breadth	Morphological feature	Euler-Poincare
Intensity features	Mean inside	Intensity features	Mean outside
Intensity features	Minimum value inside	Intensity features	Maximum value above
Intensity features	Minimum value outside	Intensity features	Maximum value below
Intensity features	Contrast	Intensity features	Standard deviation inside
Intensity features	Standard deviation outside	Intensity features	Skew inside
Intensity features	Kurtosis inside	Intensity features	Moment 2-7
Gradient features	Radial-deviation mean outside	Gradient features	XY gradient magnitude separation inside
Gradient features	Radial-deviation mean inside	Gradient features	Radial-deviation standard deviation inside
Gradient features	Radial-deviation standard deviation separation	Context features	Distance to top of the lung in Z
Context features	Distance to carina in X, Y, and Z	Context features	Distance to pleural wall

vessels, and other non-nodule tissues. In [PBR04, LD04], nodule candidates with a small size were enhanced using a series of 3D cylindrical and spherical filters. Template matching [LHF01] was also used to detect circular and semicircular nodule candidates. Bae et al. [BKN05] used spherical-shape and morphological matching filters to enhance juxtavascular nodules, with four different kernel sizes ranging from 3 to 30 mm.

Conventional methods [LHF01, LS03, LD04, BKN05, MHR10, CC12] employed feature engineering to define meaningful features to characterize nodules (Figure 2.2 illustrates lung nodules in CT images). The feature extraction step generated quantitative characteristics (features) on the lesion candidates. The classifier took these features (of the training data) as input to search for an optimal boundary in the feature space to separate nodule and non-nodules (or benign and malignant). The accuracy of the boundary highly relied on the

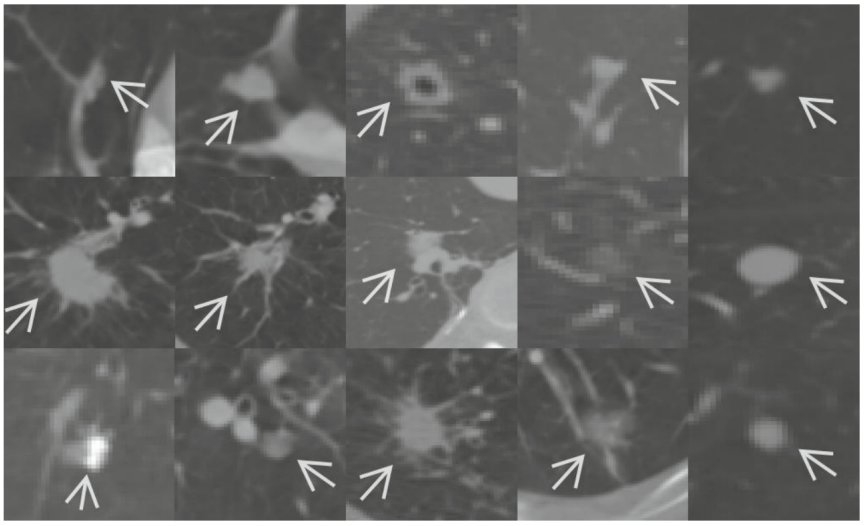


Figure 2.2: Illustrations of lung nodules in CT images [DBS15].

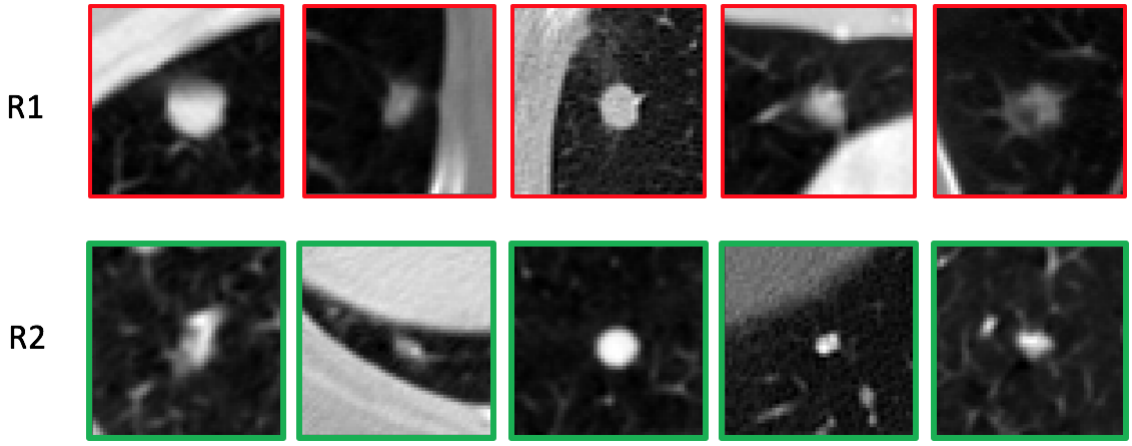


Figure 2.3: Illustrations of malignant and benign nodules: R1 are malignant nodules; R2 are benign nodules.

quality of the extracted features. For this reason, much of the actual effort for the CAD system has been put into the design of the preprocessing pipeline and feature extractor to result in representations to best support the subsequent classifier. A large number of types of features has been developed, including morphological, texture, gradient-based, and context features. Morphological features quantify shape-based information of the candidates using the candidate segmentation mask. Texture and gradient-based features captured the visual patterns and intensity information of the regions inside or outside the nodule candidate from

the voxel intensity values of the CT images. Context features measured the relative position information of the candidates. Table 2.1 summarizes some commonly used features. Feature-based classifiers were then used to perform the detection task. The purpose was to find the optimal boundary in the high-dimensional (transformed) feature space to separate the desired classes using methods such as support vector machines [YLD09], decision trees [ZFF11], artificial neural networks [CZX12], logistic regression [YLD09], linear discriminant analysis [MHW99], etc. Froz et al. [FCS17] employed a hybrid model to extract the texture features of lung nodule candidates in CT scans using artificial crawlers and rose diagram methods. The support vector machine (SVM) classifier with a radial basis kernel was adopted to detect lung nodules. Jing et al. [JBL10] also classified lung nodules using SVM. The nodule candidates were characterized based on shape, geometry, grey level and texture features after applying a rule-based approach to exclude obvious blood vessels from the images. Tartar et al. [TKA13] extracted statistical features using 2D principle component analysis and geometric features using regional descriptors for nodule candidates. The best features were selected with minimum redundancy maximum relevance method. Lung nodules were then detected using artificial neural network with the best features as input. Recently, deep learning methods have also been applied to lung nodule detection task and shown superior performance compared to conventional methods. Ginneken et. al. [GSJ15] used a CNN trained for a natural image recognition task off-the-shelf to extract image features, and a support vector machine was built for lung nodule detection. Lo et. al [LCL95] applied CNNs for the pulmonary nodule detection in 2D chest radiography images. Setio et. al. [SCL16] employed a 2D CNN and fused multi-views on CT images for nodule detection. Although these works have deployed convolutional neural network for nodule detection, but very few evaluated the transferability and robustness of the developed model using external datasets. Most studies trained and validated the models on the datasets collected from the same source, and no external validation is provided. It is challenging for conventional model-based lung nodule classifiers trained on one dataset to achieve comparable performance on a completely independent dataset. There are three reasons for such difficulties: 1) image quality (signal to noise ratio, resolution) varies significantly depending on the imaging hardware and protocol

followed; 2) lung nodules appearance are largely heterogeneous with a wide variation in shapes, sizes, and types; and 3) nodules from different categories are highly imbalanced and have different distributions depending on the datasets [SCL16]. Models with better transferability is highly desirable and needed for practical clinical usage.

After detecting the lung nodules, a lung nodule diagnosis task was usually performed after to differentiate malignant from benign ones. Figure 2.3 presents examples of benign and malignant nodules. A number of studies have looked at this task, which can be grouped into three categories based on the source of characteristics employed: 1) the diagnosis of lung nodules based on nodule features in single CT scan as shown in Table 2.1; 2) the diagnosis of lung nodules based on growth rate across multiple longitudinal CT scans; and 3) the diagnosis of lung nodules based on the fusion of positron emission tomography (PET) and CT images. This dissertation focuses mainly on the first category using single CT scans. Armato et al. [AAW03] segmented the lung nodule using multilevel thresholding techniques; extracted morphological and gray-level features; and classified nodules as benign or malignant using linear discriminant analysis. Gurney et al. [Gur93] developed a model based on Bayesian analysis to predict the malignancy probability of lung nodules with semantic radiographic features determined by radiologists and clinical findings. This study was reported to achieve significantly better results compared to human readers. However, an automated method to quantify the semantic features is missing. Kawata et al. [KNO98] presented a method to classify lung nodules as benign or malignant by measuring 3D surface characteristics using curvature and ridge lines. Kawata et al. [KNO01] developed a method employing internal and surrounding structure features to distinguish malignant and benign nodules. After selecting the optimal subset of features, classification scores were obtained using a linear discriminant classifier. Dilger et al. [DUJ15] improved nodule diagnosis accuracy by using lung parenchymal surrounding features as additional information. 32 parenchymal features, 8 nodule features, and 5 global features were employed and an artificial neural network (ANN) classifier used to classify the nodule. Zinovev et al. [ZFF11] employed both texture and intensity features using belief decision trees and a multi-label approach to perform lung nodule classification. Way et al. [WSC09] segmented lung nodules using k-

means clustering, combined nodule surface features together with texture and morphological features, and used linear discriminant analysis to diagnose malignant lung cancers. Notably, the pertinent portion of these extracted features, such as volume and shape, are sensitive to the underlying variability arising from the lung nodule segmentation results. Thus, using segmented regions may lead to inaccurate features, with downstream erroneous outputs into the classifier [SZY17]. Another critical question raised by this type of CADx design was how to define the "optimal" subset of features that can best encode characteristics of the lung nodule [CHR15]. The selected best feature set usually depends on the the training dataset, feature selection and classification methods, and is hard to achieve comparable results on different settings.

To overcome this issue, deep learning methods [SZY15, SZY17, CHR15, KWC15, HHH15], particularly convolutional neural networks (CNNs), have recently been used for lung nodule diagnosis models, with promising results. These deep learning models are capable of adaptively learning image representations data-driven, taking raw image data as input without relying on *a priori* nodule segmentation masks or handcrafted feature designs. Kumar et al. [KWC15] first trained an unsupervised deep autoencoder to extract latent features from 2D CT patches. These extracted deep features were combined with decision trees to predict lung cancer. Hua et al. [HHH15] employed supervised techniques with a deep belief network and CNN to train models to classify lung nodules as benign or malignant. Their models outperformed two baseline methods: using scale-invariant feature transform (SIFT) features and local binary patterns (LBP) [FAG11]; and using fractal analysis [LHL13]. Ciompi et al. [CHR15] used pre-trained CNN models to classify candidates as peri-fissural nodules (PFNs) or non-PFNs. Deep features were extracted from the pre-trained model for three 2D image patches in axial, coronal, and sagittal views. An ensemble of deep features and a bag of frequency features were then used to train supervised binary classifiers for the PFN classification task. Shen et. al. [SZY15] designed a multi-scale CNN using 3D nodule patches at three scales to perform the lung cancer diagnosis task. This study was further extended in [SZY17] by adding a multi-crop pooling strategy to improve model performance. Markedly, these cited works use deep learning as a "black-box" and are not able to explain what repre-

sentations have been learned or why the model generates a given prediction. This low degree of interpretability arguably hinders domain experts, such as radiologists, from understanding how the models work and ultimately impedes model adoption for clinical usage. As discussed in [JCO15], improved interpretability is helpful to improve the radiologist-CAD interaction to allow radiologists to calibrate their trust in the CAD system. Moreover, human domain knowledge regarding the differentiation of benign versus malignant nodules is able to account for the wide scope of observed lung nodule heterogeneity. Nonetheless, domain knowledge is presently not incorporated into deep learning frameworks.

A number of radiologist-quantified features, derived from CT scans, have been considered diagnostically relevant to the lung nodule malignancy assessment [KPG15, ECM00]. These features are referred to as *semantic* diagnostic features in this study. Examples of such semantic features are nodule spicularity, texture, lobularity, and margin. Although the definitions of these features are qualitative in nature, studies have shown that these semantic features can be quantified numerically using low-level image features [KC15]. Hancock et al. [HM16] demonstrated that machine learning models can achieve high prediction accuracy for lung cancer malignancy using only semantic features as inputs. In addition, these semantic features are radiologically meaningful in nature and widely accepted by radiologists. As presented in Chapter 4, an opportunity exists to incorporate these semantic features into the design of deep learning models, combining the advantages of both.

2.3 Deep Learning Methods

Traditional machine learning methods are limited in their ability to process natural data in their raw form (raw image pixels), thus relying on feature engineering and domain knowledge to identify and extract meaningful values from the raw data into learnable representations [LBH15]. In contrast, representation learning is a class of methods able to automatically discover the optimal representation of the raw data and derived features to facilitate tasks related to classification/prediction/detection [BCV13]. Deep learning is one type of representation learning method that attempts to transform raw data into hierarchical representations

by composing multiple levels of nonlinear processing modules [BCV13, Ben09]. The key aspect of deep learning is that these multiple layers of features are learned from raw data using a general-purpose learning procedure, instead of being designed in a handcrafted fashion by engineers [LBH15]. Deep learning methods have been applied to various detection and classification tasks and have significantly improved the state-of-the-art in myriad domains, such as speech and signal recognition [DMH10, DSY10, SLY11], object recognition [KSH12, KTS14] and natural language processing [MDK11]. This success also demonstrates that the adaptive representation learning ability of deep learning is better able to capture and extract the intricate structures in high-dimensional data relative to traditional feature engineering. Deep learning may achieve more success in the future due to the fact that it requires less engineering by hand [BCV13, Ben09].

Various deep learning methods have been proposed, including deep belief networks [HOT06, BLP07], deep Boltzmann machines [SL10], recurrent neural networks [FN93, Sut13], stacked autoencoders [Ben09, HS06] and convolutional neural networks [LBB98, KSH12]. The last method, which is a supervised learning method, receives the most attention in this dissertation, and is introduced in the subsequent sections.

2.3.1 Learning a deep network

Supervised learning refers to inferring functions from labeled training data [MRT12]. For example, suppose we want to build a system to classify images that contain dogs or cats. We would first collect a set of labeled images, containing dogs or cats. The system will read each image and output two classification scores, one for each category. During training, we tune the system so that it produces the highest classification score when it categorizes the image correctly. This training is performed by iteratively minimizing a loss function to adjust the weights of the deep learning system via gradient decent optimization and backpropagation. Key concepts include the following:

- **Loss function**. A loss function is a function that quantifies the agreement between the desired classes and prediction results. The training process seeks to minimize a

loss function.

- **Weights**. Weights are adjustable parameters (real numbers) that define the input-output function of nonlinear modules. In a typical deep learning system, there may be hundreds of millions of weights, which are learned using a large set of labeled training samples [LBH15].

- **Gradient descent**. To properly adjust the weights to minimize the loss function, a gradient will first be computed for each weight, aggregated into a gradient vector. The opposite direction of the gradient vector indicates the direction of the fastest descent for the averaged total loss. Thus, the weight vector is adjusted in the direction of the negative gradient. In practice, each optimization step only uses a batch of training samples, instead of all samples, to compute the outputs and errors. This procedure is named mini-batch (stochastic) gradient descent. The process is repeated for many small batches of the training data until the average loss function achieves a minimum value. Mini-batch gradient decent has been shown to have a faster convergence rate compared to the general gradient descent method [BB08].

- **Backpropagation**. The backpropagation procedure is used to compute the gradients of a loss function with respect to the weights of the multilayer stack of modules (deep learning architecture) through recursive application of the chain rule. The intuition is that the gradient of the objective with respect to the input of a module can be computed by working backwards from the gradient with respect to the output of that module [LBH15]. In a deep leaning system, the gradients from the top output layer to bottom input layer can all be computed and propagated through backpropagation equations.

2.3.2 Convolutional neural networks

The architecture of convolutional neural networks is specially designed for processing image data (multiple array data). This design is able to preserve the original data structure as well as to generate hierarchical representations. Figure 2.4 illustrates a typical CNN consisting

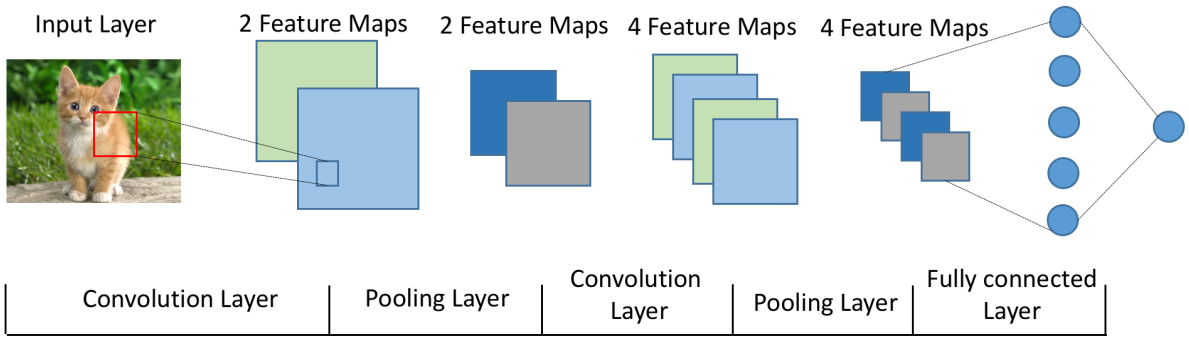


Figure 2.4: An example of a convolution neural network architecture.

of multiple stages of processing layers from left to right. The CNN typically has four types of layers: a convolutional layer, a pooling layer, a fully connected layer, and a classification layer. The convolutional and pooling layers are central to the design, and typically occur in the first few stages (as shown in Figure 2.4). The convolutional layer exploits the properties of natural signals (e.g., images) by assuming local connectivity, weight sharing, and pooling. Key concepts are as follows:

- **Convolutional layer**. As shown in Figure 2.4, convolutional layers are organized into feature maps, designed based on the principle of local connectivity and weight sharing. Local connectivity refers to the fact that each unit (neuron) within a feature map is connected only to local patches of the feature map in the previous layer through a set of weights, called a filter bank [LBD90, LBH15]. Within a feature map, one filter bank is shared among all units; this represents the weight sharing paradigm. Moreover, different filter banks are used for different feature maps. The reasoning behind local connectivity and weight sharing is to minimize the number of parameters while exploiting the fact that the local neighborhood of pixels is highly correlated, and local statistics of images are location invariant [LBB98, LBH15]. The weighted sum for each unit is then passed into a nonlinear transformation function, called an activation function. The activation function enables the nonlinear transformation for the information passed to subsequent processing layers.

- **Pooling layer.** The pooling layer performs a down-sampling operation to merge (semantically) similar features from the convolutional layer into one. A unit within a pooling layer takes a local patch from a previous feature map (convolutional layer) as input and computes the maximum or average value of the patch as its output. Thus, it reduces the dimension of the representation by reducing the number of parameters needed in subsequent layers and improves the robustness of the representations by establishing invariance to small shifts and distortions.

- **Fully connected layer.** Units in this layer are fully connected to all units in the previous layer, as seen in a regular neural network (i.e., multiple layer perceptron).

- **Classification layer.** This layer defines how to compute the prediction score using all outputs from previous layers. Typical classification layers are implemented using a softmax classifier or linear support vector machines [NNF13]. The training algorithm for CNN is the same as general supervised training steps introduced in Section 2.3.2 with gradient descent and backpropagation.

2.3.3 Deep learning in medical image analysis

Deep learning techniques have achieved significant success in object detection tasks of natural images, and more recently have been applied to a variety of medical image analysis tasks. But a major challenge in the clinical domain is the availability of labeled datasets. For natural image recognition tasks, much of the success comes from access to copious amounts of labeled training data (e.g., millions of tagged images), which enable the training of complicated and deep neural networks with millions of parameters. In the medical domain, the amount of labeled training data is much smaller (e.g., hundreds of medical imaging scans). Another issue in clinical imaging applications is the 3D structure of most medical images. A typical magnetic resonance imaging (MRI) or CT scan contains multiple slices, with disease regions inside the scan relatively small compared to the whole image stack (and of varying 3D size). Considering that cross-sectional imaging slice thickness is always much larger than the in-plane pixel size, efficiently capturing 3D information in a deep neural network design is

complex.

Several studies have used deep learning techniques to detect/segment abnormal disease regions on 2D medical images. Su et al. [SLX15] applied a CNN to segment breast cancer regions in histological images. Similarly, Cruz-Roa [CBG14] also employed a CNN to detect invasive ductal carcinoma tissue regions in whole slide images of breast cancer. Ertosun et al. [ER15] developed an automatic grading system for gliomas using assembled CNNs in digital pathology images. Cruz-Roa [COM13] et al. proposed a CNN-based deep learning architecture to detect basal-cell carcinoma in digital pathology images. Ciresan et al. [CGG12] applied similar techniques to train a CNN to segment neuronal membranes in electron microscopy images. Large numbers of 2D training data were obtained in these studies by randomly cropping the large digital images into small patches, and each patch is considered a training sample. Ronneberger et al. [RFB15] proposed a U-net architecture for segmentation of neuronal structures in electron microscopic stacks. This method was composed of a fully-convolutional structure followed by an up-sampling structure to increase the image size. Skip-connections were employed to directly connect opposing contracting and expanding convolutional layers.

For 3D medical images, developing deep-learning-based systems has been more challenging. Roth et al. [RLL16] increased the number of 3D sample sizes by aggregating random views of 2D slices to train a CNN for CAD systems, and evaluated the approach in sclerotic metastases detection and lymph node detection tasks. Suk et al. [SS13] proposed to classify Alzheimer's disease and mild cognitive impairment with deep features extracted from stacked autoencoders in MRI/PET images. To overcome the limitation of data size, low-level features are used as the input for the autoencoders instead of raw data. Ciompi et al. [CHR15] and Ginneken et al. [GSJ15] use pre-trained CNNs (OverFeat) for natural image detection as deep features extractor to perform lung nodule diagnosis tasks in CT images. Yu et al. [YYC17] proposed a volumetric convolutional neural network with mixed residual connections to automatically segment the prostate in 3D MRI studies. The combination of residual connections is employed to improve the training efficiency and discriminative capability of the network using both local and global information.

Notably, these studies using deep neural networks are a "black-box" approach and few efforts are made to make the model *interpretable*. Chapter 4 details further related deep learning works for lung nodule classification and diagnosis, and motivates the proposed novel hierarchical semantic convolutional neural network to incorporate domain knowledge and explain deep learning outputs.

2.4 Multi-state Disease Progression

A better understanding of lung cancer's progression and dynamics, such as the expected time to reach a certain disease state, may lead to more appropriate prevention, management and treatment; as well as early detection [MOH14]. Periodic screening using imaging is one of the most common ways to detect early stage lung cancer. Longitudinal data collected as a result of screening [ZL86] provides an opportunity to discover better approaches for characterizing natural disease progression and generate predictions for individualized screening or diagnostic policies [ZWH14]. Traditionally, a "one size fits all" approach has been used for screening programs; however, patients at lower risk of cancer should have longer screening intervals or not be screened at all. The MST measures how fast a disease progresses from a preclinical state (imaging detectable but without observable symptoms) to a clinical state (with observable symptoms), and has been widely used [Duf05b] to model disease progression and in the context of population screening, to calculate the optimal interval between screens and estimate the extent of overdiagnosis.

Numerous techniques for modeling multi-state disease progression, especially for MST, have been proposed. Aalen et al. modeled HIV/AIDS progression using a discrete-time Markov model [AFA97]; Chen et al. presented a three-state discrete progressive model for breast cancer [CP83]. Multi-state continuous-time Markov models can be adapted to solve the loss of continuous-time information [DCT95] due to interval censoring. In particular, they have been used to model hepatocellular carcinoma [Kay86], liver cirrhosis [AHK91], periodontal disease [MOH14], and diabetic retinopathy [MJ95]. Duffy et al. applied a three-state continuous Markov model to data from a breast cancer randomized controlled trial

to estimate the MST and the sensitivity of the screening process [DCT95]. This method assumes perfect sensitivity in estimating the transition times between states and then subsequently estimates the sensitivity using fixed transition times. Chen et al. extended and applied the continuous-time Markov model in breast screening to jointly estimate mean sojourn time, screening sensitivity, and the positive predictive value [CDT96]. Nevertheless, information from the control group (e.g., individuals who received usual care) was needed to properly estimate the desired parameters. Bayesian approaches have been increasingly applied [WRB05, CLC08, WER11, KEW12, CEW14] to infer MST and screening sensitivity. In contrast, my model is capable of modeling the situation where no control group information is available. This is especially relevant in clinical settings when it is unethical to deny treatment.

A Bayesian framework applied to breast cancer screening data was used in [WRB05] to obtain age-dependent sensitivity and estimates of transition probabilities. Chien et al. applied a Bayesian approach to validate the effectiveness of computed tomography (CT) for mortality reduction in lung cancer and to estimate the MST [CLC08]. In 2010, Wu et al. used data from the Mayo Lung Project (MLP) to estimate lung cancer screening sensitivity, age-dependent transition probability between states, and the distribution of sojourn time using a Bayesian approach [WER11]. Bayesian methods have advantages over classical techniques such as enabling small sample inference, providing appropriate measures of uncertainties, allowing inference on non-linear functions of parameters, and constructing predictive distributions to allow for additional inferences of interest [WRB05]. More recently, Jiang et al. [JWB16] used the Day and Walter model [DW84] to estimate the MST and the false negative rate from the Ontario breast cancer screening program in Canada. Taghipour et al. [TCM16] modeled the natural history of breast cancer with a 4-state hidden Markov model and analyzed the effects of covariates and over different subpopulations. Jia et al. [JBS16] used a 5-state Markov model to detect the worsening of patient symptoms in order to prioritize by symptom severity. Ma et al. [MCT16] used a Bayesian approach on a 5-state continuous time Markov model to investigate a transtheoretical model.

The advent of lung cancer and in particular lung screening trials also stimulated the

development of a number of risk models to predict lung cancer incidence from epidemiological and clinical data. Bach et al. [BKT03] developed and combined two logistic regression models that predict the 10-year cumulative probability of dying from lung cancer and dying without lung cancer. Conin et al. [CGZ06] validated this model with the placebo arm of the Alpha-Tocopherol Beta-Carotene Cancer Prevention (ATBC) study. The model underestimated the observed lung cancer risk and the observed non-lung cancer risk individuals that smoked less than 20 cigarettes per day. A Cox proportional hazards regression was developed from the COSMOS trial from epidemiological and clinical data [MBB11]. Model performance was poor on early cancers but it could identify lower risk individuals and prevent overdiagnosis. Using the PLCO dataset, Tammemagi et al. [TPC11] developed a logistic regression model that predicts the six year probability of cancer, and were validated using AUC, epidemiological and clinical factors. Petousis et al. [PHA16] developed discrete time dynamic Bayesian networks (DBNs) that predict lung cancer incidence at the different screening points of the National Lung Screening Trial (NLST). The models achieved results comparable to expert's decisions. In Chapter 5, I extend previous probabilistic models and demonstrate how my developed approach yields a more accurate picture of lung cancer progression.

