Title

Domain Knowledge-Assisted Methods for a Weakly Supervised Task: Automated Diagnosis of Idiopathic Pulmonary Fibrosis Using High Resolution Computed Tomography Scans

Author

Yu, Wenxi

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Domain Knowledge-Assisted Methods for a Weakly Supervised Task: Automated Diagnosis of Idiopathic Pulmonary Fibrosis Using High Resolution Computed Tomography Scans

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

by

Wenxi Yu

2021

ABSTRACT OF THE DISSERTATION

Domain Knowledge-Assisted Methods for a Weakly Supervised Task: Automated Diagnosis of Idiopathic Pulmonary Fibrosis Using High Resolution Computed Tomography Scans

by

Wenxi Yu
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Grace Hyun Jung Kim, Co-Chair
Professor Hua Zhou, Co-Chair

Idiopathic pulmonary fibrosis (IPF) is one type of interstitial lung disease (ILD) of unknown causes. High-resolution computed tomography (HRCT) scans play a crucial role in distinguishing IPF from non-IPF among subjects with ILD. This radiological evaluation is an important yet difficult task. In clinical practice, making a correct and reliable IPF diagnosis is critical to ensure patients with different causes of pulmonary fibrosis be treated appropriately and patients with IPF be assessed for novel therapies and lung transplantation. Therefore, this dissertation aims to build an automated IPF diagnosis system for ILD subjects, using volumetric and non-contrast chest HRCT scans.

Supervised learning methods are a type of machine learning methods that require labels of training samples as ground truth to learn a mapping function between input and output labels. Depending on the type of labels provided, if only coarse-level labels are available, rather than fine-scale labels, this task is called a *weakly supervised task.* For our example, acquiring fine-scale information of certain radiological patterns can be helpful for building the diagnostic system, but fine-scale labels are expensive to obtain. On the other hand, coarse-level labels are usually easier to acquire, such as CT scan-level information. Since we only have labels at a CT scan level, our problem is a weakly supervised task. To tackle this challenge, this dissertation leverages domain knowledge acquired from previous studies,

including IPF progression and quantification information, to provide more efficient, reliable, and explainable diagnostic support.

In project I, we used 2D deep learning models with IPF progression information and optimal design criterion to weigh HRCT samples differently. In project II, 3D deep learning models with multi-scale attention models were used with IPF quantification maps to achieve good model accuracy and explainability. Furthermore, we evaluated the robustness of these developed models under a different set of HRCT parameters, using paired HRCT scans. These proposed methodologies in project I and II can be applied to other weakly supervised tasks, where domain knowledge is available. The method used in the robustness tests can be applied to evaluate model performance if paired medical images are available.

The dissertation of Wenxi Yu is approved.

Weng Kee Wong

Jonathan Goldin

Hua Zhou, Committee Co-Chair

Grace Hyun Jung Kim, Committee Co-Chair

University of California, Los Angeles

2021

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Firstly, I would like to thank my committee co-chairs Dr. Grace Kim and Dr. Hua Zhou for their consistent support and help during the past few years. Dr. Grace Kim introduced me to this field of medical imaging and taught me how to engage in good methodological research for clinical studies. Dr. Kim kindly encouraged me to collaborate with other team members and participate in different data managing projects, which has strengthened my diverse skillset during the PhD training. Her sharp and knowledgeable insights in statistical concepts and clinical applications are the qualities that I have been striving for. In the meantime, I would like to thank Dr. Hua Zhou for always being supportive and sharing his keen insights when approaching methodological problems. Both Dr. Kim and Dr. Zhou care for their students in both study and life, especially during COVID, which I am grateful for. My appreciation goes to Dr. Jonathan Goldin for providing valuable clinical knowledge and support for us and carefully reviewing our manuscripts from clinical perspectives. I would like to thank Dr. Weng Kee Wong for serving in my committee and sharing optimal design knowledge with us, which is a critical element for project I.

Secondly, I would like to express my gratitude to the colleagues from the UCLA Computer vision and imaging biomarker (CVIB) group. Thank you Dr. Michael McNitt-Gray for kindly teaching me CT-related technical information, which is the essential component of Section 4. I appreciate the help from Dr. Pangyu Teng, Dr. Mitchell Murphy, and Dr. Koon-Pong Wong for contributing to the excellent infrastructure in the laboratory and helping me with the scalable software development. Thank you my colleagues Dr. Youngwon Choi, Wasil Wahi-Anwar, and Nastaran Emaminejad for your support and valuable discussions. Thank you Dr. Yu Shi for introducing me to this laboratory - this journey would not begin without your kind help.

I also greatly appreciate the help from Dr. Catherine Sugar for accepting me as a summer research student from the UCLA-CSST program when I was an undergraduate student, with almost zero statistics training. Without her help, I would not discover my interest in biostatistics and would have much harder difficulties when applying to graduate programs.

I would like to thank Dr. James Macinko and Dr. Gang Li for offering me graduate student researcher positions to practice my statistical training during my first and second years at UCLA.