CHAPTER 3

Automated Lung Segmentation in CT images

3.1 Overview

Computer-aided detection and diagnosis systems (CADs) have been widely investigated to improve radiologists' interpretive accuracy in finding and characterizing lung disease. Lung segmentation is a requisite preprocessing step for most lung CAD schemes. This chapter details a novel and parameter-free lung segmentation algorithm with the aim of improving lung nodule detection accuracy, focusing on juxtapleural nodules. A bidirectional chain encoding method, combined with a support vector machine (SVM) classifier, is used to selectively smooth the lung border while minimizing the over-segmentation of adjacent regions. The remainder of this chapter is organized as follows. Section 3.2 describes the dataset used for this study. Section 3.3 details the developed bidirectional chain coding and automated border correction methods. Section 3.4 describes evaluation methods and results for this technique. Section 3.5 concludes with a discussion of the strengths and limitations of this method. This Chapter is based on the content of [SBC15].

3.2 Dataset

The proposed method was validated using data from LIDC [AMM04, AMB11], available through The Cancer Imaging Archive (TCIA). LIDC contains both screening and diagnostic thoracic computed tomography scans collected from 7 academic centers and 8 medical imaging companies. For CT scans, inclusion criteria were: 1) having a collimation and reconstruction interval no greater than 3 mm; and 2) each scan approximately containing no

more than 6 lung nodules with the longest dimension ≤ 30 mm and ≥ 3 mm, as determined by a cursory review during case selection at the originating institution [AMB11]. For the whole dataset, the slice thicknesses were between 0.6 and 5 mm, and the in-plane pixel size varied from 0.461 to 0.977 mm. LIDC-IDRI comprises 1,018 cases (representing 1,010 different patients, 8 patients having 2 distinct scans), with each including images from a clinical CT scan and an associated XML file. The XML files record the reference standard for lung nodule locations, as manually annotated by four radiologists following a two-phase image annotation process.

3.3 Methods

The developed method mainly consists of three steps (Figure 3.1): 1) preprocessing to generate an initial lung lobe mask using adaptive thresholding (Figure 3.1d, e); 2) detecting inflection points (both horizontally and vertically) to obtain all major concave and convex points along the lung lobe boundary (Figure 3.1f, g); and 3) correcting the lung boundary border using a support vector machine (SVM) to identify relevant pairwise connections (Figure 3.1h) based on extracted features. The details for each step are described as follows.

3.3.1 Preprocessing

Preprocessing uses Otsu's adaptive thresholding [Ots79] method to automatically obtain an initial lung mask based on the pixel intensity distribution of the input CT image. This method uses discriminate analysis to exhaustively search for a threshold value that minimizes the intra-class variance between two regions of an image. For a given image, let L represent the gray level of all the pixels $[1, 2, \cdots, L]$. By choosing a threshold at gray level k, the pixels are divided into object class C_0 and background class C_1. Let w_0 and w_1 be the probabilities of C_0 and C_1 separated by a defined threshold and let σ_0^2 and σ_1^2 be the variances of these two classes. The intra-class variance is defined as the weighted sum of these two variances:

$$\sigma^2_{Intra}(k) = w_0(k) \times \sigma_0^2(k) + w_1(k) \times \sigma_1^2(k) \tag{3.1}$$

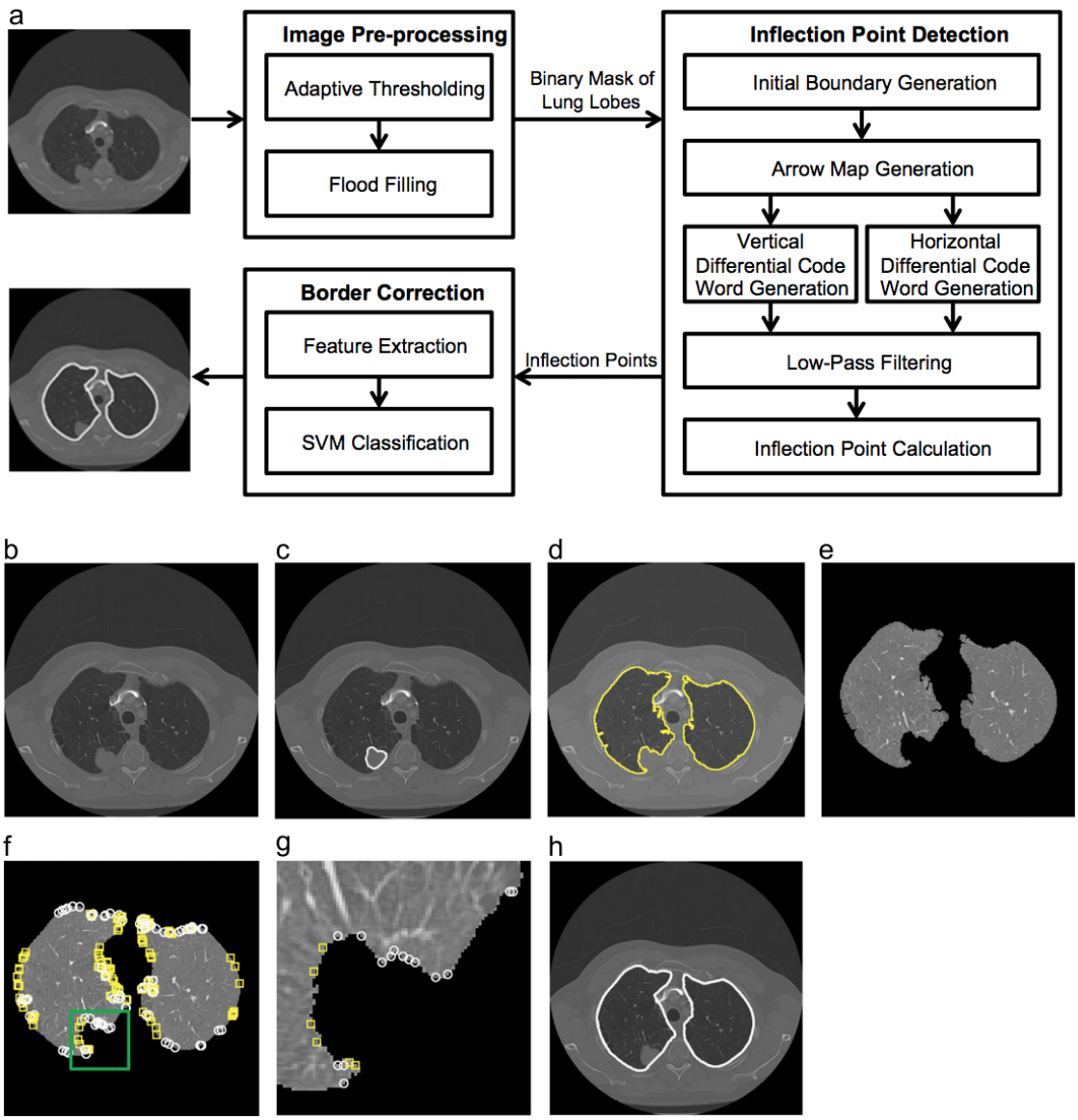


Figure 3.1: Diagrams depicting the proposed method and its outputs for a representative case. (a) Flow diagram of the proposed method; (b) original image; (c) original image with juxtapleural nodule outlined in white; (d) lung boundaries obtained after preprocessing; (e) lung lobe mask obtained after preprocessing; (f) detected inflection points shown in yellow-squares/white-circles; (g) magnified view of inflection points; and (h) results after border correction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

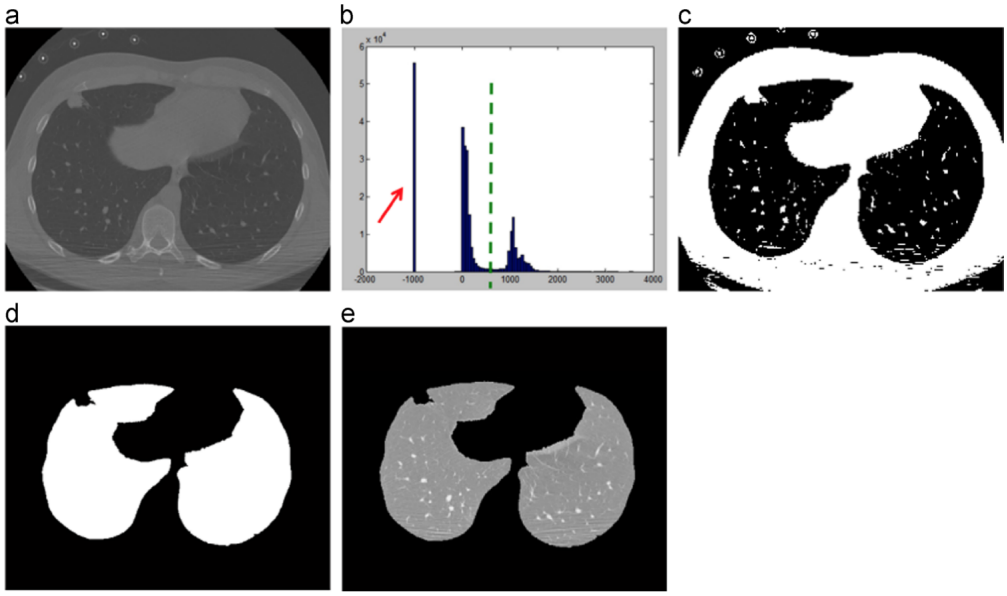


Figure 3.2: Basic steps for preprocessing. (a) Original image; (b) histogram generation of pixel value intensities; (c) adaptive thresholding to get initial segmentation result; (d) hole filling to obtain the lung lobe mask; and (e) corresponding segmented lung lobe region.

The optimal threshold T is calculated as the value minimizing $\sigma^2_{Intra}(k)$:

$$T = \arg\min_{k\in[1,L]} \sigma^2_{Intra}(k) \tag{3.2}$$

After thresholding, a flood filling method combined with 3D labeling is adopted to produce an initial lung lobe mask (Figure 3.2b-e). During initial segmentation testing, the LIDC CT imaging studies were found to have extremely low background pixel values (Figure 3.2a) that formed a peak pattern (Figure 3.2b) influencing the optimal threshold calculation. The calculated optimal threshold will not be able to differentiate the lung region from the background due to the influence of this peak pattern, leading to segmentation failure. Therefore, background pixel values are removed before calculating the intra-class variance.

3.3.2 Inflection point detection

Preprocessing generates a binary mask of the lung lobe region. To selectively revise the initial lung segmentation to re-include juxtapleural nodules, the boundary is first characterized

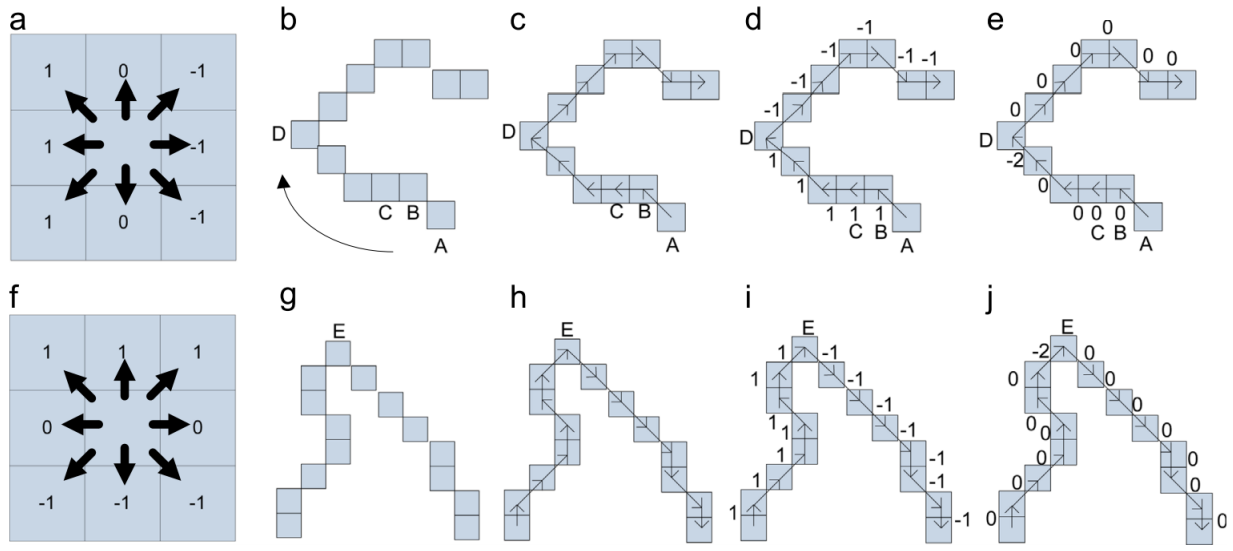


Figure 3.3: Process of encoding the bi-directional differential chain code. (a)-(e) illustrates the process of horizontal differential chain code generation, while (f)-(j) illustrates the process of vertical chain code generation. (a) Horizontal encoding coordinate system; (b) initial boundary generation; (c) arrow map generation; (d) horizontal code word assignment; (e) horizontal differential chain code generation to detect horizontal inflection points; (f) vertical encoding coordinate system; (g) initial boundary generation; (h) arrow map generation; (i) vertical code word assignment; and (j) vertical differential chain code generation to detect vertical inflection points.

using a bidirectional differential chain (BDC) encoding method to help identify inflection points. Inflection points are defined as the points where the convexity of the boundary changes. Concavities are then detected based on these inflection points. This step maximizes the sensitivity in detecting areas with juxtapleural nodules. A process for selecting critical point pairs is then followed to reduce the false positives and minimize over-segmentation. The original application of chain codes was for lossless compression of grey-scale images [GW07]. The basic principle is to separately encode the boundary coordinates (chains of pixels) for each connected component in an image. The chain is a sequence of direction codes from one pixel to the adjacent one. There are eight possible directions between two adjacent pixels. The code word $c(i)$ for a BDC is the number corresponding to the direction from one

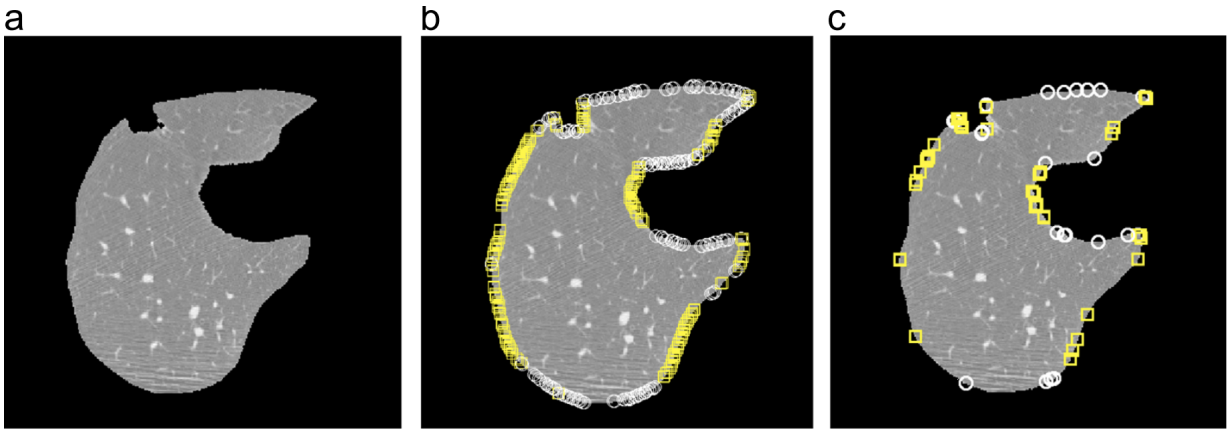


Figure 3.4: Example of detected inflection points with and without the application of a low-pass filter. (a) Right lung lobe mask; (b) detected inflection points without applying low-pass filter, the white circles represent the vertical inflection points and the yellow squares represent the horizontal inflection points; and (c) detected inflection points after applying Gaussian low-pass filter.

pixel (i) to the next $(i+1)$ in a chain, $c(i) \in A\{0, 1, 1\}$, where i represents the index value for the pixel. The assigned code word for each direction is based on the encoding coordinate system (Figure. 3.3). To detect both horizontal and vertical inflection points, this method uses two different coordinate systems for horizontal and vertical encoding. The detection of the inflection points from the BDC encoding proceeds based on the following steps:

1. **Initial boundary generation**. The lung lobe boundary pixels are extracted from the binary mask for the left and right lobes, separately.

2. **Boundary encoding**. Per lobe, both vertical and horizontal code words are obtained using the corresponding encoding coordinate systems. The encoder moves along the boundary following a (counter)clockwise path, and at each step the direction of this movement is transformed into a code word. The encoding process is illustrated in Figure. 3.3:

   (a) **Horizontal code word generation**. Figure 3.3b-d depicts the process of generating horizontal code words. In Figure 3.3b, the blue boxes represent pixels along

the lung lobe boundary. If A is the starting point, the encoder moves along the boundary in a clockwise way.

(b) **Arrow map generation**. An arrow map is generated to represent the direction that the encoder moves. For instance, in Figure 3.3c the encoder moves in the northwest direction when traversing from point A to B.

(c) **Code word assignment**. A code word is assigned to each arrow according to the encoding coordinate system given in Figure 3.3a. As shown in Figure 3.3d, the arrow that points in the northwesterly direction from A to B is assigned a code word of '1' based on the encoding coordinate system.

(d) **Vertical code word generation**. The vertical code word is generated in a similar manner, but using a vertical encoding coordinate system (Figure 3.3f). Figure 3.3g-i depicts the process of generating vertical code words.

3. **Inflection point calculation**. A differential operation is used to generate the horizontal and vertical differential chain codes, separately. Non-zero points in the differential chain are identified as inflection points. As presented in Figure 3.3e and j, the differential code is calculated using a clockwise differential operation based on the generated code words (i.e., from Steps 2a, 2d). For instance, the differential code at point A is 0; the differential code at D is 2. As can be seen, pixels D and E are the only points with nonzero differential codes; therefore, D is detected as a horizontal inflection point and E is detected as a vertical inflection point.

To overcome the influence of the small perturbations in the lobe boundary, a seven-tap Gaussian low-pass filter [GW07] is applied to smooth code words prior to the inflection point calculations in Step 3. An operator is then applied to round the smoothed code word to the nearest integer (0, 1, or 1). Figure 3.4 gives an example of the detected inflection points for a right lung lobe using the proposed method with and without the low-pass filter. Figure 3.4b shows the detected inflection points on the right lung lobe boundary without applying a low-pass filter. The white circles on the boundary represent the vertical inflection points, and the yellow squares represent the horizontal inflection points. Figure 3.4b has many

more inflection points that add unnecessary noise to the inflection detection process. After applying the Gaussian low-pass filter (Figure 3.4c), only the remaining points are deemed inflection points.

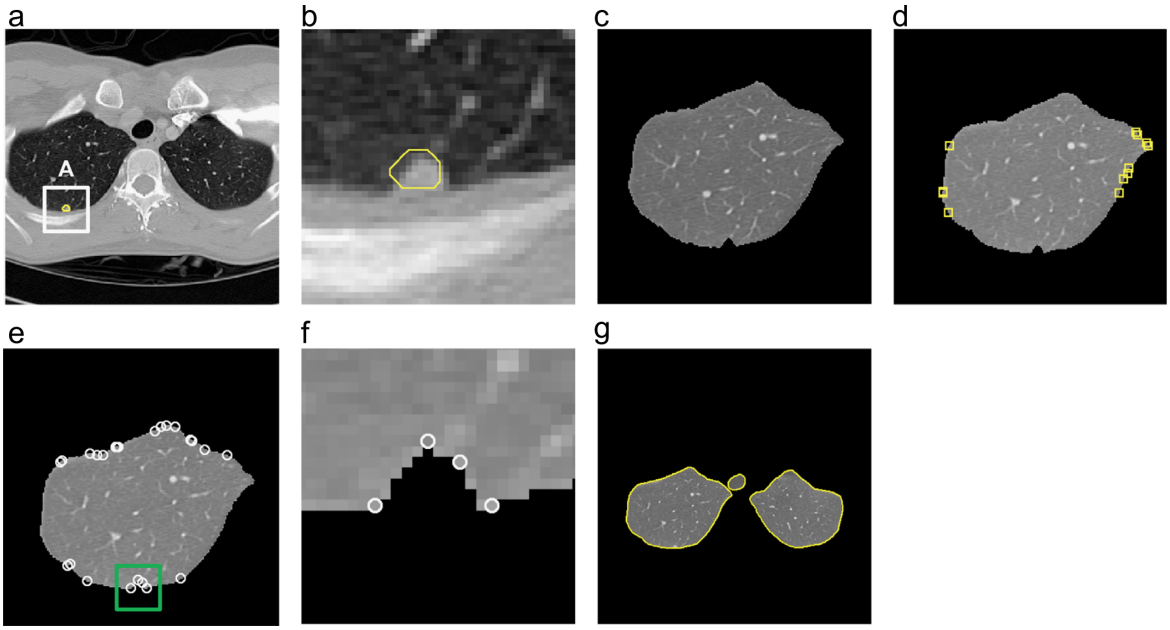


Figure 3.5: Representative results of inflection point detection. (a) Original CT slice with nodule outlines annotated by radiologists shown in yellow circle; (b) magnified view of nodule region in outlined in (a); (c) right lung mask segmented by preprocessing step; (d) detected horizontal inflection points; (e) detected vertical inflection points; (f) magnified view of vertical inflection points; and (g) lung segmentation after applying border correction.

Figure 3.5 illustrates the effectiveness of the vertical and horizontal inflection point detection process. Figure 3.5a shows an original CT slice, and the yellow contour in Figure 3.5a, b indicates the nodule in the right lung that has been manually outlined by one radiologist. After preprocessing, the segmented right lung lobe ROI does not include this nodule region; Figure 3.5c illustrates the under-segmentation problem, from which it can be observed that the convexity changes in the nodule region along the boundary are vertical. Using our approach, the detected critical horizontal and vertical inflection points are shown in Figure 3.5d-f using yellow squares and white circles, respectively. By comparing Figure 3.5d-f, the nodule region is only captured by vertical inflection points, corresponding to the

earlier observation. This observation suggests that detecting both vertical and horizontal changes simultaneously are necessary to robustly correct for undersegmentation. The final segmentation result (after border correction described in the next section) is shown in Figure 3.5g.

3.3.3 Border correction

Rather than connect all inflection point pairs, only critical point pairs are connected to correct the boundary, thereby minimizing over-segmentation. Three features are used to select critical point pairs: boundary segment concave degree, relative boundary distance, and relative position information. Let $EuclideanDistance(A, B)$ represent the Euclidean distance between two inflection points, A and B. Let $SegmentLength(A, B)$ represent the shortest boundary segment length between these two points. As shown in Figure 3.6a, ED represents $EuclideanDistance(A, B)$ and SL the $Segmentlength(A, B)$. Let $BoundaryLength$ be the total length of the lung lobe under consideration. The concave feature, $f_{concave}$, and the length feature, f_{length}, are defined as

$$f_{concave} = \frac{Segmentlength(A, B)}{EuclideanDistance(A, B)} \tag{3.3}$$

$$f_{length} = \frac{Segmentlength(A, B)}{BoundaryLength} \tag{3.4}$$

From Eq.(3.3), $f_{concave}$ increases as the boundary segment increases given a fixed geometric distance, which indicates a larger degree of concavity (Figure 3.6a). This observation implies that critical points will have larger values of $f_{concave}$. In Eq.(3.4), f_{length} increases as the boundary segment increases for a given lung lobe (with fixed total boundary length). The ratio (i.e., f_{length}) should be smaller to avoid over-segmentation. By way of illustration, a large f_{length} is shown in Figure 3.6b, where connecting the two points will cause significant over-segmentation. A third feature, $f_{position}$, indicates the relative position information of the point pair, and is defined as

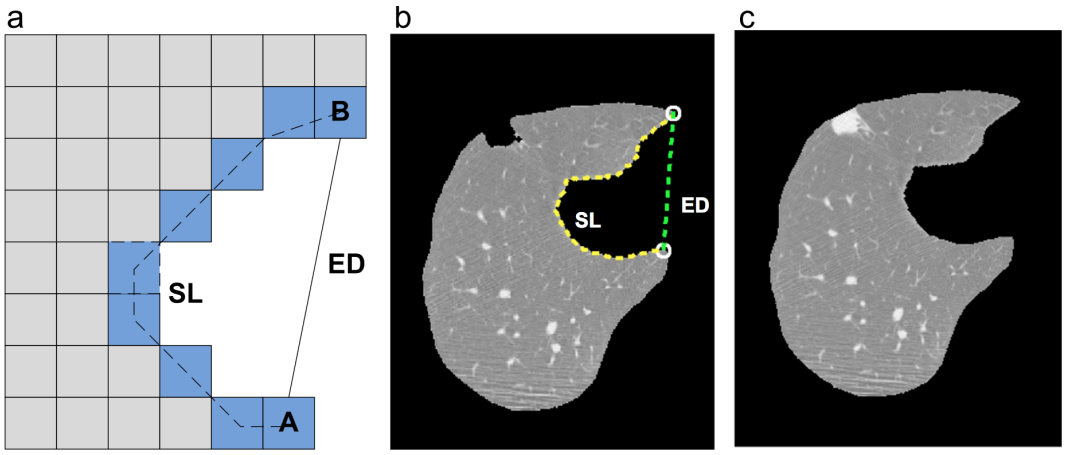


Figure 3.6: Illustration of feature definition for border correction. (a) Illustration of Euclidean distance (ED) and shortest boundary segment length (SL) between points A and B; (b) two infection points (white circles) having a large f_{length}; and (c) border correction result.

$$f_{position} = \frac{EuclideanDistance(MidPoint(A,B), CentralPoint)}{AverageDistance2CentralPoint} \tag{3.5}$$

where $MidPoint(A,B)$ is the midpoint between two inflection points, A and B; and $CentralPoint$ is the center of two lung lobe regions. $AverageDistance2CentralPoint$ is the average of all distances from lung lobe boundaries to the center.

Based on these three features, an SVM classifier is used to identify critical point pairs (instead of a threshold value or parameter). SVMs are supervised machine learning models that perform efficient non-linear classification tasks. SVMs map their inputs into higher dimension feature space to separate categories based on decision boundaries learned through training data. To train the SVM, 172 point pairs were manually selected from 42 LIDC studies, labeled as being positive examples ($n = 91$, point pairs that capture a concave region of juxtapleural nodule) or negative examples ($n = 81$, point pairs that capture a non-lung-tissue region). Finally, point pairs classified as critical are connected, resulting in the lung boundary. In our experiment, a third order polynomial kernel is chosen for the SVM classifier and a 10-fold cross validation is applied to assess the model performance using different subsets of features. Cross validation indicates that the highest accuracy (97.7%)

is achieved when all three features are employed in the training task. It is also shown that $f_{concave}$ and f_{length} are more important compared to $f_{position}$. This last observation results from the fact that $f_{concave}$ provides important information that increases sensitivity as the inflection point pairs of a juxtapleural nodule usually have larger $f_{concave}$ value. f_{length} helps to reduce false positive rate due to the limited possible size of juxtapleural nodules. To generalize for different datasets, training samples should be collected to retrain the classifier. Positive training data should be collected mainly for the juxtapleural nodules and a small portion of vessels that attach to the lung wall. Negative training data should comprise non-lung-tissue regions with a large concave rate and a moderate f_{length} value (as these point pairs can easily become false positives).

3.4 Evaluation and Results

3.4.1 Evaluation dataset

275 studies with at least one juxtapleural nodule were identified from LIDC by a trained graduate student. The entire set of 275 CT scans were divided into two subsets: 42 studies for training of the SVM; and 233 studies for testing. A total of 406 juxtapleural nodules were found in the test set, serving as the basis for evaluating the method's ability to correctly include juxtapleural nodules in the lung lobe region (i.e., re-inclusion rate [PRC08]). Additionally, 10 CT studies were randomly selected from the test set and the lung contours were manually segmented under the guidance of a practicing thoracic radiologist. The results of the manually segmented contours were used as references to validate overall segmentation accuracy. To aid in the manual segmentation task, an annotation tool was developed to enable the radiologist to automatically generate lung lobe contours first by thresholding, and then correcting any inaccuracies in the contour by adjusting the boundary.

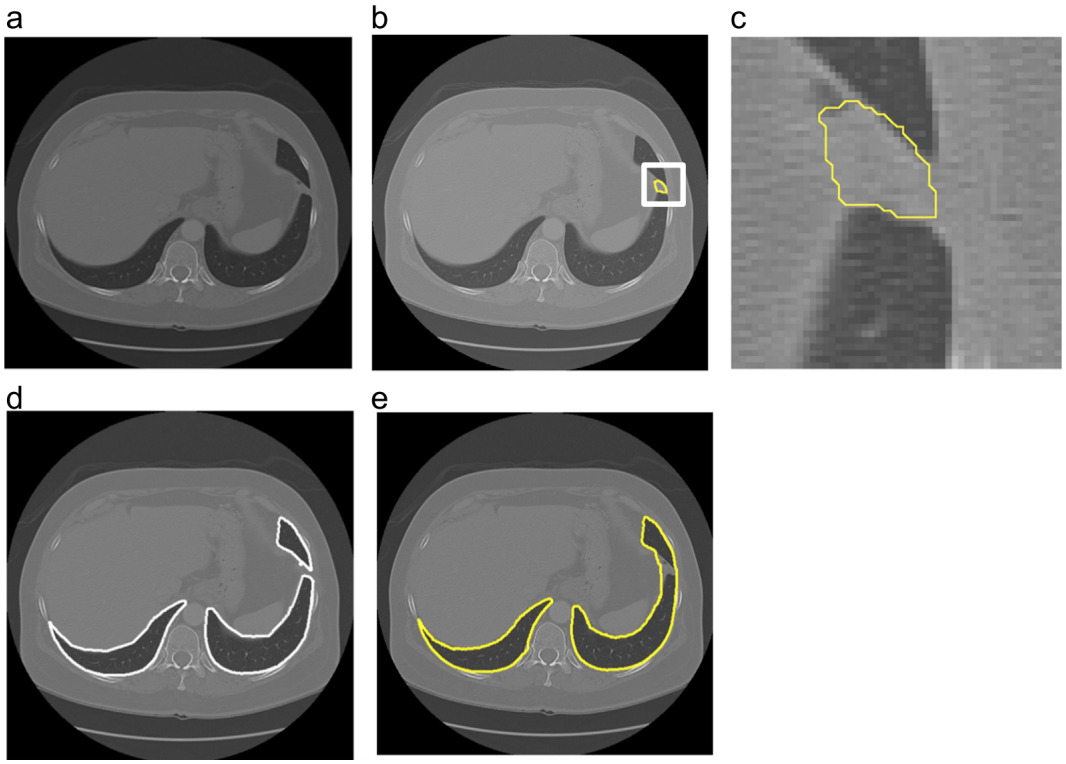


Figure 3.7: Representative case where the proposed method failed to re-include the juxtapleural nodule. (a) Original CT slice with a nodule attached to diaphragm and pleura; (b) CT slice with nodule outlines annotated by a radiologist shown in yellow circle; (c) magnified view of nodule outline annotation; (d) lung segmentation obtained by our method; and (e) reference standard lung segmentation.

3.4.2 Evaluation method

Five metrics were used to measure segmentation performance: 1) the re-inclusion ratio of juxtapleural nodules; 2) the over-segmentation rate; 3) the under-segmentation rate; 4) the volumetric overlap error ratio; and 5) the cumulative error distance distribution [DMM06, PRC08]. The re-inclusion ratio is used to assess per nodule sensitivity. Similar to [PRC08], a trained graduate student was tasked with reviewing each study to determine if a juxtapleural nodule was correctly included (or not) by identifying errors in the segmentation caused by juxtapleural nodules. For voxel-based segmentation accuracy, the volumetric overlap ratio difference, over-segmentation, and under-segmentation rates were computed to

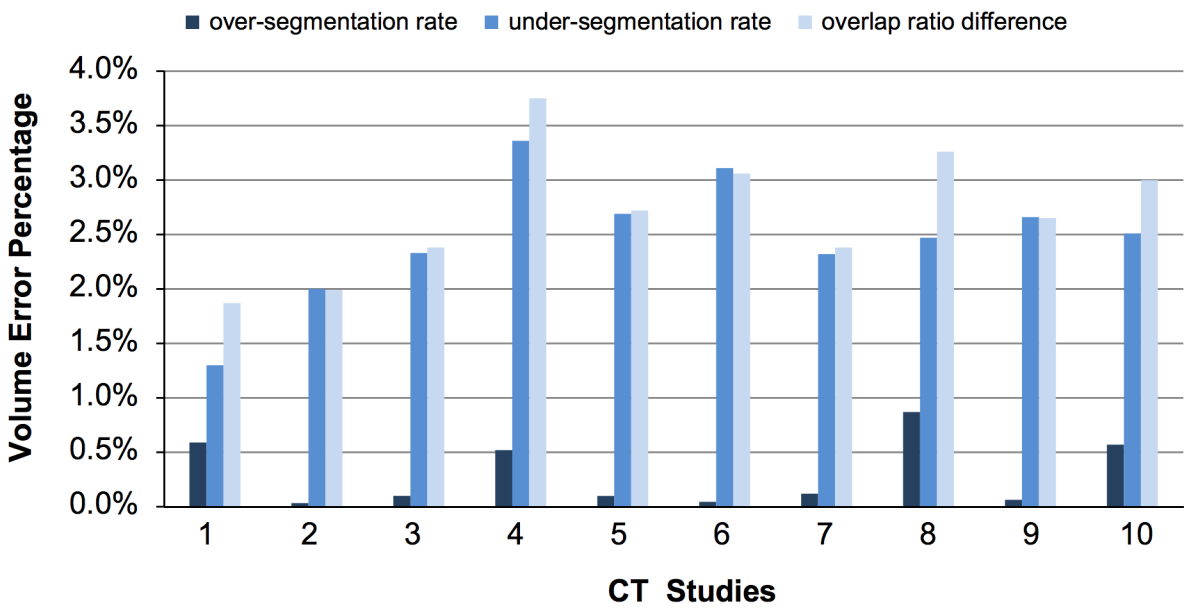


Figure 3.8: The segmentation error computed based on a comparison of lung volume: over-segmentation rate, under-segmentation rate and overlap ratio difference from Eqs. (3.6), (3.7) and (3.9). Mean errors are 0.3%, 2.4% and 2.7% respectively.

characterize differences between boundaries generated by the proposed approach and the reference boundary generated by the manual annotator.

Over-segmentation rate is defined as the number of voxels in a segmented image region that are included as part of the ROI but that are not in the reference standard [PRC08]. Let V_{auto} represent the volume of the binary mask generated using our approach and let $V_{reference}$ be the volume of the reference standard. The oversegmentation rate $OR(V_{auto}, V_{reference})$ is

$$OR(V_{auto}, V_{reference}) = \left| \frac{V_{auto} \backslash V_{manual}}{V_{manual}} \right| \tag{3.6}$$

where $V_{auto} \backslash V_{manual}$ represents the relative complement of V_{auto} in V_{manual}. Similarly, the under-segmentation rate $UR(V_{auto}, V_{reference})$ is defined as the relative lung volume amount that is regarded as lung tissue in the reference standard but not in our method:

$$UR(V_{auto}, V_{reference}) = \left| \frac{V_{manual} \backslash V_{auto}}{V_{manual}} \right| \tag{3.7}$$

The volumetric overlap ratio measures the relative overlap between the two binary segmentation masks by computing the two volumes' intersection divided by their union [SCL16]:

$$R(V_{auto}, V_{reference}) = \left| \frac{V_{manual} \cap V_{auto}}{V_{manual} \cup V_{auto}} \right| \tag{3.8}$$

Lastly, to make the overlap ratio measurement consistent with the over-segmentation and under-segmentation rates, the volumetric overlap error ratio ($DR(V_{auto}, V_{manual})$) is used:

$$DR(V_{auto}, V_{reference}) = 1 - R(V_{auto}, V_{reference}) \tag{3.9}$$

To measure the spatial similarity between the lung boundaries generated by our approach and that of the reference standard, the cumulative error distance distribution [PRC08] is computed to provide a global statistical measurement of the fitting between the lung surfaces generated by our method and the lung surfaces in the reference standard. The shortest distance between a point on the lung surface obtained by our algorithm and the lung surface of the reference standard is used to generate the error distance distribution.

3.4.3 Results

Using the 233 test studies from the LIDC dataset, our experiment shows that 373 out of total 406 juxtapleural nodules were correctly included as part of the ROI, achieving a 92.6% inclusion rate. After an error analysis, 83.3% of the missing juxtapleural nodules were found sitting in between segments of lung tissues, as shown in Figure 3.7. In this situation, the proposed method fails because each segment is processed separately.

Figure 3.8 shows the volume-based segmentation error as assessed by over-segmentation ratio, under-segmentation ratio, and overlap ratio difference. The average over-segmentation rate is 0.3%, while the average under-segmentation ratio is 2.4% and the average overlap ratio difference is 2.7%. Figure 3.9 shows the cumulative error distance distribution to assess the border positioning accuracy. The error bars in Figure 3.9 represent the standard deviation corresponding to each distance. 93% and 96% lung surfaces obtained by the proposed method

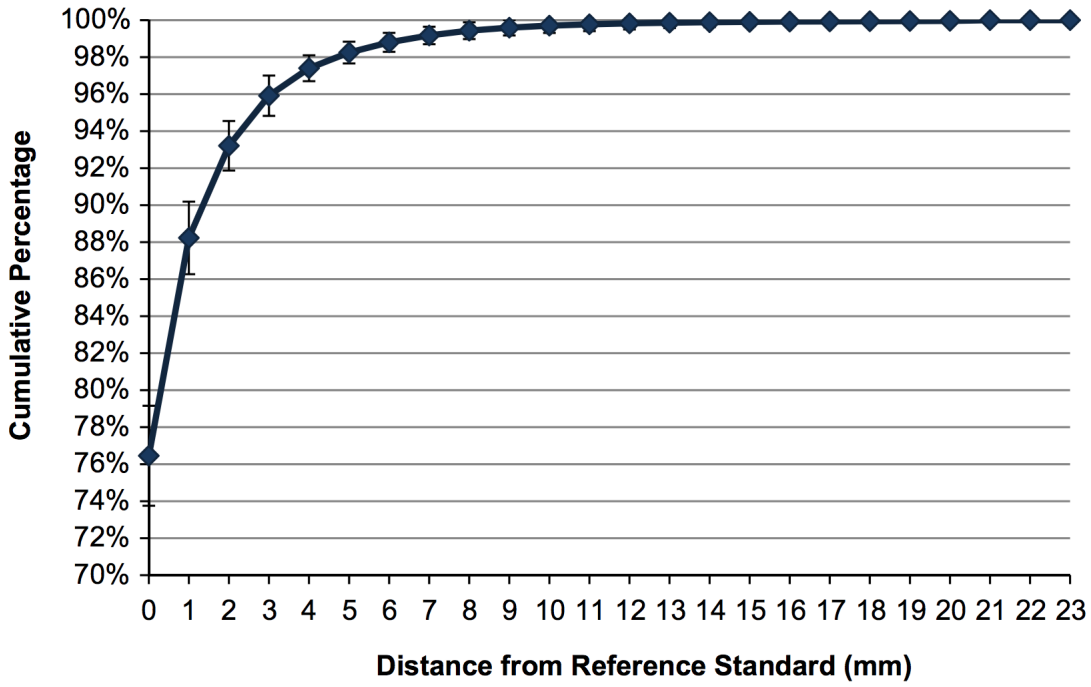


Figure 3.9: Cumulative point-wise error distance distribution of the shortest distance from proposed lung segmentation surface to lung surface of the reference standard.

are within 2-3 mm of the reference standard, respectively. The largest error distance is 22.5 mm. Relatively larger under-segmentation and error distances are mainly due to the presence of atelectasis (Figure 3.10a, b) or consolidation (Figure 3.10c, d).

Although many methods have been developed to perform automatic lung segmentation, only a few explicitly handle juxtapleural nodules and evaluate the method on actual patient data. Table 3.1 compares the proposed method to other lung segmentation algorithms that handle juxtapleural nodules. Our method achieves better average overlap ratio compared to Wei's [WSL13] method. This difference is attributed to the fact that our approach implements a point pairs selection technique, which reduces the risk of over-segmentation. Our method has similar average oversegmentation ratio and average under-segmentation ratio compared to Pu's [PRC08] method. Although both Pu [PRC08] and Wei [WSL13] report a 100% re-inclusion rate, their test sets contain a limited set of juxtapleural nodules, 67 and 32 respectively, compared to our set of 406 nodules. Similar to the limitation of our approach in detecting missing nodules that are between lung segments, Pu's and Wei's methods also

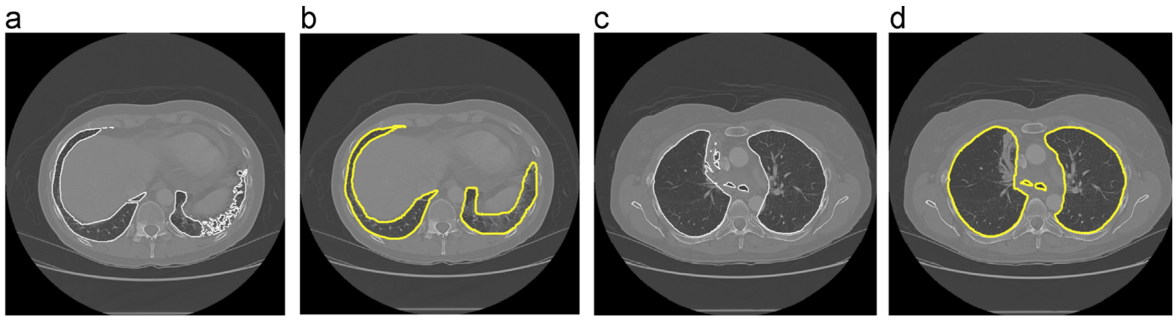


Figure 3.10: Comparison between lung segmentation obtained by our method and reference standards in cases with atelectasis or consolidation. (a) Lung segmentation obtained by our method in an atelectasis case; (b) reference standard in an atelectasis case; (c) lung segmentation obtained by our method in a consolidation case; (d) reference standard in a consolidation case.

process each isolated lung region separately and thus will likely fail in similar situations. One should note that the performance of the oft-cited rolling ball method is highly dependent on the specified parameters, which are not consistently included in publications [PRC08]. For example, Kim [KKN03] highlights the difficulty in selecting the appropriate fixed ball radius because of the large variance in juxtapleural nodule sizes. Therefore, a comparison of our method against algorithms that utilize the rolling-ball method would be inconclusive, given that the original implementation cannot be replicated effectively. Also, Stelmo et al. [NSN12] point out that a fair comparison between two methods would only be possible if the works use the same images and acquisition standards (resolution, bits per pixel, etc.); given the unavailability of others' test data, such a formal comparison is not possible. But in general, and as compared to these other methods, our described method can accurately segment the lung tissues while robustly and correctly including the juxtapleural nodules.

3.5 Discussion

A seven-tap Gaussian low-pass filter is employed to smooth the chain code to overcome the influence of the small perturbations in the lobe boundary as described in Section 3.2. This

Table 3.1: Comparison of the performance of lung segmentation methods that handling juxtapleural nodules.

Works	Pu et al. [PRC08]	Wei et al. [WSL13]	Proposed Method
Test dataset	20 scans; 67 Juxtapleural nodules	25 scans; 32 Juxtapleural nodules	233 scans; 406 Juxtapleural nodules
Average R	-	95.24%	97.3%
Average OR	0.43%	-	0.3%
Average UR	1.63%	-	2.4%
Re-inclusion Rate	100%	100%	92.6%

step is used to reduce computational complexity as fewer inflection point pairs are inputted to the SVM classifier afterwards. Removing this step will not reduce the re-inclusion rate of the juxtapleural nodules as the classifier would still identify the same points for re-inclusion but would need to consider more points. In our experiment, a Gaussian kernel with variance equal to two is employed to perform the smoothing task and a 92.6% re-inclusion rate is achieved. By examining the failure cases in our experiment, we found that none of the failure cases were caused due to inflection point detection problems. As such, we believe the adoption of this Gaussian low-pass filter did not include additional errors. However, to adapt this approach to perform on other datasets, the degree of smoothing strength may be adjusted based on preference and dataset standards.

One limitation of the proposed method is that it sometimes fails to re-include the juxtapleural nodules sitting in consolidation regions (between lung tissue segments); to overcome this problem, future work could integrate some region connection techniques as a precursor step for detected consolidation regions before border correction. With the advancement of deep learning methods, future work could also develop deep learning models to for better segmentation results, and it will be discussed in Chapter 6.

CHAPTER 4

Lung Nodule Classification and Diagnosis using Deep Convolutional Neural Network

4.1 Overview

Researchers have actively pursued methods to assist radiologists in the interpretation process to automatically classify lung nodules and diagnose lung cancer. Maintaining high sensitivity while minimizing false positives remains a challenging problem. Moreover, existing methods show a significant decrease in performance on datasets with different acquisition parameters and/or patient cohorts (compared to training data). Thus, models with better transferability are desirable and needed for practical clinical usage. Deep learning has been shown to be able to learn latent multi-level hierarchical representations adaptively from raw data, and have significantly improved the state-of-the-art performance in visual object detection [KSH12, KTS14] and a range of medical related tasks [CQY16, GPC16, EKN17]. Existing approaches do not incorporate any lung cancer domain knowledge and use the CNN as a "black-box," thus having little model interpretability and hindering it from being understood and ultimately adopted by radiologists. In contrast, radiologist-quantified image (semantic) features, such as calcification and sphericity, have been widely used by radiologists to assist the lung cancer diagnose procedure [INK05, RVF09, HM16]. These semantic features represented the domain knowledge that have been widely recognized by radiologists for lung cancer diagnosis. Many these features are currently defined only qualitatively, and are difficult to quantify from first principles [HM16].

In this chapter, I first present a robust model to automatically classify lung nodules vs. non-nodules using a hybrid ensemble of multiple deep convolutional neural networks (CNNs).

Table 4.1: Summary of LIDC and UCLA datasets.

Dataset	Type	Number of scans	Acquisition	Nodule size
LIDC	Retrospective	1018	Low dose and diagnostic dose	3 30 mm
UCLA	Retrospective	158	Diagnostic dose	5 30 mm

This model was first built based on a large, publicly available dataset, and then externally validated on an independent dataset without retraining the model. Second, I describe a novel interpretable deep hierarchical semantic convolutional neural network (HSCNN) for lung cancer prediction with two levels of output: 1) low-level radiologist semantic features; and 2) a high-level malignancy prediction score. The low-level semantic outputs quantify the diagnostic features used by radiologists and serve to explain how the model interprets the images in an expert-driven manner. In the remainder of this chapter, Section 4.2 details the developed ensemble CNN for lung nodule detection and the HSCNN model for lung cancer diagnosis. Section 4.3 summarizes the experimental results for both works. Section 4.4 discusses the limitations and future works. The content presented in this chapter have partly been published in [SBH17].

4.2 Methods

4.2.1 Dataset

The LIDC dataset (prior described in Section 3.4.3) was used to develop both the hybrid ensemble CNN model for lung nodule classification and HSCNN model for lung nodule diagnosis. A UCLA dataset was used as an external dataset to validate the ensemble nodule classification model. Table 4.1 summarizes the information for these two datasets.

In the LIDC dataset, lung nodules were manually annotated by four radiologists following a two-phase image annotation process. Pixel-level 3D contour segmentations, panel opinions on the assessment of nodule likelihood for malignancy, and eight nodule characteristics were generated for lesions categorized as nodules $\geq$ 3 mm. The eight nodule characteristics were

semantic diagnostic features, including: calcification, subtlety, lobulation, sphericity, internal structure, margin, texture, spiculation, and malignancy. Each feature was rated from 1 to 5 or 6 by radiologists. Table 4.2 lists the description and definitions for each of the labels from [AMB11, CVS13].

Table 4.2: Nodule characteristics labels in LIDC dataset.

Semantic Feature	**Description**	**Ratings**
Malignancy	Likelihood of malignancy	1. Highly unlikely 2. Moderately unlikely 3. Indeterminate 4. Moderately suspicious 5. Highly suspicious
Margin	How well defined the margins are	1. Poorly defined 2. 3. 4. 5. Sharp
Sphericity	Dimensional shape in terms of roundness	1. Linear 2. 3. Ovoid 4. 5. Round
Subtlety	Contrast between nodule and surroundings	1. Extremely subtle 2. Moderately subtle 3. Fairly subtle 4. Moderately obvious 5. Obvious

Semantic Feature	**Description**	**Ratings**
Spiculation	Degree of exhibition of spicules	1. Marked 2. 3. 4. 5. None
Texture	Internal density of nodule	1. Non-solid 2. 3. Part Solid 4. 5. Solid
Calcification	Calcification appearance in the nodule	1. Popcorn 2. Laminated 3. Solid 4. Non-central 5. Central 6. Absent
Internal structure	Expected internal composition of the nodule	1. Soft tissue 2. Fluid 3. Fat 4. 5. Air
Lobulation	Whether lobular shape is apparent from margin or not	1. Marked 2. 3. 4. 5. None

4.2.2 A hybrid ensemble CNN model for lung nodule classification

4.2.2.1 Data preprocessing

In this study, only scans with slice thickness smaller than 3 mm were included, resulting in 897 LIDC CT scans. Four radiologists annotated the nodule contour and characteristics for each CT scan in LIDC, with no enforced agreement on the existence of the nodule. Thus, each nodule may have been assigned 1-4 annotations, based on the level of agreement. Each annotations was considered as a distinct nodule (e.g., one object might be marked by all four radiologist as a nodule, resulting in four annotations), resulting in selecting 4,252 nodules from LIDC dataset. 158 lung nodules were annotated by one radiologist in UCLA dataset, and are used in this study. Each nodule was a positive sample for lung nodule classification. The negative (non-nodule) samples were extracted from the methods described in [MHR10, DBS15, SBC15]. Multi-level thresholding and morphological operations were used to detect a large set of candidates. Rule-based analysis were performed to remove extreme small or large candidates outside the target detection size range (5 mm to 30 mm). Any candidates overlapped with nodules were excluded from the non-nodule candidates. This nodule and non-nodule candidates settings for classification followed the conventions in previous work [FCS17].

The LIDC-IDRI and UCLA dataset contained a heterogeneous set of scans with various acquisition parameters. To normalize the pixel values, all CT scans were first transformed to Hounsfield (HU) scales using the information in DICOM header and then converted to the range of $(0, 1)$ from (-1000, 500 HU). A 3D cube of $40 \times 40 \times 40$ mm were extracted for each candidate. Each cube is centered around the candidate. 40 mm was chosen so that all candidates will fit in this range as the largest nodules in my subset were selected to be around 30 mm. I then rescaled each cube to a fixed size of pixels in all three dimensions, resulting in isotropic cubes for all cases. The center slices of this cube in the axial, sagittal, and coronal views were then aggregated as a 2.5D input for the network.

4.2.2.2 Hybrid ensemble CNN model

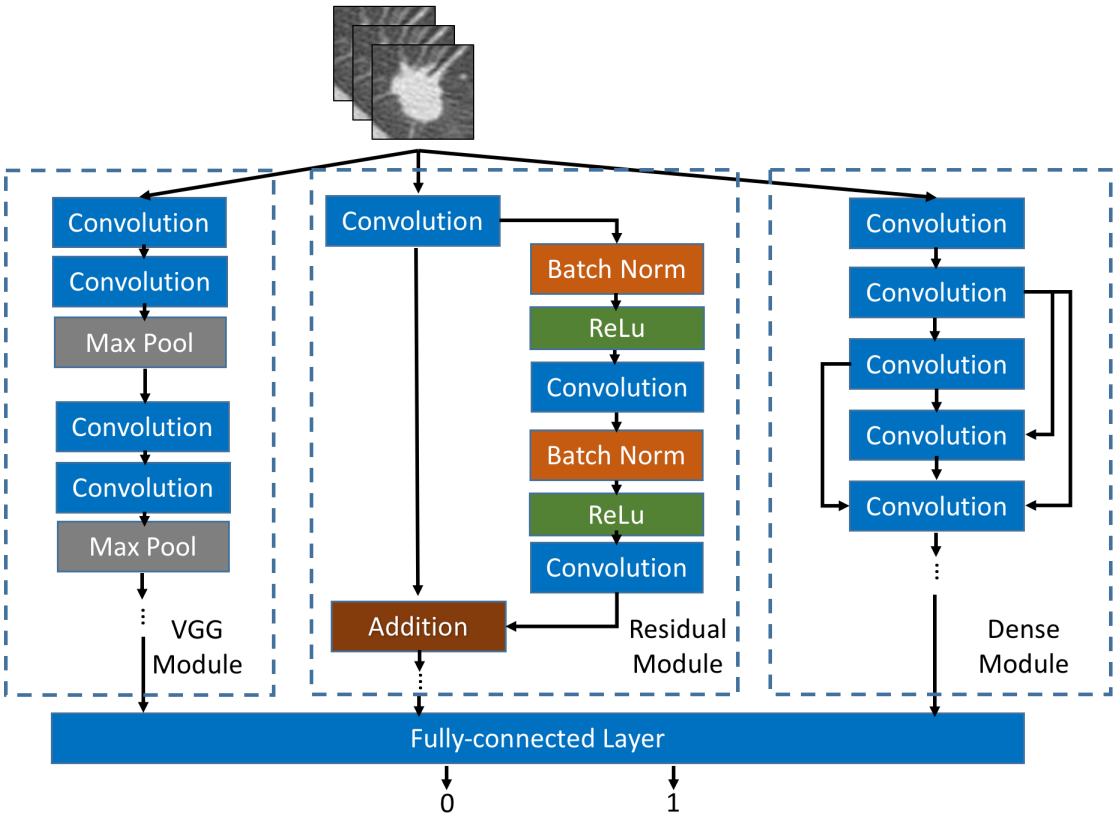


Figure 4.1: Framework of hybrid ensemble CNN model for lung nodule classification.

The CNN is constructed to perform the classification task to differentiate nodule versus non-nodule objects. I designed a hybrid ensemble CNN structure consisting of a VGG module [SZ14], residual module [HZR16a, HZR16b], and dense module [HLW16], as depicted in Figure 4.1.

2D convolutional layer and 2D pooling layer are the basic building blocks of the proposed model. Convolution layer is consisted of multiple kernels, and each kernel contains unique set of learnable parameters producing one 2D output matrix. These layers perform the convolution operation on input feature maps along both dimensions of the input 2D matrix to produce an output feature map defined by:

$$f^j = \sum_i c^j * f^i + b^j \tag{4.1}$$

where f^j and f^i are the j^{th} output feature map and i^{th} input feature map, respectively. And c^j is the j^{th} convolution kernel and $*$ represents the 2D convolution operation between the convolution kernel and input feature map. b^j is the j^{th} bias corresponding to the j^{th} convolution kernel.

The max pooling layer outputs the max value from the input window. It is used to progressively reduce the spatial size of the feature maps to reduce the number of parameters and computation for the purpose of control overfitting as:

$$\begin{aligned} \hat{f}^i_{x,y} = max\{f^i_{x',y'}; x' \in [x \cdot s_x,\ x \cdot s_x + d_x - 1], \\ y' \in [y \cdot s_y,\ y \cdot s_y + d_y - 1]\} \end{aligned} \tag{4.2}$$

where x (the row index) and y (the column index) start from zero. Here, s is the stride size (downscale factor) and d is the size of the max pooling window. In this study, I employ a pooling window size of $d = (2, 2)$ and stride of $s = (2, 2)$. This design determines that the output size will be reduced by half in width and height dimension. This pooling layer has no learnable parameters.

One convolution (conv) unit in the VGG module (Figure 4.1, VGG module) comprises two 3×3 convolution layers and one 2×2 max pooling layer. Stacking two 3×3 convolution layers achieved an effective receptive field of 5×5, significantly reducing the number of model parameters while increasing the discriminative power of the decision function by incorporating two activation function layers (versus one). Three convolution modules here were used in total. The two convolution layers in the first, second and third module have 32, 64 and 128 kernels, respectively. The output from the third convolution module is then fed into a fully connected layer with 128 neurons. These neurons were connected to the last layer with two output units.

The residual module (Figure 4.1, residual module) enforces learning of residual functions in relation to the layer inputs. The residual unit (RU) is the basic building block for the residual module and consists of two paths: 1) a direct path from the input, and 2) a batch normalization and convolution path. The output from these two paths are combined using

Table 4.3: Summary of generating binary labels from LIDC rating scales for nodule characteristics.

Nodule characteristics	**Label 0**	**Label 1**
Malignancy	Scale 1 - 3 Benign	Scale 4 - 5 Malignant
Sphericity	Scale 1 - 3 Lesser roundness	Scale 4 - 5 High degree of roundness
Margin	Scale 1 - 3 Poorly defined margin	Scale 4 - 5 Sharp margin
Subtlety	Scale 1 - 3 Poor contrast between nodule and surroundings	Scale 4 - 5 High contrast between nodule and surroundings
Texture	Scale 1 - 3 Non-solid internal density	Scale 4 - 5 Solid internal density
Calcification	Scale 1 - 5 Present of calcification	Scale 6 Absent of calcification

an addition function. Batch normalization is applied to output feature maps to accelerate the training process and reduce the internal covariate shift by normalizing the feature maps [IS15]. A 49 layer residual structure is used as in [HZR16b]. The dense module (Figure 4.1, dense module) is made up of three densely connected convolution blocks as in [HLW16]. Each block is made up of 12 batch normalization + convolution layer combinations. And all convolution layers inside a dense block are connected to every convolution layers afterward in the same block. Rectified linear units (ReLUs) [KSH12] are used as the nonlinear activation functions in all three modules. Softmax function is used as the loss function.

4.2.3 A HSCNN model for lung cancer diagnosis

4.2.3.1 Data preprocessing

Similar to the lung nodule detection task, only lung nodules were considered whose largest diameters were between 3 and 30 mm to develop lung cancer diagnosis model. As described previously, four different radiologists annotated the nodule contour and characteristics for each CT scan in LIDC. To determine which annotations refer to the same nodule, an annotation list provided in [te11] has been used. Only nodules identified by at least three radiologists in the CT scans with slice thickness smaller than 3 mm were included. This resulted in sub-selecting 4,252 nodule annotations, and I employed each annotation as a distinct nodule. One reason for this choice was that the physical nodule regions lack universal agreement, and arbitrary choices have to be made to determine the nodule contour based on the different radiologists' delineations. Another reason was to use all of the specialist expertise, following the conventions used by others [CVS13, HM16, FCS17]. Uniform labels were assigned to all annotations referring to the same nodule for each feature. As indicated in Table 4.2, LIDC employed ordinal scales from 1 to 5 to label nodule malignancy and four other semantic diagnostic features, including margin, sphericity, subtlety, and texture. For these five nodule characteristics, I obtained average ratings for each nodule as in [SZY15], and obtained a binary label for each by comparing the average scores with a threshold on 4 – that is, a "0" label was assigned for average scores smaller than 4, and a "1" label was assigned for average scores larger or equal to 4. Label 0 indicated a benign nodule, poorly defined margin, lesser roundness, poor contrast between nodule and surroundings, and non-solid internal density of nodule for malignancy, margin, sphericity, subtlety, and texture, respectively. Conversely, Label 1 denoted a malignant nodule, sharp margin, high degree of roundness, high contrast between nodule and surroundings, and solid internal density of nodule. LIDC used categorical scales from 1 to 6 to annotate calcification features; here, I averaged ratings for each nodule by majority vote per [CVS13]. For those with average ratings of 6, I labeled them as absent of calcification pattern (label 1); all other ratings represent the presence of calcification (label 0). Table 4.3 summarizes the generation of the

Table 4.4: Label counts for nodule characteristics.

Nodule characteristics	Label 0 (#)	Label 1 (#)	Total (#)
Malignancy	3212	1040	4252
Sphericity	2304	1948	4252
Margin	1640	2612	4252
Subtlety	1570	2682	4252
Texture	518	3734	4252
Calcification	496	3756	4252

binary labels from LIDC rating scales as described above. Table 4.4 lists the data counts for each label of the nodule characteristics.

I note here an important labeling error in regards to the scales and guidelines used for both the lobulation and spiculation features. It has been reported by The Cancer Imaging Archive (TCIA) that a subset of 100 among 399 known cases in the LIDC dataset were annotated using an inconsistent rating system for spiculation and lobulation [te17]. Unfortunately, precisely which 100 of the 399 cases is not known. Therefore, lobulation and spiculation are not used for this work. Hancock et al. [HM16] also deal with this mislabeling issue in their study. However, I find a few studies used the labeled lobulation and spiculation features in LIDC, but did not mention or handle these possible labeling error issues [CVS13, DMA13, NRF15, RVF09, OCR11]. The internal structure semantic feature is also removed for this study due to the fact that almost all nodule annotations are labeled on the same scale (4,242 out of the 4,252 annotations are labeled on scale 1; and the remaining 10 annotations were labeled on scale 4). Thus, this feature provides little information, as also suggested in [HM16].

Similar as in Section 4.2.2.1, all CT scans were first transformed to HU and then converted to a range of $(0, 1)$ from (-1000, 500 HU). A 3D cube of $40 \times 40 \times 40$ mm was extracted for each candidate. Each cube was centered around the candidate. I then rescaled each cube to a fixed size of pixels in all three dimensions. This 3D cube was the input for the proposed HSCNN method, detailed below.

4.2.3.2 HSCNN model

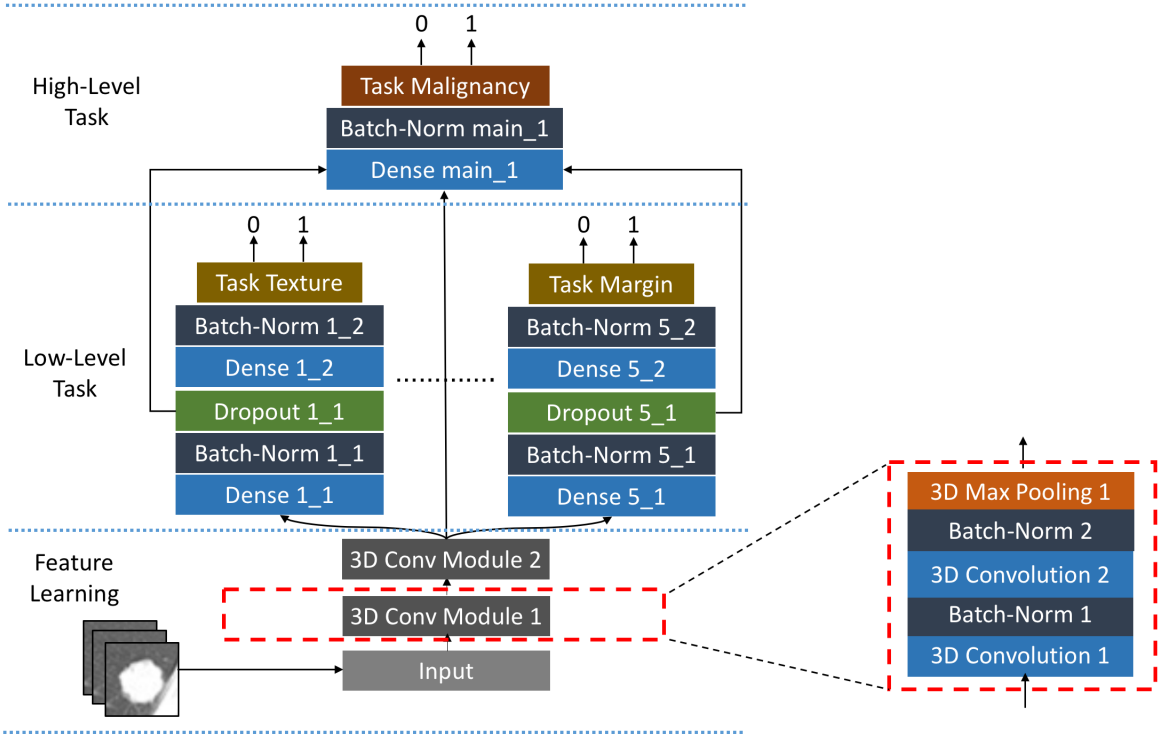


Figure 4.2: Model architecture of the hierarchical semantic convolutional neural network.

The proposed HSCNN takes the 3D lung nodule image cube as input and outputs two levels of predictions, as shown in Figure 4.2. This architecture comprises three parts: 1) a feature learning module; 2) a low-level task module; and 3) a high-level task module. The feature learning component is able to adaptively learn the image features generalizable across different tasks. The low-level task module learns to predict five explainable semantic diagnositc features, including margin, texture, sphericity, subtlety, and calcification. The high-level task component absorbs information from both the generalizable image features and the low-level tasks, producing the final prediction for lung nodule malignancy.

The feature learning component (Figure 4.2, feature learning) is the first processing unit of the proposed network. It consists of two convolution module blocks, where each block shares the same structure and contains two stacked 3D convolution layers followed by batch normalization and one 3D average pooling layer. Each convolution layer has a kernel size of $3 \times 3 \times 3$. These layers perform the convolution operation on input feature maps

along all three dimensions of the input cube. After the convolution, batch normalization is applied to all output feature maps [IS15]. ReLUs [KSH12] are used to take the output from batch normalization. 16 feature maps are used for both convolution layers in the first convolution module, and 32 feature maps are adopted for both convolution layers in the second convolution module. A 3D max pooling layer is used in the end for each convolution module block to reduce the spatial size of the feature maps. I employ a pooling window size of $d = (2, 2, 2)$ and stride size of $s = (2, 2, 2)$. This design downsamples the input feature maps by a factor of 2 across all three cube dimensions.

After the last convolutional module, output features are fed simultaneously into the low- and high-level task components. The low-level task components (Figure 4.2, low-level task) consist of five branches, each with the same architecture, addressing a distinct semantic feature task (i.e., texture, margin, sphericity, subtlety, or calcification). A fully-connected layer (densely-connected) is the major basic building block for each of these branches. One fully-connected layer connects each input unit to each output unit, designed to capture correlations from all input feature units to the output. Batch normalization and dropout techniques are both used to control model overfitting. The dropout method randomly removes connections between input and output units during network training to prevent units from co-adapting too much [SHK14]. Two fully-connected layers are employed before the final binary prediction with 256 neurons and 64 neurons for the first and second layer, respectively.

The high-level task component (Figure 4.2, high-level task) predicts the lung nodule malignancy as the final task. This module concatenates as input the output features from the feature learning component and each of the low-level task branches. As shown in Figure 4.2, the output feature maps from the last convolution module of the feature learning component is used, along with the output from the last second fully-connected layer of each subtask branch. This design makes the final prediction utilize the basic features learned from the shared convolution modules, and forces the convolution blocks to extract representations that are generalizable across tasks. It also makes use of the information learned from each related explainable subtask to ultimately infer nodule malignancy. The last fully-connected layer in each subtask branch is trained to extract representations more specific to the cor-

responding subtask compared to the second last fully-connected layer. Thus, the second last layer of the subtask branch is chosen to provide less specific but salient information for the final malignancy prediction task. The concatenated features are inputted into a fully-connected layer with 256 neurons, followed by a batch normalization operation before the final malignancy prediction.

To jointly optimize the the HSCNN during the network training, a novel global loss function is proposed to maximize the probability of predicting the correct label for each task by:

$$L_{global} = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{5}\lambda_j \cdot L_{j,i} + L_{M,i}) \tag{4.3}$$

where N is the total number of training samples and i indicates the i^{th} training sample. j is the j^{th} subtask and $j \in [1, 5]$. λ_j is the weighting hyperparameter for the j^{th} subtask. $L_{j,i}$ represents the loss for sample i and task j. $L_{M,i}$ is the loss for the malignancy prediction task for the i^{th} sample. Each loss component is defined as weighted cross entropy loss by:

$$L_{j,i} = -\log\left(e^{f_{y_i,j}} / \sum_{n} e^{f_{y_n,j}}\right) \cdot \omega_{y_i,j} \tag{4.4}$$

where y_i is true label for the i^{th} sample (x_i, y_i). Here, y_i equals 0 or 1. $f_{y_i,j}$ is the prediction score of the true class y_i for task j and $f_{y_n,j}$ represents a prediction score for class y_n. I use $\omega_{y_i,j}$ to represent the weight of class y_i for task j. The use of $\omega_{y_i,j}$ is important because the labels are unbalanced in all the tasks and $\omega_{y_i,j}$ is helpful in reducing the training bias introduced by such data imbalance. Specifically, $\omega_{y_i,j}$ weights each class loss proportional to the reciprocal of the class counts in the training data. For instance, $\omega_{y_i=0,j} = N_{y_i=1,j}/(N_{y_i=0,j} + N_{y_i=1,j})$ and $\omega_{y_i=1,j} = N_{y_i=0,j}/(N_{y_i=0,j} + N_{y_i=1,j})$. $N_{y_i=1,j}$ represents the total count of samples in the training data for task j, where the true class label equals 1. The global loss function is minimized during the training process by iteratively computing the gradient of L_{global} over the learnable parameters of HSCNN and updates the parameters through back-propagation.

To better control for model overfitting, 3D data augmentation was applied during the

training process. Data augmentation artificially inflates the dataset by using label-preserving transforms to add more invariant data examples and is considered as a model regularization scheme [KSH12]. One or more random operations are applied on each training dataset to generate artificial samples. The spatial affine operations used in this study include translating the position of the nodule within 4 mm or flipping the 3D nodule cube along one of the three axis. The translation limit is set to 4 mm to keep the boundaries of the largest nodules be captured properly in the 3D cube ($40 \times 40 \times 40$ mm).

4.3 Evaluation and Results

4.3.1 Implementation details

During training, the learnable parameters of both the hybrid ensemble CNN model and HSCNN model are initialized using the Xavier algorithm [GB10] and are updated using the Adam stochastic optimization algorithom [KB14]. To capture a majority of nodule morphology while reducing the input data dimensions, the input candidate cube size was set to be $52 \times 52 \times 52$ voxels. The hybrid CNN model uses 2.5D input with a size of $52 \times 52 \times 3$; and the HSCNN model uses the whole 3D cube as input. For both models, the learning rate was set to be 0.001; and the choices of architecture design and parameters are commonly used, as shown in [KSH12, SZ14, HLW16, HZR16a, HZR16b]. The hyperparameters were chosen by using a randomized coarse-to-fine grid search with the validation dataset in the first 20 epochs [BB12]. Both models are implemented in Python 2.7 with TensorFlow [ABC16] and the Keras toolkit [Cho15]. All experiments were performed on a server with 16 Intel Xeon E5-2630 CPU processors, 32GB memory, and one NVIDIA TITAN Xp GPU (12GB on-board memory).

4.3.2 Hybrid ensemble CNN experimental results

To train and validate the proposed model, I divided the LIDC-IDRI scans into four folds, two folds for training the model, one fold for validation and one fold for test. This scan level

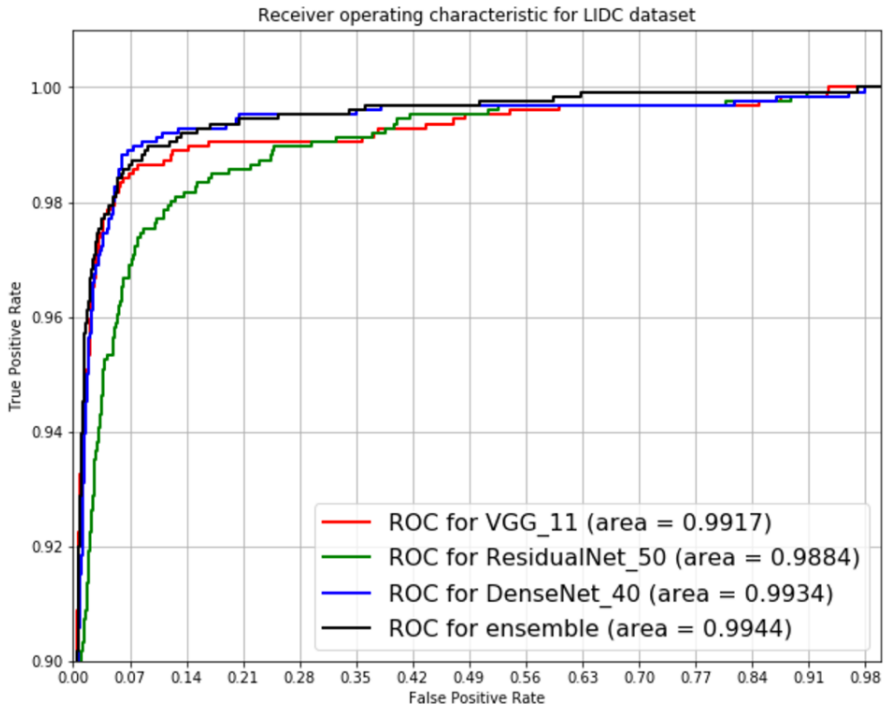


Figure 4.3: ROC plot on LIDC datasets for hybrid ensemble CNN model comparison.

splits made sure that there would be no information overlap between training/validation and test due to multiple annotations. The scans in the LIDC dataset were randomly divided into five subsets of similar size. Three subsets were used for training, one for validation, and one for testing. All five folds contain a total of 6,776 nodules. The test subset contained 207 scans with 1,262 nodules and 8,281 non-nodules. The UCLA dataset, which included 158 nodules and 3,938 non-nodules, was used as an independent dataset to validate the model performance.

Figure 4.3 and Figure 4.4 show the receiver operating characteristic (ROC) curve plots comparing the hybrid ensemble CNN model versus single VGG model, single residual model and single densely connected model on LIDC dataset and UCLA dataset, respectively. These plots represent the intuitive trade-off between sensitivity and specificity. By visual inspection of the ROC curves, the hybrid ensemble CNN model performs better than all the other three models on both the LIDC test set and external UCLA dataset. The area under the ROC curve (AUC) quantitatively compares the model classification overall performance

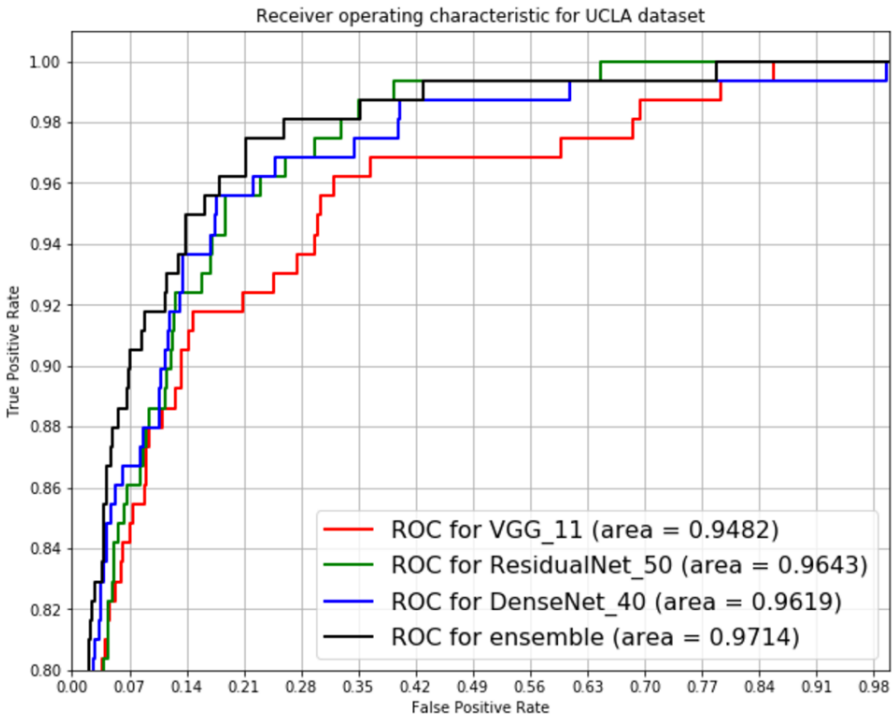


Figure 4.4: ROC plot on UCLA datasets for hybrid ensemble CNN model comparison.

and is frequently used as the metric to access the model performance in classifying nodules [SZY17, CHR15, HM16, CVS13, FCS17]. On LIDC dataset, the AUC score is 0.9944, 0.9917, 0.9884 and 0.9934 for the ensemble model, VGG model, residual model and densely connected model, respectively. On UCLA dataset, the AUC score is 0.9714, 0.9482, 0.9643 and 0.9619 for the ensemble model, VGG model, residual model and densely connected model, respectively.

A recent study by Froz et al. [FCS17] developed nodule classification model also using LIDC dataset with similar data size, and have reported state-of-art performances compared to previous works. My method obtained better results in all metrics compared to [FCS17]: an AUC of 0.994, sensitivity of 0.970, specificity of 0.973, and accuracy of 0.974 (vs. AUC of 0.922, sensitivity of 0.919, specificity of 0.948, and accuracy of 0.943. On UCLA dataset, My method achieved an AUC of 0.971, sensitivity of 0.886, specificity of 0.939, and accuracy of 0.942. The external validation results show that the ensemble model has robust performance.

Table 4.5: Results comparison: HSCNN versus 3D CNN.

Model	AUC (SD)	Accuracy (SD)	Sensitivity (SD)	Specificity (SD)
3D CNN	0.847 (0.024)	0.834 (0.022)	0.668 (0.040)	0.889 (0.022)
HSCNN	0.856 (0.026)	0.842 (0.025)	0.705 (0.045)	0.889 (0.022)

4.3.3 HSCNN results

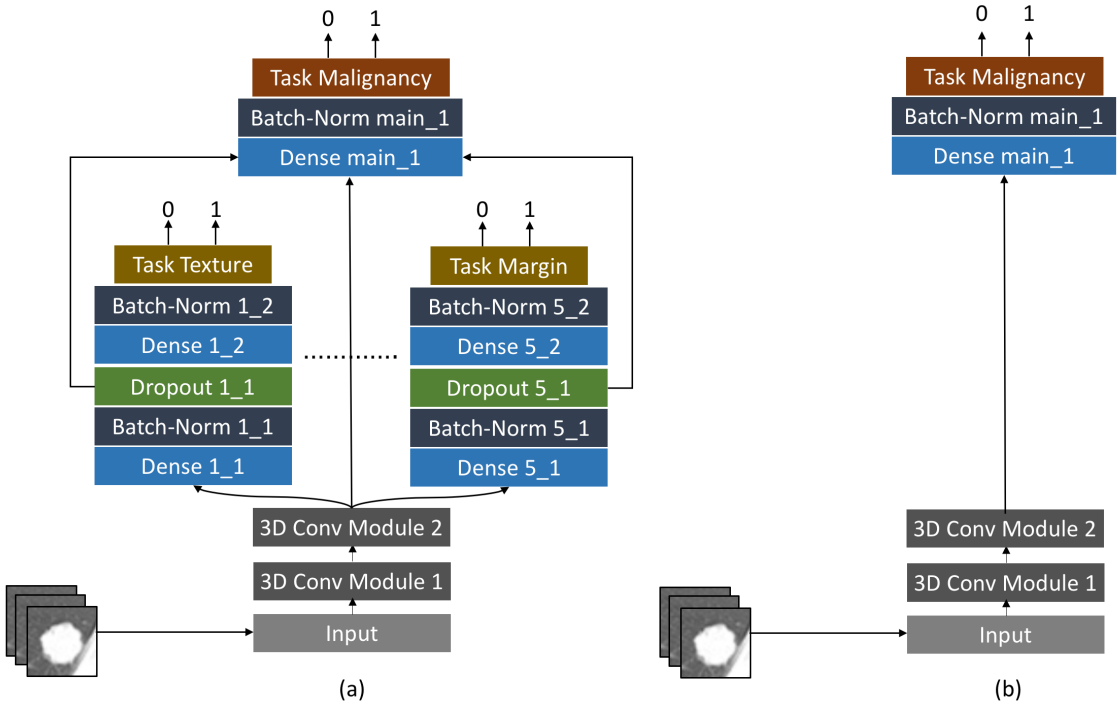


Figure 4.5: Framework comparison between proposed HSCNN and baseline 3D CNN. (a) proposed HSCNN architecture; (b) baseline 3D CNN architecture. Compared with the proposed HSCNN, baseline model has the same structure but without the low-level semantic task component.

I performed model training, validation, and testing using 897 LIDC cases, selected as described in Section 4.2.3.1. I split these cases into four subsets, where each subset had a similar number of nodules. A 4-fold cross validation study design was employed to obtain the final assessment of the model performance (i.e., for each fold, 2 subsets are used for training, 1 subset for validation, and 1 subset for holdout testing). Each subset is used as the test

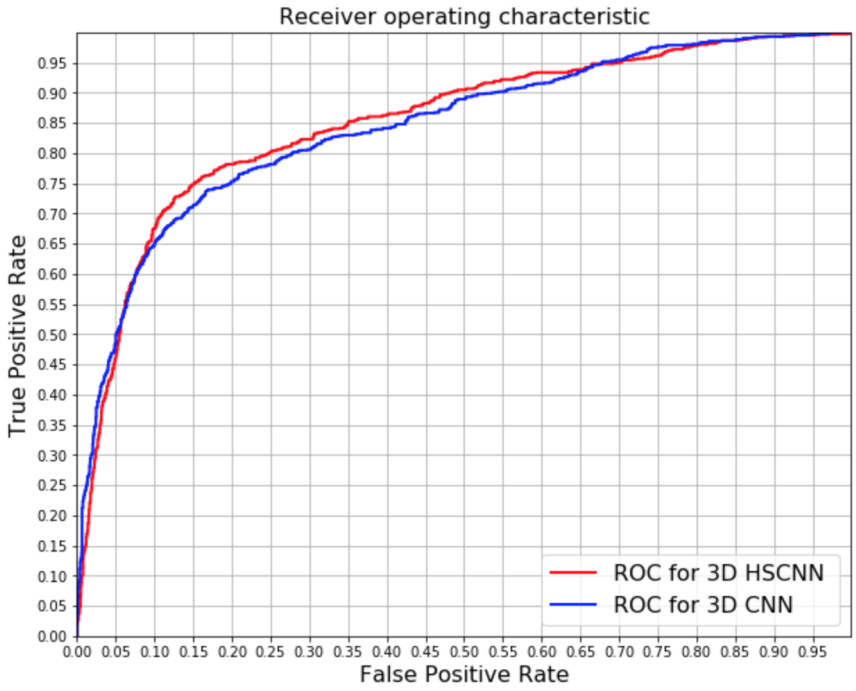


Figure 4.6: Receiver operating characteristic curve comparison: HSCNN versus 3D CNN.

set once during the cross validation. This design ensures that the test set is independent of model training and parameter optimizations, and should better reflect the true model performance without information leakage.

To evaluate and compare the HSCNN performance on lung nodule malignancy prediction, a 3D convolutional neural network (3D_CNN) was first implemented as a baseline model, shown in Figure 4.5b. This 3D CNN uses the same feature learning and high-level task components architecture as the HSCNN but has the low-level task component removed. The baseline model was trained and evaluated using the same 4-fold cross validation process and with the same data splitting for each fold (using the same randomization seed).

Figure 4.6 shows the ROC curve plots comparing HSCNN versus 3D CNN performance. By visual inspection of the ROC curves, HSCNN performs better than the traditional 3D CNN model. Table 4.6 summarizes the mean AUC score, accuracy, sensitivity, and specificity for both models. The HSCNN model achieved mean AUC 0.856, mean accuracy 0.842, mean sensitivity 0.705 and mean specificity 0.889; while the 3D CNN model achieved mean AUC 0.847, mean accuracy 0.834, mean sensitivity 0.668 and mean specificity 0.889. Both ROC

Table 4.6: Paired T-Test summarizes for AUC scores between HSCNN and 3D CNN model. CI represents for confidence interval.

Test Fold	**HSCNN AUC**	**3D CNN AUC**	**AUC Difference (HSCNN - 3D_CNN)**	**Paired T-Test**
Fold 1	0.878	0.869	0.009	P-value=0.005, Mean_difference=0.009, CI = [0.0051, 0.0129]
Fold 2	0.813	0.807	0.006	
Fold 3	0.874	0.862	0.012	
Fold 4	0.860	0.851	0.009	

plots and metric assessments show that the proposed HSCNN achieved better performance for malignancy prediction compared with the conventional 3D CNN approach.

To assess the statistical significance of model performance improvements, I conducted a paired sample t-test to evaluate the mean differences in AUC scores between the HSCNN and 3D CNN model. Let Group 1 be the AUC score of the HSCNN model for each holdout test fold during the cross validation, and then paired Group 2 consists of the corresponding AUC score for the 3D CNN for the same fold. The null hypothesis is that the mean difference of AUC score between these two models equals 0. Table 4.6 summarizes the AUC scores for these two paired groups and the t-test results. The test obtained a p-value of 0.005 and confidence interval of $[0.0051, 0.0129]$; this rejects the null hypothesis and indicates that the HSCNN model achieved a statistically significantly better AUC relative to the 3D CNN. The mean improvement of the AUC score is 0.009. This finding demonstrates that adding a low-level task component on existing CNN structure could improve the lung nodule malignancy prediction results.

I also compared this results with current deep learning models for lung nodule malignancy prediction, as reported in the literature to date. Kumar et al. [KWC15] developed a deep autoencoder-based model with 4,323 nodules of the LIDC dataset, achieving model accuracy of 0.7501. Hua et al. [HHH15] presented a CNN model and deep belief network (DBN) model. Both models were trained and validated using 2,545 lung nodule samples from LIDC. The CNN model had specificity of 0.787 and sensitivity 0.737; and the DBN model obtained

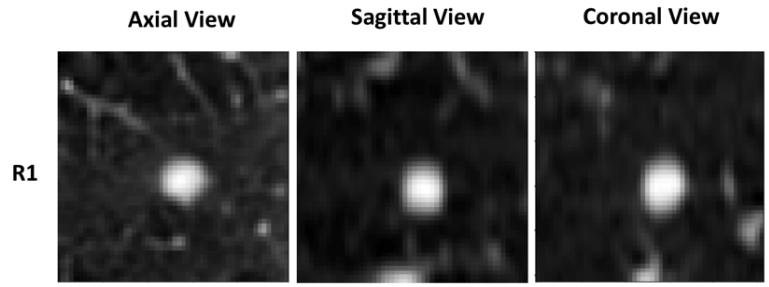
Table 4.7: Classification performance for semantic feature predictions.

Semantic Features	**Accuracy (SD)**	**AUC (SD)**	**Specificity (SD)**	**Sensitivity (SD)**
Calcification	0.908 (0.050)	0.930 (0.034)	0.763 (0.092)	0.930 (0.067)
Margin	0.725 (0.049)	0.776 (0.033)	0.632 (0.109)	0.758 (0.091)
Subtlety	0.719 (0.019)	0.803 (0.015)	0.796 (0.045)	0.673 (0.044)
Texture	0.834 (0.086)	0.850 (0.042)	0.636 (0.199)	0.855 (0.108)
Sphericity	0.552 (0.027)	0.568 (0.015)	0.554 (0.076)	0.552 (0.095)

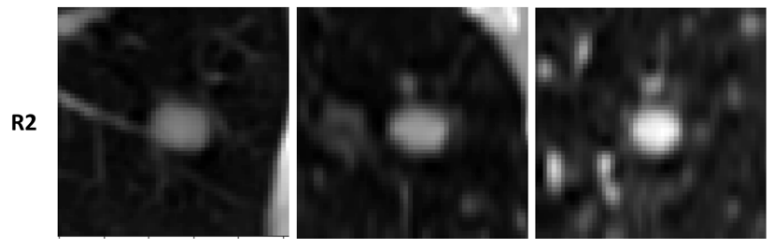
specificity of 0.822 and sensitivity 0.734. Shen et al. [SZY15] used a model based on multi-scale 3D CNN. Developed with 1,375 LIDC nodule samples, the average accuracy is reported above 0.84 with different configurations. Shen et al. [SZY17] extended this multi-scale model using a multi-crop approach and achieved accuracy of 0.839, 0.8636, and 0.8714 with 340, 1030 and 1375 nodules of LIDC, respectively. However, unlike my method, all these methods are evaluated with only training and validation data splits without independent holdout test dataset. This evaluation design might lead to information leakage due to using the validation data for optimal model parameters selection. Thus, this setting tends to over-estimate the model performance. In general, my model achieved better or similar performances compared with these reported methods.

Table 4.7 presents the classification performance for each of the low-level tasks (i.e., semantic features). Proposed model achieved mean accuracy of 0.908, 0.725, 0.719, 0.834 and 0.552; mean AUC score of 0.930, 0.776, 0.803, 0.850 and 0.568; mean sensitivity of 0.930, 0.758, 0.673, 0.855 and 0.552; and mean specificity of 0.763, 0.632, 0.796, 0.636 and 0.554 for calcification, margin, subtlety, texture, and sphericity, respectively. These results suggest that the HSCNN model is able to extract representations that are predictable for semantic features while achieving high nodule malignancy prediction power.

Figure 4.7 demonstrates the interpretability of the HSCNN model by visualizing the central slices of the lung nodule samples in axial, coronal, and sagittal views while presenting the predicted interpretable semantic labels along with the malignancy classification results.

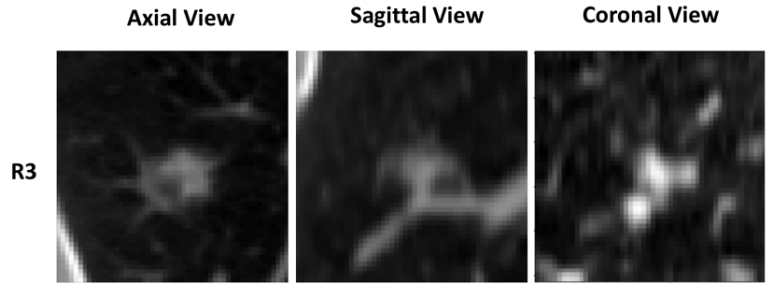


	True Label	Prediction
Calcification	1	1
Margin	1	1
Texture	1	1
Sphericity	1	1
Subtlety	1	1
Malignancy	0	0

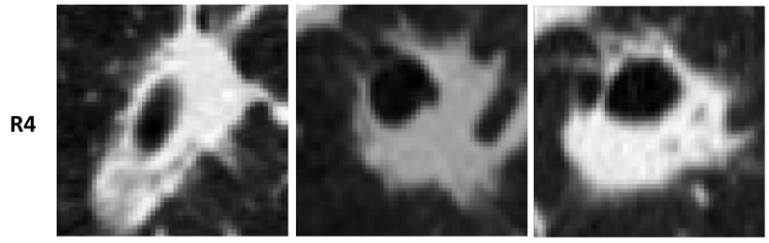


	True Label	Prediction
Calcification	1	1
Margin	1	1
Texture	1	1
Sphericity	1	1
Subtlety	1	1
Malignancy	0	0

(a)



	True Label	Prediction
Calcification	1	1
Margin	0	0
Texture	0	0
Sphericity	1	0
Subtlety	1	1
Malignancy	1	1



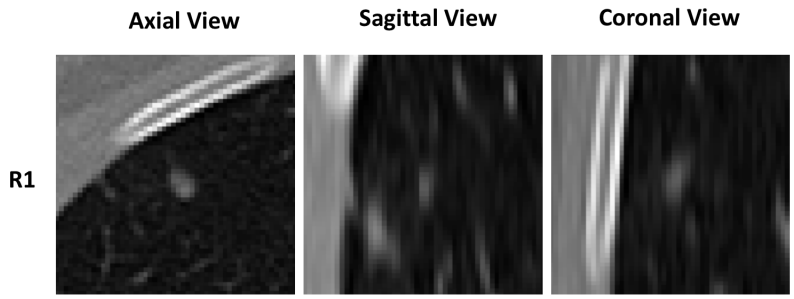
	True Label	Prediction
Calcification	1	1
Margin	0	0
Texture	0	0
Sphericity	0	0
Subtlety	1	1
Malignancy	1	1

(b)

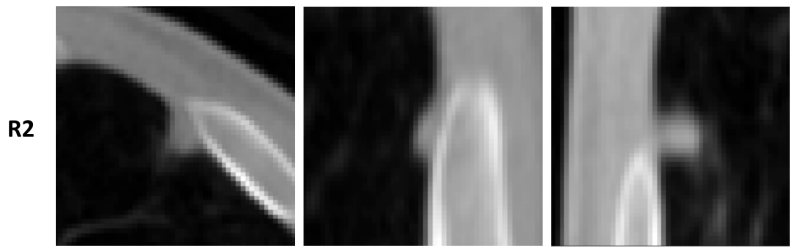
Figure 4.7: Illustrating the HSCNN model interpretability: lung nodule central slices, interpretable semantic feature prediction and malignancy prediction. R1, R2, R3 and R4 are four different nodules. (a) Central slices of axial, coronal and sagittal view of two benign nodule samples; true and predicted labels for interpretable semantic features and malignancy. (b) Central slices of axial, coronal and sagittal view of two malignant nodule samples; true and predicted labels for interpretable semantic features and malignancy.

Figure 4.7a-R1 shows that the HSCNN model classifies the lung nodule as benign (the true label is also benign), with this decision correlated to predictions of this nodule as having no calcification, sharp margins, roundness, obvious contrast between nodule and surroudings, and solid texture. The predictions of these five semantic characteristics are the same as the true label and corresponds to our knowledge about benign lung nodules. Compared to a 3D CNN malignancy prediction model, the HSCNN provides more insight for interpreting its predictions. Similarly, in Figure 4.7b-R3, the proposed model predicts the lung nodule as malignant (true label is also malignant). Different from the benign case, the HSCNN model predicts this nodule having poorly defined margins, non-solid texture, and non-round shape. This partly explains why the HSCNN makes a malignancy classification, with such nodule characteristics corresponding to our expert knowledge about typically malignant nodules. We note that the sphericity predictions made by the model are different from the true label. This result is explained by the fact that this nodule has a more regular round shape in axial view, but less round shape in the two other views, as shown in Figure 4.7b-R3.

Figure 4.8 shows representative cases where the HSCNN fails on predictions of semantic feature or cancer malignancy. Figure 4.8-R1 shows that the HSCNN model classifies the lung nodule correctly as benign, but incorrectly for four semantic features of this nodule (margin, texture, sphericity and subtlety). In Figure 4.8-R2, the HSCNN model incorrectly classify the lung nodule as malignant (the true label is benign). However, all semantic features of this nodule have been predicted correct as the true label. These two cases present the situation where the correctness of the predictions are inconsistent between the malignancy and semantic predictions. One possible solution to reduce this consistence and to improve the model performance is to add more semantic features, which are predictive for cancer malignancy, to the HSCNN model. Such semantic features could be the nodule size, spiculation and lobulation.



	True Label	Prediction
Calcification	1	1
Margin	1	0
Texture	0	1
Sphericity	0	1
Subtlety	1	0
Malignancy	0	0



	True Label	Prediction
Calcification	1	1
Margin	1	1
Texture	1	1
Sphericity	0	0
Subtlety	1	1
Malignancy	0	1

Figure 4.8: Represented cases where the HSCNN model predict incorrectly for semantic features or cancer malignancy. R1 and R2 are two different nodules. R1: one case has four incorrect semantic feature predictions, and the correct malignancy prediction. R2: one case have all correct semantic predictions, but incorrect malignancy prediction.

4.4 Discussion

One limitation of the proposed hybrid CNN method is that the model depends on the nodule candidate generation step to obtain the segmented candidate masks as the input data. Thus, the previous candidate generation step sets the upper bound of detection sensitivity for the following nodule detection (classification). If nodules are missed by the candidate generation step, they cannot be detected by the proposed model. In current settings, nodule candidate generation is a separate pre-processing component. Future work will explore build integrated and end-to-end lung nodule detection pipeline, which the nodule candidate generation and nodule detection steps will be combined in one joint deep learning framework. In this joint setting, both steps could be optimized and improved by minimizing training error in one unified framework.

There are some limitations to the HSCNN study. First, nodule characteristics of lobula-

tion and spiculation are well-known features correlated to lung cancer malignancy, but could not be used in this study given the labeling errors present in LIDC. Second, the original semantic features have scales of 5 or 6; changing the label into a binary variable may lose some label information. Still, using a binary label rather than the original labels is useful in this case. First, it overcomes a label sparsity issue, where the number of cases labeled for certain scales might be very small compared with the other scales (e.g., only 4 cases are labeled as popcorn calcification pattern, while 3,018 cases are labeled as absent of calcification pattern). Second, analysis shows that the inter-reader agreement is much lower for 5 or 6 scales compared with the proposed binary labels. Thus, binary labeling helps to reduce labeling noise caused by inter-reader variability. One way to overcome these two limitations is to collect a large dataset with more semantic labels. Although only five subtask modules are presented for the HSCNN architecture in this paper, the HSCNN framework and global loss function could be easily extended to increase or decrease the number of low-level semantic features. Future work will explore model improvement by adding more labeled semantic features and investigate model variability using different semantic labeling schema. More details about future works will be presented in Chapter 6.

CHAPTER 5

Lung Cancer Disease Progression Estimation

5.1 Overview

A growing number of individuals who are considered at high risk of cancer are now routinely undergoing population-level screening. However, noted harms such as radiation exposure, overdiagnosis, and overtreatment underscore the need for better temporal models that predict who should be screened and at what frequency. The mean sojourn time (MST), an average duration period when a tumor can be detected by imaging but with no observable clinical symptoms, is a critical variable for formulating screening policy. Estimation of MST has been long studied using continuous Markov models (CMM) with Maximum likelihood estimation (MLE). But traditional methods assume no observation error in interpreting the imaging data, which is unlikely and can bias the estimation of the MST. In addition, the MLE may not be stably estimated when data is sparse. Addressing these shortcomings, this chapter presents a probabilistic modeling approach for periodic cancer screening data. I first model the cancer state transition using a three state CMM model, while simultaneously considering observation error. Then, a novel Bayesian framework is presented to jointly estimate the MST and observation error. This study also considers the inclusion of covariates to estimate individualized rates of disease progression. In the remainder of this chapter, Section 5.2 details the developed Bayesian framework and CMM modeling approach. Section 5.3 describes the evaluation metrics and Section 5.4 shows the validation results, demonstrating that this method achieves more accurate and sensible estimates of MST in comparison to MLE. Section 5.5 discusses the limitations and compare this study with other works. The content of this chapter is based on my prior publication [SHP17].

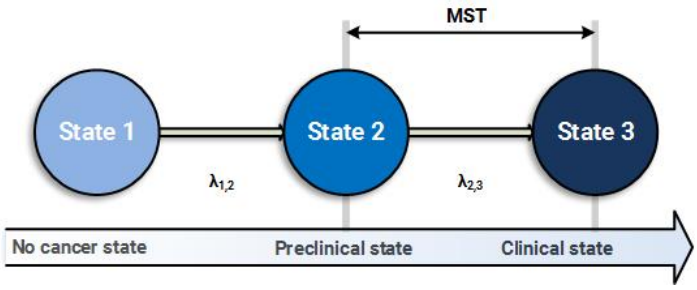


Figure 5.1: Model state transition diagram. State 1 is the disease-free state, State 2 is the preclinical state and State 3 is the clinical state. Parameters λ_{12} and λ_{23} are the transition intensities for transitioning from State 1 to State 2 and State 2 to State 3, respectively.

5.2 Materials and Methods

5.2.1 Overview

As with prior work, I model the natural progression of lung cancer as transitioning through three states (see Figure 5.1): a disease-free state (State 1), a preclinical state detectable via screening but asymptomatic (State 2), and a symptomatic state (State 3) [UHC10, DCT95, CDT96, CLC08]. The model assumes that a patient in State 1 must go through State 2 to reach State 3. When a patient undergoes screening, one of two states can be observed: if the screening result is positive and confirmed by a diagnostic evaluation, such as a biopsy, the patient is in the preclinical state; otherwise, the patient is in the disease-free state. Thus, the second state (preclinical) is identified under two conditions: 1) a positive screening test; and 2) a confirmed positive pathology diagnosis. However, patients in the preclinical state include both false-negatives due to interpretation error, and true-negatives, both of whom can progress to the clinical state. When cancer is first detected by emerging lung cancer symptoms (not through screening), the patient is in the clinical state. In multi-round screening settings, patients who do not progress to the clinical state and are not found to be preclinical during screening will repeat the process in subsequent rounds. Those found to be symptomatic of lung cancer prior to another round of screening are considered to be interval cases. Figure 5.2 illustrates this process. There is observation error when I observe State 1 (i.e., the underlying real state could be either State 1 or State 2), but no observation error

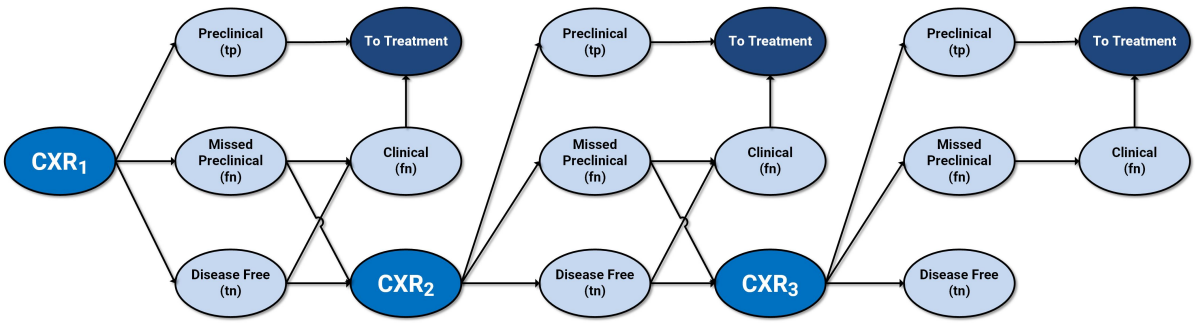


Figure 5.2: An illustration of possible outcomes from periodic CXR screening, where CXR_j represents the jth screening. CXR_2 and following screening will have similar possible outcomes and procedure as with CXR_1. If the subjects are observed in the preclinical state in the first screening, they will enter treatment (and stop periodic screening CXR). Otherwise, subjects are observed to be in the disease free state. However, these observed disease-free subjects include both false-negatives (missed preclinical cases) and true-negatives. Some subjects, who are found at the clinical state (lung cancer symptoms emerge) prior to another round of screening, are called interval cases and also will not undergo additional screening. These interval cases may come from missed preclinical subjects or true disease-free subjects. Subjects who do not progress to the clinical state repeat the process in subsequent rounds.

is assumed when I observe State 2 and State 3, as both are confirmed clinically.

MST is difficult to estimate because the direct transition from the disease-free state to the preclinical state is clinically unobservable. Patients will undergo intervention/treatment after being observed in a preclinical state (a positive cancer screening), thus obviating the natural progression from a preclinical state to a clinical state. Therefore, interval cases become the only source of information for estimating MST if no control group (individuals who never undergo screening) is available. As the discovery of interval cancers is affected by false-negative screening results, estimation of MST is affected by detection sensitivity; a biased estimate of sensitivity can influence the estimate of MST [CLC08]. Sensitivity is the unknown probability of screening detecting preclinical cancer.

Table 5.1: Detailed chest x-ray participant breakdown

Screening	**Detection Mode**	**Number**	**Transition Types**
First screening	Participants	25244	
	Screening-Detected Cases	132	No disease → Preclinical
	Negative Screening Cases	25112	No disease → No disease
	Interval Cancers (Between 1st/2nd screenings)	48	No disease → Clinical
Second screening	Participants	23506	
	Screening-Detected Cases	64	No disease → Preclinical
	Negative Screening Cases	23442	No disease → No disease
	Interval Cancers (Between 2nd/3rd screenings)	42	No disease → Clinical
Third Screening	Participants	22411	
	Screening-Detected Cases	74	No disease → Preclinical
	Negative Screening Cases	22337	No disease → No disease
	Post-screening Cancers (after 3rd screening)	484	No disease → Clinical

5.2.2 National lung screening trial (NLST) data

The National Lung Screening Trial was a large multi-center randomized controlled trial (RCT) of over $53,000$ high-risk current or former smokers. Participants were initially between 55 and 74 years old, had smoking histories of at least 30 pack-years and were cancer-free at the start of the trial. The study followed participants between $2002 - 2007$, with follow-up through 2009. Each participant had up to three rounds of screening, with roughly one year between screenings. The study consisted of two arms, chest x-ray (CXR) and low-dose computed tomography (CT). If at any point in the study the participant was found to have cancer, he/she did not receive further screenings and was removed from the trial. In this study, I utilize data from the CXR arm. Of the $26,730$ total patients originally in the CXR arm, 807 were removed from my analysis due to withdrawal from the study or

loss of contact, and 100 were removed because they were discovered to have been ineligible after enrollment (e.g., patient had a CT within 18 months of enrollment). A further 579 patients who did not receive the first round of screening were removed. Only clinically confirmed positive screenings are considered preclinical (State 2). Here, interval cancers are cases detected after a negative screen, but before the next screen (between first and second screen or between second and third screen). Post-screening cancer cases are those detected after a negative third screen during follow-up, and the follow-up time is up to 5.09 years. Both interval and post-screening cancers are assumed to be symptomatic cancers and defined as clinical (State 3). False-positives were considered to not be cancer (State 1). Table 5.1 presents a detailed breakdown of events.

5.2.3 Continuous-time Markov model

Let k and l denote one of the three disease states, where $k, l \in \{1, 2, 3\}$. Suppose the disease is in State k at time t and let $p_{kl}(t, t + \Delta t)$ denote the probability of transition from State k to l during time period Δt. Then, the instantaneous $\lambda_{kl}(t)$ transition intensity, which represents the instantaneous hazard rate of progression to State l [JST03], is

$$\lambda_{kl}(t) = \lim_{\Delta t \to 0} p_{kl}(t, t + \Delta t)/\Delta t. \tag{5.1}$$

Using a time-homogeneous model, both the transition intensity and transition probability is independent of t, where $\lambda_{kl}(t) = \lambda_{kl}$. In this case, the process is stationary and the transition probability $p_{kl}(\Delta t) = p_{kl}(t, t + \Delta t) = p_{kl}(0, \Delta t)$. The three-state instantaneous transition rate matrix Q is:

$$Q = \begin{bmatrix} -\lambda_{12} & \lambda_{12} & 0 \\ 0 & -\lambda_{23} & \lambda_{23} \\ 0 & 0 & 0 \end{bmatrix}, \tag{5.2}$$

whose rows sum to 0, so that the diagonal entries are [KL85, JST03]:

$$\lambda_{kk} = -\sum_{k \neq l} \lambda_{kl}. \tag{5.3}$$

As shown in Figure 5.1, transitions could only happen from State 1 to State 2 and from State 2 to State 3. For other undefined transitions, transition rates are 0. Transition rate

λ_{12} represents the instantaneous hazard rate to the preclinical state from the disease free state and λ_{23} is the instantaneous hazard rate of transitioning to clinical state from the preclinical state. In the CMM, MST is calculated as $1/\lambda_{23}$. The transition probability matrix in time Δt is the matrix exponential [KL85] $P(\Delta t) = \exp(Q\Delta t)$:

$$P(\Delta t) = \begin{bmatrix} \exp(-\lambda_{12}\Delta t) & \frac{\lambda_{12}(\exp(-\lambda_{23}\Delta t)-\exp(-\lambda_{12}\Delta t))}{\lambda_{12}-\lambda_{23}} & 1+\frac{\lambda_{23}\exp(-\lambda_{12}\Delta t)-\lambda_{12}\exp(\lambda_{23}\Delta t)}{\lambda_{12}-\lambda_{23}} \\ 0 & \exp(-\lambda_{23}\Delta t) & 1-\exp(-\lambda_{23}\Delta t) \\ 0 & 0 & 1 \end{bmatrix} \quad (5.4)$$

whose $(k,l)^{th}$ entry is $p_{kl}(\Delta t)$. I first assume no observation error, that is, sensitivity equals 1. I can easily write the likelihood function for each observation, and the parameters can be estimated using maximum likelihood. Transition probabilities from baseline to the first screening time point are conditional on no cancer at baseline. Probabilities for different transitions can be computed and are given in Table 5.2. The Markov assumption states that transitions only depend on the previous state. Patients are independent given the parameters and thus the log-likelihood function for all subjects is equal to the summation over all participants.

5.2.4 Modeling imperfect screening sensitivity

Equation (5.4) and Table (5.2) assume that sensitivity of the low-dose lung cancer screening exam is 100%. However, in practice, false negatives reduce the test sensitivity. In this section, I introduce a Bayesian model for inferring both sensitivity and transition probabilities simultaneously.

I develop the model for three rounds of screening, including interval cancer and post-screening cancer cases. Time intervals between the first and second screenings and second and third screenings are Δt_{12} and Δt_{23}, respectively. I assume Δt_{12} and Δt_{23} are the same for all participants. Let A be the average age of all participants at first screening and A is used as the time interval for the first screening (participants are assumed to be disease-free at birth and the first observation time is at first screening) [UHC10, DCT95, CLC08, SCL07].

Table 5.2: Likelihood function for the Markov model

Observation Type	**Probability**
Disease free at 1st screening	$P_1 = \frac{\exp(-\lambda_{12}\Delta t)}{\exp(-\lambda_{12}\Delta t)+\frac{\lambda_{12}(\exp(-\lambda_{23}\Delta t)-\exp(-\lambda_{12}\Delta t))}{\lambda_{12}-\lambda_{23}}}$
Disease free at 2nd or 3rd screening	$P_2 = \exp(-\lambda_{12}\Delta t)$
Preclinical disease at 1st screening	$P_3 = 1 - P_1$
Preclinical disease at 2nd or 3rd screening	$P_4 = \frac{\lambda_{12}(\exp(-\lambda_{23}\Delta t)-\exp(-\lambda_{12}\Delta t))}{\lambda_{12}-\lambda_{23}}$
Clinical disease (interval and post-screening cases)	$P_5 = 1 - \frac{\lambda_{23}\exp(-\lambda_{12}\Delta t)-\lambda_{12}\exp(\lambda_{23}\Delta t)}{\lambda_{23}-\lambda_{12}}$

The real state at time t is Y_t, a random variable with three possible states $\{1, 2, 3\}$ modeled by the 3-state homogeneous Markov process (Figure 5.1) with transition hazard matrix Q and transition probability matrix $P(\Delta t)$. The observed state at time t is denoted as Z_t, also with three possible states $\{1, 2, 3\}$. At screening, the observed state Z_t is subject to error due to false-negatives, meaning a preclinical state may be incorrectly observed as disease-free. The sensitivity S is $S = Pr(Z_t = 2|Y_t = 2)$ and $1 - S = Pr(Z_t = 1|Y_t = 2)$. The observation error for the other two states are assumed to be zero because they are confirmed clinically: $Pr(Z_t = 1|Y_t = 1) = 1$ and $Pr(Z_t = 3|Y_t = 3) = 1$. Suppose there are T rounds of screening in total (T observations), I make two conditional independence assumptions about Y_t and Z_t: 1) an observational independence assumption, such that the t^{th} observation given the t^{th} real state is independent of all other observations and real states on that subject

$$P(Z_t|Y_T, Z_T, Y_{T-1}, Z_{T-1}, ..., Y_{t+1}, Z_{t+1}, Y_t, Y_{t-1}, Z_{t-1}, ..., Y_1, Z_1) = P(Z_t|Y_t),$$

; and 2) a Markov assumption, such that the t^{th} real state given previous real states is independent of all previous observations and real states except the most recent real state,

$$P(Y_t|Y_{t-1}, Z_{t-1}, Y_{t-2}, Z_{t-2}, ..., Y_1, Z_1) = P(Y_t|Y_{t-1}).$$

Given the total number of participants in the j^{th} round of screening and the probability that a participant will be observed as preclinical this round, the number of observed preclinical cases at this screening follows a binomial distribution [CC08]. Let N_j be the total number of attendees at screening j; let π_j be the probability of being observed at the preclinical state at the j^{th} screening, and let n_j be the number of subjects observed at the preclinical state at the j^{th} screening, then I have:

$$n_j|\pi_j \sim B(N_j, \pi_j). \tag{5.5}$$

The probability π_1 that a subject is observed in the preclinical state on the first screening given the subject is not in the clinical state is:

$$\begin{aligned}
\pi_1 &= Pr(Z_A = 2|Y_A \neq 3) = Pr(Z_A = 2, Y_A \neq 3)/Pr(Y_A \neq 3) \\
&= (Pr(Z_A = 2|Y_A = 2)Pr(Y_A = 2) + Pr(Z_A = 1|Y_A = 2)Pr(Y_A = 2))/ \\
&\quad (Pr(Y_A = 2) + Pr(Y_A = 1)) \\
&= (S \cdot p_{12}(A))/(p_{12}(A) + p_{11}(A)).
\end{aligned} \tag{5.6}$$

Next π_2 is the probability that a subject is observed in the preclinical state on the second screening given that the patient is not observed in the clinical state after the first screen and was disease-free on the first screening. The true preclinical case here may come from two sources: 1) the patient is disease-free at the first screening and progresses to preclinical at the second screening; 2) the false-negative patient whose real state is preclinical at the first screening and stays in the preclinical state for the second screening.

$$\begin{aligned}
\pi_2 &= Pr(Z_{A+\Delta t_{12}} = 2|Z_A = 1, Y_A \neq 3, Y_{A+\Delta t_{12}} \neq 3) \\
&= Pr(Z_{A+\Delta t_{12}} = 2, Y_{A+\Delta t_{12}} \neq 3|Z_A = 1, Y_A \neq 3)/Pr(Y_{A+\Delta t_{12}} \neq 3|Z_A = 1, Y_A \neq 3) \\
&= (Pr(Z_{A+\Delta t_{12}} = 2, Y_{A+\Delta t_{12}} = 2|Z_A = 1, Y_A \neq 3) + Pr(Z_{A+\Delta t_{12}} = 2, Y_{A+\Delta t_{12}} = 1| \\
&\quad Z_A = 1, Y_A \neq 3))/(Pr(Y_{A+\Delta t_{12}} \neq 3|Z_A = 1, Y_A \neq 3)) \\
&= (Pr(Z_{A+\Delta t_{12}} = 2|Y_{A+\Delta t_{12}} = 2, Z_A = 1, Y_A \neq 3)Pr(Y_{A+\Delta t_{12}} = 2|Z_A = 1, Y_A \neq 3)
\end{aligned}$$

$$
\begin{aligned}
&+Pr(Z_{A+\Delta t_{12}}=2|Y_{A+\Delta t_{12}}=1,Z_A=1,Y_A\neq 3)Pr(Y_{A+\Delta t_{12}}=1|Z_A=1,Y_A\neq 3))/\\
&Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3).\\
=\;& S\cdot Pr(Y_{A+\Delta t_{12}}=2|Z_A=1,Y_A\neq 3)/Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3)\\
=\;& S\cdot(\beta_1\cdot p_{12}(\Delta t_{12})+\beta_2\cdot p_{22}(\Delta t_{12}))/\theta_1 \qquad (5.7)
\end{aligned}
$$

where (5.7) follows from assumptions (a) and (b) and

$$
\begin{aligned}
\beta_2 &= Pr(Y_A=2|Z_A=1,Y_A\neq 3)=Pr(Z_A=1|Y_A=2,Y_A\neq 3)Pr(Y_A=2|Y_A\neq 3)/\\
&\quad Pr(Z_A=1|Y_A\neq 3)\\
&= (1-S)\cdot\pi_1/((1-\pi_1)\cdot S), \qquad (5.8)\\
\beta_1 &= Pr(Y_A=1|Z_A=1,Y_A\neq 3)=1-\beta_2, \qquad (5.9)\\
\theta_1 &= Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3)\\
&= 1-Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3)\\
&= 1-Pr(Y_A=1|Z_A=1,Y_A\neq 3)Pr(Y_{A+\Delta t_{12}}=3|Y_A=1,Z_A=1,Y_A\neq 3)-Pr(\\
&\quad Y_A=2|Z_A=1,Y_A\neq 3)Pr(Y_{A+\Delta t_{12}}=3|Y_A=2,Z_A=1,Y_A\neq 3)\\
&= 1-\beta_1\cdot p_{13}(\Delta t_{12})-\beta_2\cdot p_{23}(\Delta t_{12}). \qquad (5.10)
\end{aligned}
$$

Finally, π_3 is the probability that one is observed as preclinical on the third screening given disease-free observations in the two previous screenings and not in the clinical state in any screening. The formula for π_3 is similar to that of π_2:

$$
\begin{aligned}
\pi_3 &= Pr(Z_{A+\Delta t_{12}+\Delta t_{23}}=2|Z_{A+\Delta t_{12}}=1,Z_A=1,Y_A\neq 3,Y_{A+\Delta t_{12}}\neq 3,Y_{A+\Delta t_{12}+\Delta t_{23}}\neq 3)\\
&= S\cdot(\gamma_1\cdot p_{12}(\Delta t_{23})+\gamma_2\cdot p_{22}(\Delta t_{23}))/\theta_2 \qquad (5.11)
\end{aligned}
$$

where

$$
\begin{aligned}
\gamma_2 &= Pr(Y_{A+\Delta t_{12}}=2|Z_{A+\Delta t_{12}}=1,Y_{A+\Delta t_{12}}\neq 3,Z_A=1,Y_A\neq 3)\\
&= Pr(Y_{A+\Delta t_{12}}=2,Z_{A+\Delta t_{12}}=1|Z_A=1,Y_A\neq 3)/(Pr(Y_{A+\Delta t_{12}}=2,Z_{A+\Delta t_{12}}=1|Z_A=1,\\
&\quad Y_A\neq 3)+Pr(Y_{A+\Delta t_{12}}=1,Z_{A+\Delta t_{12}}=1|Z_A=1,Y_A\neq 3))\\
&= (Pr(Z_{A+\Delta t_{12}}=1|Y_{A+\Delta t_{12}}=2,Z_A=1,Y_A\neq 3,Y_A\ =2)Pr(Y_{A+\Delta t_{12}}=2|Z_A=1,Y_A
\end{aligned}
$$

$$
\begin{aligned}
& \neq 3, Y_A = 2)Pr(Y_A = 2|Z_A = 1, Y_A \neq 3) + Pr(Z_{A+\Delta t_{12}} = 1|Y_{A+\Delta t_{12}} = 2, Z_A = 1, Y_A \\
& \neq 3, Y_A = 1)Pr(Y_{A+\Delta t_{12}} = 2|Z_A = 1, Y_A \neq 3, Y_A = 1)Pr(Y_A = 1|Z_A = 1, Y_A \neq 3))/ \\
& (Pr(Y_{A+\Delta t_{12}} = 2, Z_{A+\Delta t_{12}} = 1|Z_A = 1, Y_A \neq 3) + Pr(Y_{A+\Delta t_{12}} = 1, Z_{A+\Delta t_{12}} = 1|Z_A \\
& = 1, Y_A \neq 3)) \\
= \; & (1 - S) \cdot (p_{22}(\Delta t_{12})\beta_2 + p_{12}(\Delta t_{12})\beta_1)/((1 - S) \cdot (p_{22}(\Delta t_{12})\beta_2 + p_{12}(\Delta t_{12})\beta_1) + \\
& p_{1,1}(\Delta t_{12})\beta_1), \qquad (5.12)
\end{aligned}
$$

$$
\gamma_1 = Pr(Y_{A+\Delta t_{12}} = 1|Z_{A+\Delta t_{12}} = 1, Y_{A+\Delta t_{12}} \neq 3, Z_A = 1, Y_A \neq 3) = 1 - \gamma_2, \qquad (5.13)
$$

$$
\begin{aligned}
\theta_2 &= Pr(Y_{A+\Delta t_{12}+\Delta t_{23}} \neq 3|Z_{A+\Delta t_{12}} = 1, Y_{A+\Delta t_{12}} \neq 3, Z_A = 1, Y_A \neq 3) \\
&= 1 - \gamma_2 \cdot p_{23}(\Delta t_{23}) - \gamma_1 \cdot p_{13}(\Delta t_{23}). \qquad (5.14)
\end{aligned}
$$

Observed interval and follow-up cancers are assumed to be clinical. Let $N^{(3)}_{j\Delta t}$ be the number of subjects observed in the clinical state between the j^{th} and $(j+1)^{th}$ screening within time interval Δt. Then $N^{(3)}_{j\Delta t}$ is a random variable that can be modeled as a Poisson process [CC08]. Let $M_j(\Delta t)$ be the mean of the Poisson distribution. Then $M_j(\Delta t)$ is the sum of two parts: 1) the number of patients who progress to the clinical state from the disease-free state within Δt; and 2) the number of progressions from the preclinical states, which are false-negatives of the previous screening that transit into the clinical state within Δt [CC08]. The first part is the product of the number of disease-free patients at j^{th} screening $N^{(1)}_j$ times $p_{13}(\Delta t)$; and the second part is the product of the number of false-negative subjects at j^{th} screening times $p_{23}(\Delta t)$.

Section 5.4 discusses the priors adopted for the model. I use Markov Chain Monte Carlo (MCMC) to generate random samples from the joint posterior distribution of the parameters with the WinBUGS software program [LTB00].

5.2.5 Considering covariates

This model can be extended to evaluate the effects of covariates, such as gender and age, on parameter estimation. In this study, the effects of covariates are investigated with a stratified analysis by fitting my model separately for age groups > 60 and ≤ 60 to yield

independent estimates of parameters (λ_{12}, λ_{23}, S) for each age group [SCL07]. The same analysis is also performed on different gender groups. The NLST CXR dataset used in this work enrolled only high-risk lung cancer subjects, who are former or current smokers and have a minimum of 30 pack-years of cigarette smoking history. To further divide the dataset into sub-cohorts and investigate the cancer progression differences, covariates are identified from demographics, smoking history and disease history, including age, gender, family history of cancer, body mass index, disease history, cancer history, current or former smoker, number of packs of cigarette per year, and smoke years. Distributions of each covariate within no-cancer, non-symptomatic cancer, symptomatic cancer, and post-screening cancer groups are plotted and compared to identify the significant covariates in this high-risk cohort for further stratification. Age and gender are two significant and important factors that have also been used by previous lung cancer studies [ZW96, WER11], and are thus used here and reported in Section 5.4.

5.3 Evaluation

To validate the proposed model and compare with other methods, two metrics are employed. First, I use Pearson's chi-square to check the adequacy of the proposed model and validate parameter estimates by checking whether there are significant differences between the observed and expected counts [UHC10, CLC08, SCL07, CC08, Jac11]. A p-value larger than 0.05 suggests no significant difference indicating a good fit and accurate estimation of parameters. Then, posterior predictive p-values (PPPV) [GMS96] are employed to check the inconsistency between model predictions and observed counts for the Bayesian approach. A p-value away from 0 for the PPPV indicates a good fit.

In Section 5.4.1, I first present the results of a CMM model (no observation error) fit with MLE using NLST dataset for estimation of MST. I then present the results of the model including observation error using the proposed Bayesian approach in Section 5.4.2. Using Pearson's chi-square and PPPV, I show that the proposed model fits the data better than the model without observation error.

Table 5.3: Summaries of the posterior

Parameter	**Mean**	**SD**	**2.5%**	**97.5%**
λ_{12}	0.00525	0.000185	0.00489	0.00562
λ_{23}	0.927	0.0889	0.748	1.09
MST (year)	1.09	0.108	0.914	1.34
Sensitivity	0.899	0.0589	0.761	0.984

5.4 Results

Section 5.4.1 gives results using the simple CMM model assuming 100% sensitivity. Pearson's chi-square reveals a poor fit using simple CMM model. Section 5.4.2 gives results for transition intensities and sensitivity using the Bayesian approach with and without covariates. Both Pearson's chi-square test and the PPPV suggest a good fit for the proposed model. Based on my covariate analysis in Section 5.4.3, the MST is longer in the older population.

5.4.1 Maximum likelihood without observation error

MLE is used to estimate the parameters of the three-state CMM when assuming sensitivity is 1. Two parameters $\theta = (\lambda_{12}, \lambda_{23})$ need to be estimated. The likelihood calculation is implemented in R and quasi-Newton function maximization is used for the optimization step. The initial value for λ_{12}, the incidence rate of preclinical disease, is set to 0.00552 based on a study done by Manser et al. [CC08, MDC05]. The initial value for λ_{23}, the incidence rate of clinical disease and the inverse of MST, is set to 0.52 based on the inverse of the average reported CXR MST range $(0.46 - 3.35)$ [CLC08, KEW12, CEW14, CC08]. With these initial values, λ_{12} is estimated at 0.0154 (95% CI: $0.0143 - 0.0164$), λ_{23} is estimated at 3.31 (95% CI: $2.90 - 3.72$) and MST is estimated as 0.302 years (95% CI: $0.269 - 0.345$). However, the chi-square is 610.7 with p smaller than 0.00001, indicating a poor fit for the model.

Table 5.4: Goodness of fit with sensitivity < 1

	Observed	Expected	Residual
First screening negative	25112	25115.1	-3.1
First screening positive	132	128.9	3.1
Interval cancers after first screening	48	55	-7
Second screening negative	23442	23428.9	13.1
Second screening positive	64	77.1	-13.1
Interval cancers after second screening	42	47	-5
Third screening negative	22337	22339.2	2.2
Third screening positive	74	71.8	-2.2
Post-screening cancers after third screening	484	464.2	19.8
$\chi^2 = 4.643$, $P = 0.590$			

5.4.2 Bayesian Approach

In my analysis of the CXR data using the proposed Bayesian approach, there are now three parameters to be estimated, $\theta = (\lambda_{12}, \lambda_{23}, S)$. In previous studies, sensitivity of CXR in lung cancer is reported as being in the range of $69 - 90\%$ [WER11, CC08, TNK08] with a mean around 80%. Thus, a beta distribution with $\alpha = 5$ and $\beta = 2$ is adopted as the prior for sensitivity with a mode at 80%. Uniform distributions are employed as priors for λ_{12} and λ_{23}. A range from 0.0005 to 0.05 is selected as the prior for λ_{12} to allow enough flexibility given previously reported values in studies [CC08, MDC05] as described in section 5.4.1. Similarly, the prior of λ_{23} is chosen to be uniform of 0.2 to 5 based on previous studies [CLC08, KEW12, CEW14, CC08].

I use two sub-chains and WinBUGS [LTB00] to sample from the posterior for $\theta = (\lambda_{12}, \lambda_{23}, S)$. Each MCMC simulation is run for $45,000$ iterations, with a burn-in of $5,000$ iterations. Convergence was essentially immediate. After the burn-in, the posteriors are sampled and stored every 8 iterations, generating $5,000$ posterior samples per chain. The $10,000$ posterior samples, $\theta^i, i = 1, ..., 10000$, are pooled for analysis. The program is running on

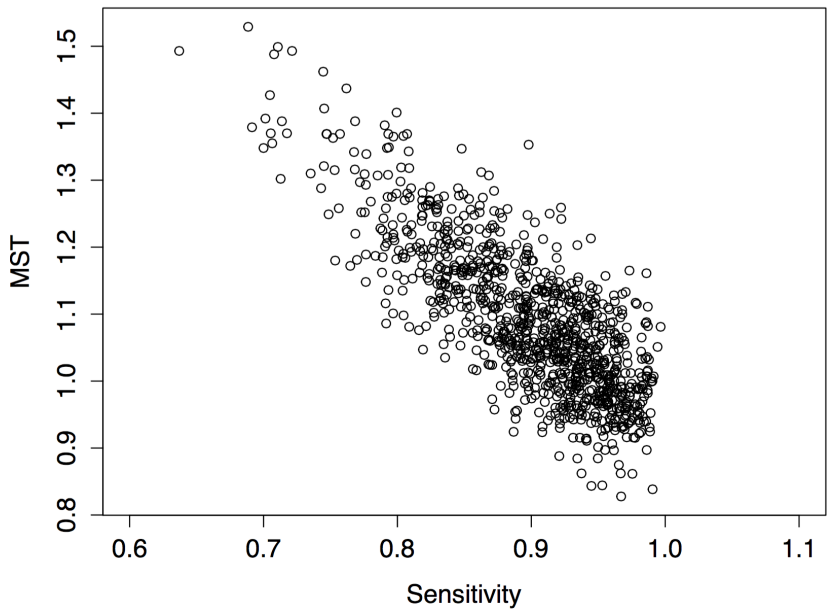


Figure 5.3: Scatter plot of 1000 randomly selected posterior samples of sensitivity and corresponding MST.

a Windows 8.1 desktop with a Intel Xeon CPU (3.3.GHz and 3.30GHz) and 16GB RAM. The WinBUGS program running time is 3.2 minutes.

Table 5.3 shows summaries of the posterior for the parameters. MST and sensitivity are estimated as 1.09 years (95% CI: $0.914 - 1.34$) and 0.899 (95% CI: $0.761 - 0.984$), respectively. Table 5.4 shows results of the chi-square test. There is no significant difference between observed and expected values, indicating a good fit to the empirical data. Compared to the model with perfect sensitivity, the expanded model fits the data better. The estimated MST is much longer compared to the reduced model fit assuming a sensitivity of 1. This corresponds to the expectation that lower sensitivity will lead to higher MST. This trend is also demonstrated in Figure 5.3, which plots 1000 randomly selected posterior samples of MST and sensitivity.

To further evaluate the expanded model, I also employ a posterior predictive p-value [GMS96] to assess the model fit. Let $y = (y_1, ..., y_9)$ denote the observed data, where y_a is the number of positive, negative and interval cases or post-screening cancer subjects for all three screenings. Let $y_{rep} = (y_{rep1}, ..., y_{rep9})$ be the replicated data that could have been

observed. A χ^2 discrepancy is the sum of squares of standardized residuals of the data with respect to their expectations under a posterior model and defined as [GMS96]

$$\chi^2(y;\theta) = \sum_{a=1}^{9} \frac{(y_a - E(y_a|\theta))^2}{Var(y_a|\theta)}. \tag{5.15}$$

where $Var(y_a|\theta)$ represents the variance of y_a given the parameter vector θ and $E(y_a|\theta)$ represents the expectation. For each posterior sample $\theta^{(b)}, b = 1, ..., 10000$, draw a simulated replicated data set, $y_{rep}^{(b)}$, from the sampling distribution $p(y_{rep}^{(b)}|\theta^{(b)})$. Then, calculate $\chi^2(y;\theta^{(b)})$ and $\chi^2(y_{rep}^{(b)};\theta^{(b)})$. Figure 5.4 plots $\{(\chi^2(y;\theta^{(b)}), \chi^2(y_{rep}^{(b)};\theta^{(b)})), b = 1, ..., 10000\}$. The estimated PPPV is calculated as the proportion of the 10,000 pairs for which $\chi^2(y_{rep}^{(b)};\theta^{(b)})$ exceeds $\chi^2(y;\theta^{(b)})$ [GMS96]. The estimated PPPV is 0.381 as shown in the figure and it does not indicate a lack fit for the model.

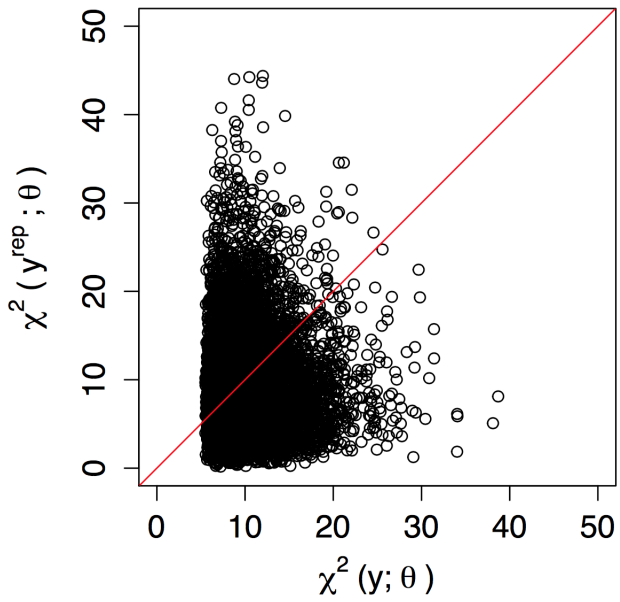


Figure 5.4: Scatter plot of predictive and realized log likelihood ratio discrepancies for the proposed Bayesian model using the whole CXR data set; the proportion of points above the red 45^o line represents the proportion of $\chi^2(y_{rep}^{(b)};\theta^{(b)})$ exceeding $\chi^2(y;\theta^{(b)})$ and is the posterior predictive p-value (PPPV). A PPPV away from 0 indicates a good model fit. The PPPV is 0.381.

Table 5.5: Summaries of the posterior for the two gender groups

Parameter	**Gender**	**Mean**	**SD**	**2.5%**	**97.5%**
λ_{12}	Male	0.00528	0.000241	0.00483	0.00577
	Female	0.00519	0.000294	0.00463	0.00578
λ_{23}	Male	0.908	0.112	0.694	1.13
	Female	0.889	0.125	0.652	1.14
MST	Male	1.12	0.143	0.884	1.44
	Female	1.15	0.171	0.878	1.54
Sensitivity	Male	0.871	0.0728	0.705	0.980
	Female	0.866	0.0801	0.677	0.980

Table 5.6: Summaries of the posterior for the two age groups

Parameter	**Age Group**	**Mean**	**SD**	**2.5%**	**97.5%**
λ_{12}	≤ 60	0.00322	0.000205	0.00284	0.00364
	> 60	0.00732	0.000314	0.00670	0.00795
λ_{23}	≤ 60	0.986	0.164	0.687	1.33
	> 60	0.872	0.0988	0.676	1.07
MST	≤ 60	1.04	0.181	0.753	1.46
	> 60	1.16	0.137	0.937	1.48
Sensitivity	≤ 60	0.848	0.0883	0.646	0.976
	> 60	0.881	0.0684	0.723	0.982

5.4.3 Covariate analysis

The explanatory variables (covariates) can be included using a stratified analysis. For gender, the data is divided into male and female and the model refit for each group. At the first screening, there are $14,936$ males and $10,308$ females. Table 5.5 shows posterior summaries for these two groups. There is little difference between females and males in terms of MST, sensitivity, and incidence rate of preclinical disease (λ_{12}). For age, the sample is divided into two groups: age ≤ 60 (12,669 participants) and those older than 60 (12,575 participants).

Table 5.7: Goodness of fit by age group

	Observed	Expected	Residual
age ≤ 60			
First screen negative	12631	12633.5	-2.5
First screen positive	38	35.5	2.5
Interval cancers after first screen	14	19.1	-5.1
Second screen negative	11828	11830.2	2.2
Second screen positive	18	15.8	-2.2
Interval cancers after second screen	16	15.2	0.8
Third screen negative	11258	11257.2	0.8
Third screen positive	20	20.8	-0.8
Post-screening cancers after third screen	154	146.9	7.1
$\chi^2 = 2.261$, $P = 0.894$			
age > 60			
First screen negative	12481	12481.3	-0.3
First screen positive	94	93.7	0.3
Interval cancers after first screen	34	38	-4
Second screen negative	11614	11605.3	8.7
Second screen positive	46	54.7	-8.7
Interval cancers after second screen	26	31.9	-5.9
Third screen negative	11079	11082.5	-3.5
Third screen positive	54	50.5	3.5
Post-screening cancers after third screen	330	316.8	13.2
$\chi^2 = 3.697$, $P = 0.718$			

Table 5.6 shows the posterior summaries for these two groups. The incidence rate λ_{12} of preclinical disease is two times larger in the older age group, and the 95% confidence intervals do not overlap. However, the MST and sensitivity are similar. This indicates that subjects (heavy smokers) older than 60 are twice as likely to progress to the preclinical state com-

Table 5.8: Goodness of fit by gender group

	Observed	Expected	Residual
Male			
First screen negative	14858	14859.7	-1.7
First screen positive	78	76.3	1.7
Interval cancers after first screen	30	34	-4
Second screen negative	13932	13927.8	4.2
Second screen positive	42	46.2	-4.2
Interval cancers after second screen	23	29	-6
Third screen negative	13307	13303.3	3.7
Third screen positive	39	42.7	-3.7
Post-screening cancers after third screen	293	277.9	15.1
$\chi^2 = 3.275$, $P = 0.774$			
Female			
First screen negative	10254	10255.4	-1.4
First screen positive	54	52.6	1.4
Interval cancers after first screen	18	22.8	-4.8
Second screen negative	9510	9500.8	9.2
Second screen positive	22	31.2	-9.2
Interval cancers after second screen	19	18.6	0.4
Third screen negative	9030	9036.3	-6.3
Third screen positive	35	28.7	6.3
Post-screening cancers after third screen	191	185.6	5.4
$\chi^2 = 5.323$, $P = 0.503$			

pared to those who are younger, consistent with the observations that there are significantly more detected preclinical cases as shown in Table 5.7. There is also a significantly higher percentage of observed interval/post-screening cancers in the older group. This is reasonable given that there are more subjects in the preclinical state. Table 5.7 and table 5.8 give the

chi-square test results for the models fit separetely by age and gender. The results suggest that the model fits well in all sub-groups. Compared to the whole population, fit is improved in each of the age groups. The estimated PPPV for the age sub-groups ≤ 60, > 60, male and female are $0.604, 0.507, 0.610$ and 0.259, respectively. All values indicate no lack of fit.

5.5 Discussion

This chapter presents a CMM-based model that incorporates observation error to model estimate multi-state disease progression. Applied to lung cancer screening data from the NLST dataset, the model produces results that are plausible and consistent with published literature [CLC08, WER11, CC08, HRK15]. The CMM is a natural approach to take for modeling the transitions of discrete health states [SCL07]. It can model transitions over time by incorporating longitudinal patient data and variable observation intervals. Sensitivity and MST are two important unknown parameters in the model. However, these two parameters are correlated and difficult to untangle as shown in figure 5.3, especially when no information is available for the incidence rate from a control group [UHC10]. Without a control group, MST can only be estimated from the occurrence of interval cancer cases. On the other hand, more false-negative cases leads to more occurrences of interval cases, resulting in a shorter MST estimate. Thus, MST and sensitivity should be modeled jointly [WVA05, WLV08] and estimates are sensitive to small changes in interval cancer counts [UHC10]. In [CLC08] and [SCL07], MST is estimated assuming sensitivity is 1, which is quite optimistic in reality. Duffy and Chen et al. [DCT95] proposed a two step method: firstly, MST is estimated by assuming sensitivity to be 1, and secondly, sensitivity is re-estimated using the obtained MST. This method is similarly still subject to error due to not estimating both jointly. In my study, I first investigated MST assuming sensitivity to be 1 similar to Step 1 in [DCT95]. As shown in Section 5.4.1, the model did not fit well. In [UHC10, CDT96, JST03, Jac11], sensitivity is modeled as part of the likelihood function, and MLE is adopted to estimate the parameters.

To compare with other methods, I implemented a three-state model from [UHC10] using

Table 5.9: Comparison between modeling approches

Model	**Comment**
Shih et al. [SCL07]	Uses a Markov model in conjunction with the prevalence pool concept, but has the limitations of assuming a steady state disease rate. Parameters are estimated using an Expectation-Maximum likelihood algorithm.
Chien et al. [CLC08]	Uses a Bayesian approach to estimate parameters of a 3-state lung cancer Markov model. The authors assumes imaging exams have 100% sensitivity.
Petousis et al. [PHA16]	Uses a discrete time dynamic Bayesian network to predict lung cancer incidence across time points. But the discrete time nature make it impossible to estimate MST.
Proposed method	Uses continuous-time Markov model and developed a Bayesian framework for parameter estimation. Provides analytical solutions to model observed occurrences for each state and jointly estimated MST and sensitivity. It is able to use all observed data including data from the third screening and post- screening cancer cases.

the R programming language [R C13] to model the same NLST CXR data. A general multi-state Markov model software package developed in [Jac11] was also used to try to fit the data. However, no stable estimates were obtained in either case. Overall, my method uses the probabilistic Bayesian approach to model the observed occurrences for each state and jointly estimated MST and sensitivity and provides improved fit. An additional benefit is that the likelihood is able to use all the data including data from the third screening and post-screening cancer cases.

There are some limitations to this work. First, the preclinical state is defined as the state in which the disease is detectable by screening. Therefore, depending on the screening

modality, the probability of transition into the preclinical state will also vary. For instance, CT has better resolution for detecting lung cancer, and a CT-screened patient might enter the preclinical state earlier relative to a CXR-screened patient. Therefore, MST and sensitivity are specific to each screening modality. Second, my current model models population level information by using average transition times between states as in [UHC10, CLC08, SCL07, CC08]. By using individual data, I can make inferences on the MST distribution across the patient population, possibly improving accuracy of the estimates. This also opens the path towards using patients' electronic health record (EHR) data for individualized screening schedules. Future work will focus on an individualized Bayesian framework that models each patient's information separately. In the reduced model where sensitivity is 1, both average and individualized patient transition times were investigated and the estimates for MST were very similar. Thus, the fit of my proposed model does not lose much generalizability using average transition time.

CHAPTER 6

Conclusion

6.1 Overview

This chapter summarizes the results and findings of this dissertation. It also presents the potential avenues of research and future studies as a result of this work.

6.2 Summary & Results

This dissertation addresses the needs for methods and tools that assist physicians to optimize the lung cancer screening and diagnosis workflow towards improving cancer diagnosis accuracy and reduce cost and radiation exposure. My approach was to develop methods for automated lung segmentation, nodule classification, cancer diagnosis and personalized periodic screening interval estimation. The specific contributions of this dissertation are as follows:

- *A parameter-free lung segmentation method to improve juxtapleural nodule detection accuracy.* I presented a novel approach to segment the lung using a bidirectional differential chain code combined with a machine learning framework.
- *A robust hybrid ensemble convolutional neural network for lung nodule classification.* I developed a hybrid convolutional neural network to differentiate lung nodules vs. non-nodules employing an ensemble of VGG module, residual module and densely connected module designs to improve the model robustness.
- *An interpretable hierarchical semantic convolutional neural network for lung cancer*

diagnosis. I described how to incorporate domain knowledge into the design of deep learning framework to enable model interpretation and improve lung cancer diagnosis.

- *A statistical multi-state disease estimation model*. I showed how to achieve more accurate and robust disease progression estimation using a novel Bayesian framework while considering observation error.

I performed a rigorous evaluation of the novel lung segmentation approach. The results in Section 3.4.3 have demonstrated the method was able to correctly include the juxtapleural nodules into the lung tissue while minimizing over- and under-segmentation. In addition, this method is generalizable to any task that involves concave/convex detection. For example, in magnetic resonance angiography, detection of concave/convex regions may be able to identify and accurately segment incidental aneurysms to assess their risk of rupture. The hybrid ensemble CNN model was tested using both a developed dataset and an external dataset as in Section 4.3.2, and has shown great robustness for dataset collected with different acquisition parameters. This model can be used for lung nodule classification, but also could serve as a robust feature extractor. The HSCNN model predicted semantic nodule characteristics along with the primary task of nodule malignancy diagnosis. The results in Section 4.3.3 suggest that the HSCNN model was able to quantify these nodule characteristics in a fully data-driven way and in one joint model for malignancy prediction. The predicted semantic labels were useful in interpreting the model's predictions for malignancy and correlated with known medical domain knowledge about pulmonary nodules and their relation to lung cancer. The produced model can be used as a semantic feature generator for unlabeled cases. In Chapter 5, the Bayesian-based multi-state disease progression estimation model builds the basis for providing individualized screening recommendations for a specific group of individuals stratified by their covariates. This work provides a way to more robustly calculate MST for specific cohorts with sparse observations, while considering the observation error. In clinical practice, with collected screening data (may be noisy and sparse), this work makes it possible to more accurately determine suitable screening periods. It also serves as the foundation to move towards individualized screening, where a personalized screening

paradigm will be provided for each subject.

In this dissertation work, the developed machine and deep learning models were trained and evaluated by splitting the datasets into training, validation and test sets. Training sets were used to minimize the training error and update the learnable model parameters. The validation sets were employed to tune the model hyper-parameters. The final assessment of the model performances were reported on the test sets as external holdouts. This design ensures that the test set is independent of model training and parameter optimizations, and should better reflect the true model performance without information leakage. We note that earlier studies in [SZY15, SZY17, KWC15, HHH15] only use training and validation splits during the cross validation process, without consideration for holdout test sets; such designs arguably have information leakage, and thus tend to over-estimate model performance.

6.3 Future Work

While a substantial amount of work has been done to develop the methodologies presented in Chapters 3, 4 and 5, these approaches can be refined through additional studies. As discussed in Section 3.5, the developed lung segmentation sometimes failed to re-include the juxtapleural nodules sitting in consolidation regions. One way to overcome this problem is to introduce region-based post-processing after obtaining the initial lung segmentation mask. Conditional random Markov field [CEE16] is one of the potential methods that could be used for this purpose. On the other hand, with the advancement of deep learning methods, region- and pixel-based convolutional neural networks could further be explored for lung segmentation to handle the juxtapleural nodules. Another limitation of the segmentation approach is that current inflection point detection and border correction stems were performed on each 2D image slice independently. One possible extension of this work is to perform 3D inflection point detection and 3D border correction for both lobes. With 3D information, the over-/under-segmentation error could be further reduced.

Deep learning models have been developed for lung nodule classification and diagnosis and prove effective. However, current lung segmentation, lung nodule candidate segmentation,

lung nodule classification and lung nodule diagnosis models were developed independently (i.e., as separate components). Recent progress in multi-task learning and semantic segmentation provide opportunities to solve lung segmentation, nodule detection, cancer diagnosis and nodule segmentation within one single framework. For example, He et al. [HGD17] developed a mask R-CNN approach to perform object detection, localization and semantic segmentation within one deep learning model. Mask R-CNN could serve as the basis to develop the multi-task model for lung segmentation, nodule detection, cancer diagnosis and nodule segmentation. Additionally, the lung nodule detection model currently only has binary labels: nodule and non-nodule. A future extension of this work could use additional labels (e.g., blood vessel segments) to perform multi-class classification. This multi-class setting will provide further supervision that could potentially improve the model performance to detect lung nodules. Chapter 4 has presented the importance of building interpretable CNN approach for lung cancer diagnosis and the developed HSCNN model has enabled this model interpretation by predicting the nodule's semantic features along with its malignancy. With these semantic feature outputs, human experts can now easily read and provide feedback about the correctness of these semantic outputs. Opportunity now exists to extend the model using active learning approach to incorporate the feedback from human experts to improve the model accuracy for cancer diagnosis. As discussed in Chapter 4, current semantic feature labels are binarized due to sparsity issues. With more labels and data collected in the future, multi-labels or continuous labels could be used to provide more information. In this case, the information of label distributions could be incorporated into the model's design to boost performance. Furthermore, Chapter 4 also presented that HSCNN could be easily extended to incorporate more semantic features. Nonetheless, too many semantic features (e.g., more than 20) could make the model convergence more challenging. Thus, improving the model scalability to account for large number of semantic features should be studied in future works. Notably, not all combinations of semantic labels could co-occur in reality given known domain knowledge. Therefore, this observation could be employed to improve model design. In the end, the inputs of current models were the 3D cubes centered at each nodule with all background pixel intensities. The background object, such as the lung walls

of the juxtapleural nodules, might prevent the model from learning useful information for the classification task. A possible future work is to explore feeding the deep learning model with nodule versus perinodular (background) regions as two distinct separated inputs for each input data.

Finally, as discussed in Section 5.5, my disease progression estimation approach modeled population level information by using average transition times between states. Compared with using individualized data, this setting reduced the computation complexity, but may lose individual information. Future work could use individual transition times for the model, and thus could possibly improve accuracy of the estimates. Another limitation of the work is that only two covariates were validated for the model. The influences of more factors could be explored in future works. Lastly, current approach made a time-homogeneous assumption, where both the states transition probability matrix and screening sensitivity were assumed to be invariant across time. Future works could relax this assumption and investigate the influence on the estimated parameters.

6.4 Concluding Remarks

Existing low-dose CT screening programs generate large volumes of screening data, but has challenges of high over-diagnosis rates, high cost and increased radiation exposure. These challenges highlight the need for assisting physicians with an automated lung cancer detection and diagnosis system to reduce false positive findings and make the screening program more cost-effective. It emphasizes the need to provide personalized screening recommendations for individuals with various risk factors to reduce unnecessary cost and radiation exposure while maintaining high cancer detection power. A first potential extension of this dissertation study was to build an interpretable automated clinical diagnosis pipeline to assist radiologists' reading procedure in an interactive way to improve the efficiency and accuracy of lung cancer early detection and diagnosis. This pipeline would provide: 1) lung segmentation and lobe volume quantification; 2) lung nodule localization and segmentation; 3) lung nodule malignancy prediction and characterization; and 4) interactive feedback loop to incorporate

opinions from physicians. Another potential extension of this work is to develop a personalized lung cancer screening recommendation system for individual subjects with different risk factors utilizing the multi-state disease progression estimation model. This system would estimate individual disease states' transition times by including various factors, such as age, gender, smoking history, and disease history. These two potential products would be significant steps towards optimizing the current low-dose CT screening program and improve early lung cancer detection. The contributions of this dissertation lay the foundation towards enabling creating such products.

[AAW03] Samuel G Armato, Michael B Altman, Joel Wilkie, Shusuke Sone, Feng Li, Arunabha S Roy, et al. "Automated lung nodule classification following automated nodule detection on CT: A serial approach." *Medical Physics*, **30**(6):1188–1197, 2003.

[ABC16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. "TensorFlow: A System for Large-Scale Machine Learning." In *OSDI*, volume 16, pp. 265–283, 2016.

[AEF07] Asem M Ali, Ayman S El-Baz, and Aly A Farag. "A Novel Framework for Accurate Lung Segmentation using Graph Cuts." In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pp. 908–911. IEEE, 2007.

[AF08] Asem M Ali and Aly A Farag. "Automatic Lung Segmentation of Volumetric Low-dose CT Scans using Graph Cuts." In *International Symposium on Visual Computing*, pp. 258–267. Springer, 2008.

[AFA97] Odd O Aalen, Vernon T Farewell, Daniela de Angelis, Nicholas E Day, and O Nöel Gill. "A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales." *Statistics in Medicine*, **16**(19):2191–2210, 1997.

[AGS10] Elisabeth Arnoldi, Mulugeta Gebregziabher, U Joseph Schoepf, Roman Goldenberg, Luis Ramos-Duran, Peter L Zwerner, Konstantin Nikolaou, Maximilian F Reiser, Philip Costello, and Christian Thilo. "Automated Computer-aided Stenosis Detection at Coronary CT Angiography: Initial Experience." *European radiology*, **20**(5):1160–1167, 2010.

[AHK91] Per Kragh Andersen, Lars Sommer Hansen, and Niels Keiding. "Assessing the influence of reversible disease indicators on survival." *Statistics in Medicine*, **10**(7):1061–1067, 1991.

[AMB11] Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." *Medical physics*, **38**(2):915–931, 2011.

[AMM04] Samuel G Armato III, Geoffrey McLennan, Michael F McNitt-Gray, Charles R Meyer, David Yankelevitz, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, et al. "Lung image database consortium: developing a resource for the medical imaging research community." *Radiology*, **232**(3):739–748, 2004.

[ARK09] Samuel G Armato, Rachael Y Roberts, Masha Kocherginsky, Denise R Aberle, Ella A Kazerooni, Heber MacMahon, Edwin JR van Beek, David Yankelevitz, Geoffrey McLennan, Michael F McNitt-Gray, et al. "Assessment of Radiologist Performance in the Detection of Lung Nodules: Dependence on the Definition of ?Truth?" *Academic radiology*, **16**(1):28–38, 2009.

[AS04] Samuel G Armato and William F Sensakovic. "Automated Lung Segmentation for Thoracic CT: Impact on Computer-aided Diagnosis1." *Academic Radiology*, **11**(9):1011–1021, 2004.

[ATR10] Pavan Annangi, Sheshadri Thiruvenkadam, A Raja, Hao Xu, XiWen Sun, and Ling Mao. "A region based active contour method for x-ray lung segmentation using prior shape and low level features." In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pp. 892–895. IEEE, 2010.

[BB08] Olivier Bousquet and Léon Bottou. "The tradeoffs of large scale learning." In *Advances in neural information processing systems*, pp. 161–168, 2008.

[BB12] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization." *Journal of Machine Learning Research*, **13**(Feb):281–305, 2012.

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives." *IEEE transactions on pattern analysis and machine intelligence*, **35**(8):1798–1828, 2013.

[Ben09] Yoshua Bengio et al. "Learning deep architectures for AI." *Foundations and trends® in Machine Learning*, **2**(1):1–127, 2009.

[BGM00] Matthew S Brown, Jonathan G Goldin, Michael F McNitt-Gray, Lloyd E Greaser, Amita Sapra, Kuo-Tung Li, James W Sayre, Katherine Martin, and Denise R Aberle. "Knowledge-based segmentation of thoracic computed tomography images for assessment of split lung function." *Medical physics*, **27**(3):592–598, 2000.

[BGS03] Matthew S Brown, Jonathan G Goldin, Robert D Suh, Michael F McNitt-Gray, James W Sayre, and Denise R Aberle. "Lung micronodules: automated method for detection at thin-section CT?initial experience." *Radiology*, **226**(1):256–262, 2003.

[BGS14] William C Black, Ilana F Gareen, Samir S Soneji, JoRean D Sicks, Emmett B Keeler, Denise R Aberle, Arash Naeim, Timothy R Church, Gerard A Silvestri, Jeremy Gorelick, et al. "Cost-effectiveness of CT screening in the National Lung Screening Trial." *New England Journal of Medicine*, **371**(19):1793–1802, 2014.

[BKN05] Kyongtae T Bae, Jin-Sung Kim, Yong-Hum Na, Kwang Gi Kim, and Jin-Hwan Kim. "Pulmonary Nodules: Automated Detection on CT Images with Morphologic Matching Algorithm?Preliminary Results." *Radiology*, **236**(1):286–293, 2005.

[BKT03] Peter B Bach, Michael W Kattan, Mark D Thornquist, Mark G Kris, Ramsey C Tate, Matt J Barnett, Lillian J Hsieh, and Colin B Begg. "Variations in lung cancer risk among smokers." *Journal of the National Cancer Institute*, **95**(6):470–478, 2003.

[BLP07] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. "Greedy layer-wise training of deep networks." In *Advances in neural information processing systems*, pp. 153–160, 2007.

[CBG14] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks." In *Medical Imaging 2014: Digital Pathology*, volume 9041, p. 904103. International Society for Optics and Photonics, 2014.

[CC08] Chun Ru Chien and Tony Hsiu Hsi Chen. "Mean sojourn time and effectiveness of mortality reduction for lung cancer screening with computed tomography." *International Journal of Cancer*, **122**(11):2594–2599, 2008.

[CC12] Wook-Jin Choi and Tae-Sun Choi. "Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images." *Information Sciences*, **212**:57–78, 2012.

[CC13] Wook-Jin Choi and Tae-Sun Choi. "Automated Pulmonary Nodule Detection System in Computed Tomography Images: A Hierarchical Block Classification Approach." *Entropy*, **15**(2):507–523, 2013.

[CDT96] HH Chen, SW Duffy, and Laszlo Tabar. "A Markov Chain Method to Estimate the Tumour Progression Rate from Preclinical to Clinical Phase, Sensitivity and Positive Predictive Value for Mammography in Breast Cancer Screening." *The statistician*, pp. 307–317, 1996.

[CEE16] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin DAnastasi, et al. "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 415–423. Springer, 2016.

[CEW14] YT Chen, D Erwin, and D Wu. "Over-diagnosis in Lung Cancer Screening using the MSKC-LCSP Data." *J Biomet Biostat*, **5**(201):2, 2014.

[CGG12] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. "Deep neural networks segment neuronal membranes in electron microscopy images." In *Advances in neural information processing systems*, pp. 2843–2851, 2012.

[CGZ06] Kathleen A Cronin, Mitchell H Gail, Zhaohui Zou, Peter B Bach, Jarmo Virtamo, and Demetrius Albanes. "Validation of a model of lung cancer risk prediction among smokers." *Journal of the National Cancer Institute*, **98**(9):637–640, 2006.

[Cho15] François Chollet et al. "Keras.", 2015.

[CHR15] Francesco Ciompi, Bartjan de Hoop, Sarah J van Riel, Kaman Chung, Ernst Th Scholten, Matthijs Oudkerk, Pim A de Jong, Mathias Prokop, and Bram van Ginneken. "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box." *Medical image analysis*, **26**(1):195–202, 2015.

[CLC08] Chun Ru Chien, Mei Shu Lai, and Tony Hsiu Hsi Chen. "Estimation of mean sojourn time for lung cancer by chest X-ray screening with a Bayesian approach." *Lung Cancer*, **62**(2):215–220, 2008.

[COM13] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 403–410. Springer, 2013.

[CP83] Judy S Chen and Philip C Prorok. "Lead time estimation in a controlled screening program." *American Journal of Epidemiology*, **118**(5):740–751, 1983.

[CQY16] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. "Dcan: Deep contour-aware networks for accurate gland segmentation." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2487–2496, 2016.

[CV01] Tony F Chan and Luminita A Vese. "Active contours without edges." *IEEE Transactions on image processing*, **10**(2):266–277, 2001.

[CVS13] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository." *Journal of digital imaging*, **26**(6):1045–1057, 2013.

[CZX12] Hui Chen, Jing Zhang, Yan Xu, Budong Chen, and Kuan Zhang. "Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans." *Expert Systems with Applications*, **39**(13):11503–11509, 2012.

[DBS15] Nóirín Duggan, Egil Bae, Shiwen Shen, William Hsu, Alex Bui, Edward Jones, Martin Glavin, and Luminita Vese. "A technique for lung nodule candidate detection in CT using global minimization methods." In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 478–491. Springer, 2015.

[DCT95] Stephen W Duffy, Hsiu-Hsi Chen, Laszlo Tabar, and Nicholas E Day. "Estimation of Mean Sojourn Time in Breast Cancer Screening using a Markov Chain Model of Both Entry to and Exit from the Preclinical Detectable Phase." *Statistics in medicine*, **14**(14):1531–1543, 1995.

[DMA13] Ashis Kumar Dhara, Sudipta Mukhopadhyay, Naved Alam, and Niranjan Khandelwal. "Measurement of spiculation index in 3D for solitary pulmonary nodules in volumetric lung CT images." In *Medical Imaging 2013: Computer-Aided Diagnosis*, volume 8670, p. 86700K. International Society for Optics and Photonics, 2013.

[DMH10] George Dahl, Abdel-rahman Mohamed, Geoffrey E Hinton, et al. "Phone recognition with the mean-covariance restricted Boltzmann machine." In *Advances in neural information processing systems*, pp. 469–477, 2010.

[DMM06] Marco Das, Georg Mhlenbruch, Andreas H Mahnken, Thomas G Flohr, Lutz Gndel, Sven Stanzel, Thomas Kraus, Rolf W Gnther, and Joachim E Wildberger. "Small Pulmonary Nodules: Effect of Two Computer-aided Detection Systems on Radiologist Performance 1." *Radiology*, **241**(2):564–571, 2006.

[Doi05] Kunio Doi. "Current Status and Future Potential of Computer-aided Diagnosis in Medical Imaging." *The British journal of radiology*, **78**(suppl_1):s3–s19, 2005.

[Doi07] Kunio Doi. "Computer-aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." *Computerized medical imaging and graphics*, **31**(4):198–211, 2007.

[DSY10] Li Deng, Michael L Seltzer, Dong Yu, Alex Acero, Abdel-rahman Mohamed, and Geoff Hinton. "Binary coding of speech spectrograms using a deep auto-encoder." In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[Duf05a] S. W Duffy. "Screening, Sojourn Time." *Encyclopedia of Biostatistics*, **7**, 2005.

[Duf05b] Stephen W Duffy. *Screening, Sojourn Time*. Wiley Online Library, 2005.

[DUJ15] Samantha K Dilger, Johanna Uthoff, Alexandra Judisch, Emily Hammond, Sarah L Mott, Brian J Smith, John D Newell, Eric A Hoffman, and Jessica C Sieren. "Improved pulmonary nodule classification utilizing quantitative lung parenchyma features." *Journal of Medical Imaging*, **2**(4):041004, 2015.

[DW84] Nicholas E Day and Stephen D Walter. "Simplified models of screening for chronic disease: estimation procedures from mass screening programmes." *Biometrics*, pp. 1–13, 1984.

[ECM00] Jeremy J Erasmus, John E Connolly, H Page McAdams, and Victor L Roggli. "Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions." *Radiographics*, **20**(1):43–58, 2000.

[EKN17] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature*, **542**(7639):115–118, 2017.

[ER15] Mehmet Günhan Ertosun and Daniel L Rubin. "Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A Modular Approach with Ensemble of Convolutional Neural Networks." In *AMIA Annual Symposium Proceedings*, volume 2015, p. 1899. American Medical Informatics Association, 2015.

[FAG11] Amal Farag, Asem Ali, James Graham, Aly Farag, Salwa Elshazly, and Robert Falk. "Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose CT scans of the chest." In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 169–172. IEEE, 2011.

[FCS17] Bruno Rodrigues Froz, Antonio Oseas de Carvalho Filho, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. "Lung nodule classification using artificial crawlers, directional texture and support vector machine." *Expert Systems with Applications*, **69**:176–188, 2017.

[FN93] Ken-ichi Funahashi and Yuichi Nakamura. "Approximation of dynamical systems by continuous time recurrent neural networks." *Neural networks*, **6**(6):801–806, 1993.

[For15] U.S. Preventive Services Task Force. "Final Update Summary: Lung Cancer: Screening." 2015.

[GAG08] Fiona J Gilbert, Susan M Astley, Maureen GC Gillan, Olorunsola F Agbaje, Matthew G Wallis, Jonathan James, Caroline RM Boggis, and Stephen W Duffy. "Single Reading with Computer-aided Detection for Screening Mammography." *New England Journal of Medicine*, **359**(16):1675–1684, 2008.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[GMS96] Andrew Gelman, Xiao Li Meng, and Hal Stern. "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, **6**(4):733–760, 1996.

[GPC16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *JAMA*, **316**(22):2402–2410, 2016.

[GSC05] Zhanyu Ge, Berkman Sahiner, Heang-Ping Chan, Lubomir M Hadjiiski, Philip N Cascade, Naama Bogot, Ella A Kazerooni, Jun Wei, and Chuan Zhou. "Computer-aided detection of lung nodules: False positive reduction using a 3D

gradient field method and 3D ellipsoid fitting." *Medical physics*, **32**(8):2443–2454, 2005.

[GSJ15] Bram van Ginneken, Arnaud AA Setio, Colin Jacobs, and Francesco Ciompi. "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans." In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pp. 286–289. IEEE, 2015.

[Gur93] JW Gurney. "Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory." *Radiology*, **186**(2):405–413, 1993.

[GW07] Rafael C Gonzalez and Richard E Woods. "Image processing." *Digital image processing*, **2**, 2007.

[HAG82] LW Hedlund, Roger Fabian Anderson, PL Goulding, JW Beck, EL Effmann, and CE Putman. "Two methods for isolating the lung area of a CT scan for density information." *Radiology*, **144**(2):353–357, 1982.

[HBM08] David M Hansell, Alexander A Bankier, Heber MacMahon, Theresa C McLoud, Nestor L Muller, and Jacques Remy. "Fleischner Society: Glossary of Terms for Thoracic Imaging." *Radiology*, **246**(3):697–722, 2008.

[HGD17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988. IEEE, 2017.

[HHH15] Kai-Lung Hua, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Yu-Jen Chen. "Computer-aided classification of lung nodules on computed tomography images via deep learning technique." *OncoTargets and therapy*, **8**, 2015.

[HHR01] Shiying Hu, Eric A Hoffman, and Joseph M Reinhardt. "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images." *IEEE transactions on medical imaging*, **20**(6):490–498, 2001.

[HKB14] Samuel H Hawkins, John N Korecki, Yoganand Balagurunathan, Yuhua Gu, Virendra Kumar, Satrajit Basu, Lawrence O Hall, Dmitry B Goldgof, Robert A Gatenby, and Robert J Gillies. "Predicting outcomes of nonsmall cell lung cancer using CT image features." *IEEE Access*, **2**:1418–1426, 2014.

[HLW16] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. "Densely connected convolutional networks." *arXiv preprint arXiv:1608.06993*, 2016.

[HM16] Matthew C Hancock and Jerry F Magnan. "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods." *Journal of Medical Imaging*, **3**(4):044504, 2016.

[HOT06] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural computation*, **18**(7):1527–1554, 2006.

[HPY17] Peng Huang, Seyoun Park, Rongkai Yan, Junghoon Lee, Linda C Chu, Cheng T Lin, Amira Hussien, Joshua Rathmell, Brett Thomas, Chen Chen, et al. "Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study." *Radiology*, **286**(1):286–295, 2017.

[HRK15] Kevin ten Haaf, Joost van Rosmalen, and Harry J de Koning. "Lung Cancer Detectability by Test, Histology, Stage, and Gender: Estimates from the NLST and the PLCO Trials." *Cancer Epidemiology Biomarkers & Prevention*, **24**(1):154–161, 2015.

[HS06] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks." *science*, **313**(5786):504–507, 2006.

[HZR16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[HZR16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks." In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016.

[INK05] Shingo Iwano, Tatsuya Nakamura, Yuko Kamioka, and Takeo Ishigaki. "Computer-aided diagnosis: a shape classification of pulmonary nodules imaged by high-resolution CT." *Computerized Medical Imaging and Graphics*, **29**(7):565–570, 2005.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167*, 2015.

[Jac11] Christopher H Jackson. "Multi-state models for panel data: the MSM package for R." *Journal of Statistical Software*, **38**(8):1–29, 2011.

[JBL10] Zhang Jing, Li Bin, and Tian Lianfang. "Lung nodule classification combining rule-based and SVM." In *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on*, pp. 1033–1036. IEEE, 2010.

[JBS16] Jing Jia, Lisa Barbera, and Rinku Sutradhar. "Using Markov Multistate Models to Examine the Progression of Symptom Severity Among an Ambulatory Population of Cancer Patients: Are Certain Symptoms Better Managed Than Others?" *Journal of pain and symptom management*, **51**(2):232–239, 2016.

[JCO15] W Jorritsma, F Cnossen, and PMA van Ooijen. "Improving the radiologist–CAD interaction: designing for appropriate trust." *Clinical radiology*, **70**(2):115–122, 2015.

[JST03] Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. "Multistate Markov models for disease progression with classification error." *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**(2):193–209, 2003.

[JWB16] Huan Jiang, SD Walter, PE Brown, and AM Chiarelli. "Estimation of screening sensitivity and sojourn time from an organized screening program." *Cancer Epidemiology*, **44**:178–185, 2016.

[Kay86] Richard Kay. "A Markov model for analysing cancer markers and disease states in survival studies." *Biometrics*, **44**:855–865, 1986.

[KB01] Jane P Ko and Margrit Betke. "Chest CT: automated nodule detection and assessment of change over time?preliminary experience." *Radiology*, **218**(1):267–273, 2001.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[KBC12] Elizabeth A Krupinski, Kevin S Berbaum, Robert Caldwell, and Kevin M Schartz. "Is Diagnostic Accuracy for Detecting Pulmonary Nodules in Chest CT Reduced after a Long Day of Reading?" In *SPIE Medical Imaging*, pp. 83180X–83180X. International Society for Optics and Photonics, 2012.

[KC15] Aydın Kaya and Ahmet Burak Can. "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics." *Journal of biomedical informatics*, **56**:69–79, 2015.

[KEW12] Seongho Kim, Diane Erwin, and Dongfeng Wu. "Efficacy of dual lung cancer screening by chest X-ray and sputum cytology using Johns Hopkins Lung Project data." *Journal of Biometrics & Biostatistics*, **3**(4), 2012.

[KKC05] Jin-Sung Kim, Jin-Hwan Kim, Gyuseung Cho, and Kyongtae T Bae. "Automated detection of pulmonary nodules on CT images: effect of section thickness and reconstruction interval?initial results." *Radiology*, **236**(1):295–299, 2005.

[KKN03] D-Y Kim, J-H Kim, S-M Noh, and J-W Park. "Pulmonary nodule detection using chest CT images." *Acta Radiologica*, **44**(3):252–257, 2003.

[KL85] JD Kalbfleisch and Jerald F Lawless. "The analysis of panel data under a Markov assumption." *Journal of the American Statistical Association*, **80**(392):863–871, 1985.

[KNO98] Y Kawata, N Niki, H Ohmatsu, R Kakinuma, K Eguchi, M Kaneko, and N Moriyama. "Quantitative surface characterization of pulmonary nodules based on thin-section CT images." *IEEE Transactions on nuclear science*, **45**(4):2132–2138, 1998.

[KNO01] Yoshiki Kawata, Noboru Niki, Hironobu Ohmatsu, Masahiko Kusumoto, Ryutaro Kakinuma, Kensaku Mori, Hiroyuki Nishiyama, Kenji Eguchi, Masahiro Kaneko, and Noriyuki Moriyama. "Computerized analysis of 3-D pulmonary nodule images in surrounding and internal structure feature spaces." In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pp. 889–892. IEEE, 2001.

[KPG15] Hyungjin Kim, Chang Min Park, Jin Mo Goo, Joachim E Wildberger, and Hans-Ulrich Kauczor. "Quantitative computed tomography imaging biomarkers in the diagnosis and management of lung cancer." *Investigative radiology*, **50**(9):571–583, 2015.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[KTS14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[KWC15] Devinder Kumar, Alexander Wong, and David A Clausi. "Lung nodule classification using deep features in CT images." In *Computer and Robot Vision (CRV), 2015 12th Conference on*, pp. 133–138. IEEE, 2015.

[LBB98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, **86**(11):2278–2324, 1998.

[LBD90] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. "Handwritten digit recognition with a back-propagation network." In *Advances in neural information processing systems*, pp. 396–404, 1990.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature*, **521**(7553):436, 2015.

[LCL95] Shih-Chung B Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T Freedman, and Seong K Mun. "Artificial convolution neural network for medical image pattern recognition." *Neural networks*, **8**(7):1201–1214, 1995.

[LD04] Qiang Li and Kunio Doi. "New selective nodule enhancement filter and its application for significant improvement of nodule detection on computed tomography."

In *Medical Imaging 2004: Image Processing*, volume 5370, pp. 1–10. International Society for Optics and Photonics, 2004.

[LHF01] Yongbum Lee, Takeshi Hara, Hiroshi Fujita, Shigeki Itoh, and Takeo Ishigaki. "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique." *IEEE Transactions on medical imaging*, **20**(7):595–604, 2001.

[LHL13] Phen-Lan Lin, Po-Whei Huang, Cheng-Hsiung Lee, and Ming-Ting Wu. "Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model." *Pattern Recognition*, **46**(12):3279–3287, 2013.

[Li07] Qiang Li. "Recent Progress in Computer-aided Diagnosis of Lung Nodules on Thin-section CT." *Computerized Medical Imaging and Graphics*, **31**(4):248–257, 2007.

[Lip16] Zachary C Lipton. "The mythos of model interpretability." *arXiv preprint arXiv:1606.03490*, 2016.

[LKH12] Shu Ling Alycia Lee, Abbas Z Kouzani, and Eric J Hu. "Automated Detection of Lung Nodules in Computed Tomography Images: a Review." *Machine vision and applications*, **23**(1):151–163, 2012.

[LS03] Qiang Li, Shusuke Sone, et al. "Selective enhancement filters for nodules, vessels, and airway walls in two-and three-dimensional CT scans." *Medical physics*, **30**(8):2040–2051, 2003.

[LTB00] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. "WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility." *Statistics and Computing*, **10**(4):325–337, 2000.

[MBB11] Patrick Maisonneuve, Vincenzo Bagnardi, Massimo Bellomi, Lorenzo Spaggiari, Giuseppe Pelosi, Cristiano Rampinelli, Raffaella Bertolotti, Nicole Rotmensz, John K Field, Andrea DeCensi, et al. "Lung cancer risk prediction to select smokers for screening CT?a model based on the Italian COSMOS trial." *Cancer Prevention Research*, **4**(11):1778–1789, 2011.

[MCT16] Junsheng Ma, Wenyaw Chan, and Barbara C Tilley. "Continuous time Markov chain approaches for analyzing transtheoretical models of health behavioral change: A case study and comparison of model estimations." *Statistical methods in medical research*, pp. 1–14, 2016.

[MDC05] Renee Manser, Andrew Dalton, Rob Carter, Graham Byrnes, Mark Elwood, and Donald A Campbell. "Cost-effectiveness analysis of screening for lung cancer with low dose spiral CT (computed tomography) in the Australian setting." *Lung Cancer*, **48**(2):171–185, 2005.

[MDK11] Tomáš Mikolov, Anoop Deoras, Stefan Kombrink, Lukáš Burget, and Jan Černockỳ. "Empirical evaluation and combination of advanced language modeling techniques." In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[MFF03] Parthiv J Mahadevia, Lee A Fleisher, Kevin D Frick, John Eng, Steven N Goodman, and Neil R Powe. "Lung Cancer Screening with Helical Computed Tomography in Older Adult Smokers: a Decision and Cost-effectiveness Analysis." *Jama*, **289**(3):313–322, 2003.

[MFF14] Alexander V Mamonov, Isabel N Figueiredo, Pedro N Figueiredo, and Yen-Hsi Richard Tsai. "Automated Polyp Detection in Colon Capsule Endoscopy." *IEEE transactions on medical imaging*, **33**(7):1488–1502, 2014.

[MHR10] Temesguen Messay, Russell C Hardie, and Steven K Rogers. "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery." *Medical image analysis*, **14**(3):390–406, 2010.

[MHW99] Michael F McNitt-Gray, Eric M Hart, Nathaniel Wyckoff, James W Sayre, Jonathan G Goldin, and Denise R Aberle. "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results." *Medical physics*, **26**(6):880–888, 1999.

[MJ95] Guillermo Marshall and Richard H Jones. "Multi-state models and diabetic retinopathy." *Statistics in Medicine*, **14**(18):1975–1983, 1995.

[MKM04] Colin C McCulloch, Robert A Kaucic, Paulo RS Mendonça, Deborah J Walter, and Ricardo S Avila. "Model-based detection of lung nodules in computed tomography exams1: Thoracic computer-aided diagnosis." *Academic radiology*, **11**(3):258–266, 2004.

[MOH14] Ibrahimu Mdala, Ingar Olsen, Anne D Haffajee, Sigmund S Socransky, Magne Thoresen, and Birgitte Freiesleben Blasio. "Comparing clinical attachment level and pocket depth for predicting periodontal disease progression in healthy sites of patients with chronic periodontitis using multi-state Markov models." *Journal of Clinical Periodontology*, **41**(9):837–845, 2014.

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[MW14] Pierre P Massion and Ronald C Walker. "Indeterminate Pulmonary Nodules: Risk for Having or for Developing Lung Cancer?" *Cancer prevention research*, **7**(12):1173–1178, 2014.

[NNC10] Dustin Newell, Ke Nie, Jeon-Hor Chen, Chieh-Chih Hsu, J Yu Hon, Orhan Nalcioglu, and Min-Ying Su. "Selection of Diagnostic Features on Breast MRI to Differentiate between Malignant and Benign Lesions using Computer-aided Diagnosis: Differences in Lesions Presenting as Mass and Non-mass-like Enhancement." *European radiology*, **20**(4):771–781, 2010.

[NNF13] Andrew Ng, Jiquan Ngiam, Chuan Yu Foo, Yifan Mai, and Caroline Suen. "Unsupervised feature learning and deep learning tutorial." *h ttp://deeplearning. stanford. edu/tutorial*, 2013.

[NRF15] Ron Niehaus, Daniela Stan Raicu, Jacob Furst, and Samuel Armato. "Toward understanding the size dependence of shape features for predicting spiculation in lung nodules for computer-aided diagnosis." *Journal of digital imaging*, **28**(6):704–717, 2015.

[NSN12] Stelmo Magalhães Barros Netto, Aristófanes Corrêa Silva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. "Automatic segmentation of lung nodules with growing neural gas and support vector machine." *Computers in biology and medicine*, **42**(11):1110–1121, 2012.

[NY07] Janne Näppi and Hiroyuki Yoshida. "Fully Automated Three-dimensional Detection of Polyps in Fecal-tagging CT Colonography." *Academic radiology*, **14**(3):287–300, 2007.

[OCR11] Pia Opulencia, David S Channin, Daniela S Raicu, and Jacob D Furst. "Mapping LIDC, RadLex?, and lung nodule image features." *Journal of digital imaging*, **24**(2):256–270, 2011.

[Ots79] Nobuyuki Otsu. "A threshold selection method from gray-level histograms." *IEEE transactions on systems, man, and cybernetics*, **9**(1):62–66, 1979.

[PBR04] David S Paik, Christopher F Beaulieu, Geoffrey D Rubin, Burak Acar, R Brooke Jeffrey, Judy Yee, Joyoni Dey, and Sandy Napel. "Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT." *IEEE transactions on medical imaging*, **23**(6):661–675, 2004.

[PHA16] Panayiotis Petousis, Simon X Han, Denise Aberle, and Alex AT Bui. "Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network." *Artificial Intelligence in Medicine*, **72**:42–55, 2016.

[Pin14] Paul F Pinsky. "Assessing the benefits and harms of low-dose computed tomography screening for lung cancer." *Lung cancer management*, **3**(6):491–498, 2014.

[PRC08] Jiantao Pu, Justus Roos, A Yi Chin, Sandy Napel, Geoffrey D Rubin, and David S Paik. "Adaptive Border Marching Algorithm: Automatic Lung Segmentation on Chest CT Images." *Computerized Medical Imaging and Graphics*, **32**(6):452–462, 2008.

[PZL08] Jiantao Pu, Bin Zheng, Joseph Ken Leader, Xiao-Hui Wang, and David Gur. "An automated CT based lung nodule detection scheme using geometric analysis of signed distance field." *Medical physics*, **35**(8):3453–3461, 2008.

[R C13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[RDF08] Alessandra Retico, Pasquale Delogu, Maria Evelina Fantacci, Ilaria Gori, and A Preite Martinez. "Lung Nodule Detection in Low-dose and Thin-slice Computed Tomography." *Computers in biology and medicine*, **38**(4):525–534, 2008.

[RED09] James C Ross, Raúl San José Estépar, Alejandro Díaz, Carl-Fredrik Westin, Ron Kikinis, Edwin K Silverman, and George R Washko. "Lung extraction, lobe segmentation and hierarchical region assessment for quantitative analysis on high resolution computed tomography images." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 690–698. Springer, 2009.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[RLL16] Holger R Roth, Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald M Summers. "Improving computer-aided detection using convolutional neural networks and random view aggregation." *IEEE transactions on medical imaging*, **35**(5):1170–1181, 2016.

[RLP05] Geoffrey D Rubin, John K Lyo, David S Paik, Anthony J Sherbondy, Lawrence C Chow, Ann N Leung, Robert Mindelzun, Pamela K Schraedley-Desmond, Steven E Zinck, David P Naidich, et al. "Pulmonary nodules on multi–detector row CT scans: performance comparison of radiologists and computer-aided detection." *Radiology*, **234**(1):274–283, 2005.

[RVF09] Daniela S Raicu, Ekarin Varutbangkul, Jacob D Furst, and Samuel G Armato III. "Modelling semantics from image data: opportunities from LIDC." *International Journal of Biomedical Engineering and Technology*, **3**(1-2):83–113, 2009.

[SBC15] Shiwen Shen, Alex AT Bui, Jason Cong, and William Hsu. "An automated lung segmentation approach using bidirectional chain codes to improve nodule detection accuracy." *Computers in biology and medicine*, **57**:139–149, 2015.

[SBH17] Shiwen Shen, Alex A. T. Bui, and William Hsu. "Robust Lung Nodule Classification using 2.5D Convolutional Neural Network." In *AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, November 4-8, 2017*, 2017.

[SCL07] Hui Chuan Shih, Pesus Chou, Chi Ming Liu, and Tao Hsin Tung. "Estimation of progression of multi-state chronic disease using the Markov model and prevalence pool concept." *BMC Medical Informatics and Decision Making*, **7**(1):34, 2007.

[SCL16] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram van Ginneken. "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks." *IEEE transactions on medical imaging*, **35**(5):1160–1169, 2016.

[SHK14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research*, **15**(1):1929–1958, 2014.

[SHP17] Shiwen Shen, Simon X Han, Panayiotis Petousis, Robert E Weiss, Frank Meng, Alex AT Bui, and William Hsu. "A Bayesian model for estimating multi-state disease progression." *Computers in biology and medicine*, **81**:111–120, 2017.

[SJ15] Rebecca Siegel and Ahmedin Jemal. "Cancer facts & figures 2015." *American Cancer Society. Cancer Facts & Figures*, 2015.

[SJG15] Arnaud AA Setio, Colin Jacobs, Jaap Gelderblom, and Bram Ginneken. "Automatic detection of large pulmonary solid nodules in thoracic CT images." *Medical physics*, **42**(10):5642–5653, 2015.

[SL10] Ruslan Salakhutdinov and Hugo Larochelle. "Efficient learning of deep Boltzmann machines." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 693–700, 2010.

[SLX15] Hai Su, Fujun Liu, Yuanpu Xie, Fuyong Xing, Sreenivasan Meyyappan, and Lin Yang. "Region segmentation in histopathological breast cancer images using deep convolutional neural network." In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pp. 55–58. IEEE, 2015.

[SLY11] Frank Seide, Gang Li, and Dong Yu. "Conversational speech transcription using context-dependent deep neural networks." In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[SNM07] Margarida Silveira, Jacinto Nascimento, and Jorge Marques. "Automatic segmentation of the lungs using robust level sets." In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 4414–4417. IEEE, 2007.

[SS13] Heung-Il Suk and Dinggang Shen. "Deep Learning-based Feature Representation for AD/MCI Classification." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 583–590. Springer, 2013.

[Sut13] Ilya Sutskever. "Training recurrent neural networks." *University of Toronto, Toronto, Ont., Canada*, 2013.

[SZ14] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *CoRR*, **abs/1409.1556**, 2014.

[SZY15] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. "Multi-scale convolutional neural networks for lung nodule classification." In *International Conference on Information Processing in Medical Imaging*, pp. 588–599. Springer, 2015.

[SZY17] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification." *Pattern Recognition*, **61**:663–673, 2017.

[TCM16] Sharareh Taghipour, Laurent N Caudrelier, Anthony B Miller, and Bart Harvey. "Using Simulation to Model and Validate Invasive Breast Cancer Progression in Women in the Study and Control Groups of the Canadian National Breast Screening Studies I and II." *Medical Decision Making*, pp. 1–12, 2016.

[TCM17] Sharareh Taghipour, Laurent N Caudrelier, Anthony B Miller, and Bart Harvey. "Using Simulation to Model and Validate Invasive Breast Cancer Progression in Women in the Study and Control Groups of the Canadian National Breast Screening Studies i and ii." *Medical Decision Making*, **37**(2):212–223, 2017.

[te11] A. P. Reeves, A. M. Biancardi. "The Lung Image Database Consortium (LIDC) Nodule Size Report." http://www.via.cornell.edu/lidc/, oct 2011. Release: 2011-10-27-2.

[te17] The Cancer Imaging Archive. "Lung image database consortium - reader annotation and markup - annotation and markup issues/comments." https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI, may 2017.

[Tea11] National Lung Screening Trial Research Team et al. "Reduced lung-cancer mortality with low-dose computed tomographic screening." *N Engl J Med*, **2011**(365):395–409, 2011.

[TKA13] Ahmet Tartar, Niyazi Kilic, and Aydin Akan. "Classification of pulmonary nodules by using hybrid features." *Computational and Mathematical Methods in Medicine*, **2013**, 2013.

[TNK08] Y Toyoda, T Nakayama, Y Kusunoki, H Iso, and T Suzuki. "Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography." *British Journal of Cancer*, **98**(10):1602–1607, 2008.

[TPC11] C Martin Tammemagi, Paul F Pinsky, Neil E Caporaso, Paul A Kvale, William G Hocking, Timothy R Church, Thomas L Riley, John Commins, Martin M Oken, Christine D Berg, et al. "Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation." *Journal of the national cancer institute*, **103**(13):1058–1068, 2011.

[TSJ16] Lindsey A Torre, Rebecca L Siegel, and Ahmedin Jemal. "Lung cancer statistics." In *Lung Cancer and Personalized Medicine*, pp. 1–19. Springer, 2016.

[UHC10] Z Uhry, G Hédelin, M Colonna, B Asselain, P Arveux, A Rogel, C Exbrayat, C Guldenfels, I Courtial, P Soler-Michel, et al. "Multi-state Markov Models in Cancer Screening Evaluation: a Brief Review and Case Study." *Statistical methods in medical research*, **19**(5):463–486, 2010.

[VBA12] PR Varshini, S Baskar, and S Alagappan. "An improved adaptive border marching algorithm for inclusion of juxtapleural nodule in lung segmentation of CT-images." In *Wireless Networks and Computational Intelligence*, pp. 230–235. Springer, 2012.

[WER11] Dongfeng Wu, Diane Erwin, and Gary L Rosner. "Sojourn Time and Lead Time Projection in Lung Cancer Screening." *Lung Cancer*, **72**(3):322–326, 2011.

[WLV08] Harald Weedon-Fekjær, Bo H Lindqvist, Lars J Vatten, Odd O Aalen, and Steinar Tretli. "Estimating mean sojourn time and screening sensitivity using questionnaire data on time since previous screening." *Journal of Medical Screening*, **15**(2):83–90, 2008.

[WRB05] Dongfeng Wu, Gary L Rosner, and Lyle Broemeling. "MLE and Bayesian Inference of Age-Dependent Sensitivity and Transition Probability in Periodic Screening." *Biometrics*, **61**(4):1056–1063, 2005.

[WSC09] Ted W Way, Berkman Sahiner, Heang-Ping Chan, Lubomir Hadjiiski, Philip N Cascade, Aamer Chughtai, Naama Bogot, and Ella Kazerooni. "Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features." *Medical physics*, **36**(7):3086–3098, 2009.

[WSL13] Ying Wei, Guo Shen, and Juan-juan Li. "A fully automatic method for lung parenchyma segmentation and repairing." *Journal of digital imaging*, **26**(3):483–495, 2013.

[WVA05] Harald Weedon-Fekjær, Lars J Vatten, Odd O Aalen, Bo Lindqvist, and Steinar Tretli. "Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results." *Journal of Medical Screening*, **12**(4):172–178, 2005.

[YHS05] Yeny Yim, Helen Hong, and Yeong Gil Shin. "Hybrid lung segmentation in chest CT images for computer-aided diagnosis." In *Enterprise networking and Computing in Healthcare Industry, 2005. HEALTHCOM 2005. Proceedings of 7th International Workshop on*, pp. 378–383. IEEE, 2005.

[YLD09] Xujiong Ye, Xinyu Lin, Jamshid Dehmeshki, Greg Slabaugh, and Gareth Beddoe. "Shape-based computer-aided detection of lung nodules in thoracic CT images." *IEEE Transactions on Biomedical Engineering*, **56**(7):1810–1820, 2009.

[YYC17] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. "Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images." In *AAAI*, pp. 66–72, 2017.

[ZDJ15] Yudong Zhang, Zhengchao Dong, Genlin Ji, and Shuihua Wang. "Effect of Spider-web-plot in MR Brain Image Classification." *Pattern recognition letters*, **62**:14–16, 2015.

[ZDP15] Yudong Zhang, Zhengchao Dong, Preetha Phillips, Shuihua Wang, Genlin Ji, Jiquan Yang, and Ti-Fei Yuan. "Detection of Subjects and Brain Regions Related to Alzheimer's Disease using 3D MRI Scans Based on Eigenbrain and Machine Learning." *Frontiers in Computational Neuroscience*, **9**, 2015.

[ZFF11] Dmitriy Zinovev, Jonathan Feigenbaum, Jacob Furst, and Daniela Raicu. "Probabilistic lung nodule classification with belief decision trees." In *Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE*, pp. 4493–4498. IEEE, 2011.

[ZL86] Scott L Zeger and Kung Yee Liang. "Longitudinal data analysis for discrete and continuous outcomes." *Biometrics*, pp. 121–130, 1986.

[ZTB13] Binsheng Zhao, Yongqiang Tan, Daniel J Bell, Sarah E Marley, Pingzhen Guo, Helen Mann, Marietta LJ Scott, Lawrence H Schwartz, and Dana C Ghiorghiu. "Exploring intra-and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals." *European journal of radiology*, **82**(6):959–968, 2013.

[ZW96] Edith A Zang and Ernst L Wynder. "Differences in lung cancer risk between men and women: examination of the evidence." *Journal of the National Cancer Institute*, **88**(3-4):183–192, 1996.

[ZWD14] Yudong Zhang, Shuihua Wang, and Zhengchao Dong. "Classification of Alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree." *Progress In Electromagnetics Research*, **144**:171–184, 2014.

[ZWH14] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144. ACM, 2014.

[ZWW10] Yudong Zhang, Shuihua Wang, and Lenan Wu. "A Novel Method for Magnetic Resonance Brain Image Classification Based on Adaptive Chaotic PSO." *Progress In Electromagnetics Research*, **109**:325–343, 2010.

[ZWW11] Yudong Zhang, Lenan Wu, and Shuihua Wang. "Magnetic Resonance Brain Image Classification by an Improved Artificial Bee Colony Algorithm." *Progress In Electromagnetics Research*, **116**:65–79, 2011.