I deeply appreciate the support and help from my parents. Thank you dad for supporting every single decision that I make and always encouraging me to explore the beauty of the world. Thank you, mom, for your unconditional love. There is one sentence that I am always too shy to tell you: I am strong because a strong woman raised me.

My appreciation also goes to my boyfriend Jian, who supports and helps me during the PhD study. His kindness, patience (he helped me practice my entire defense twice!), and intelligence are inspiring to me. Thank you Shuang and Yiyi for being my lifelong friends. Thanks Jane, Leiwen, Ye, Zhenyu, Lei, and Jianchao for our memorable experience at UCLA. Every single one of you means a lot to me and I wish all the best things for you.

Lastly, thank you to my pets Lulu and Mona for bringing so much joy to my life. Having you sleeping under my study desk makes the thesis writing process much more enjoyable.

This research is supported by NIH, NHLBI-R21-HL140465.

Section 2 was adapted from *Wenxi Yu, Hua Zhou, Jonathan G. Goldin, Weng Kee Wong, and Grace Hyun J. Kim. "End-to-end domain knowledge-assisted automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using computed tomography (CT)." Medical Physics (2021).* Section 3 was adapted from *Wenxi Yu, Hua Zhou, Youngwon Choi, Jonathan G. Goldin, Pangyu Teng, Weng Kee Wong, Michael F. McNitt-Gray, Matthew S. Brown, and Grace Hyun J. Kim. "MSGA+RF: Two-stage Deep Learning-based Multi-scale Guided Attention Models to Diagnose Idiopathic Pulmonary Fibrosis from CT Images" (Under review).* Section 4 was adapted from *Wenxi Yu, Michael McNitt-Gray, Jin Woo Song, Jonathan G. Goldin, Hua Zhou, and Grace Hyun J. Kim. "Evaluating the robustness of several high-performing deep learning-based models for idiopathic pulmonary fibrosis (IPF) diagnosis within an interstitial lung disease (ILD) population under different CT imaging protocols" (In preparation).*

VITA

2016 B.S. (Environmental Science), Nanjing University.

2019-2021 Graduate Student Researcher, Department of Radiological Science, University of California, Los Angeles.

2016-2020 Graduate Student Researcher, Department of Biostatistics, University of California, Los Angeles.

PUBLICATIONS

A. Peer-reviewed journals

Wenxi Yu, Hua Zhou, Jonathan G. Goldin, Weng Kee Wong, and Grace Hyun J. Kim. "End-to-end domain knowledge-assisted automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using computed tomography (CT)." Med Phys. 2021 Feb 5. doi: 10.1002/mp.14754. Epub ahead of print. PMID: 33547645.

Grace Hyun J. Kim, Yu Shi, Wenxi Yu, and Weng Kee Wong. "A study design for statistical learning technique to predict radiological progression with an application of idiopathic pulmonary fibrosis using chest CT images." Contemp Clin Trials. 2021 Mar 19;104:106333. doi: 10.1016/j.cct.2021.106333. Epub ahead of print. PMID: 33753286.

B. Conference papers and abstracts

Wenxi Yu, Hua Zhou, Youngwon Choi, Jonathan G. Goldin, and Grace Hyun J. Kim. (2021, April). "MGA-NET: Multi-scale Guided Attention Models for an Automated Diagnosis of Idiopathic Pulmonary Fibrosis". In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1777-1780). IEEE.

Wenxi Yu, Hua Zhou, Youngwon Choi, Jonathan G. Goldin, Pangyu Teng, and Grace Hyun J. Kim. "An automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using domain knowledge-guided attention models in HRCT images." In Medical Imaging 2021: Computer-Aided Diagnosis, vol. 11597, p. 115971Y. International Society for Optics and Photonics, 2021. Awarded as **Medical Imaging 2021 Cum Laude Poster Award.**

Wenxi Yu, Jonathan G. Goldin, Hua Zhou, Youngwon Choi, Pangyu Teng, Matthew S. Brown, and Grace Hyun J. Kim. "Developing 2D and 3D deep learning-based models for an automated diagnosis of idiopathic pulmonary fibrosis (IPF) using chest CT scans." In American Thoracic Society (ATS), 2021. Awarded as **2021 ATS Abstract Scholarship.**

Wenxi Yu, Hua Zhou, Jonathan G. Goldin, and Grace Hyun J. Kim. "Domain Knowledge-Assisted Automatic Diagnosis of Idiopathic Pulmonary Fibrosis (IPF) Using High Resolution Computed Tomography (HRCT)(Student Abstract)." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 10, pp. 13979-13980. 2020.

C. Under review

Wenxi Yu, Hua Zhou, Youngwon Choi, Jonathan G. Goldin, Pangyu Teng, Weng Kee Wong, Michael F. McNitt-Gray, Matthew S. Brown, and Grace Hyun J. Kim. "MSGA+RF: Two-stage Deep Learning-based Multi-scale Guided Attention Models to Diagnose Idiopathic Pulmonary Fibrosis from CT Images" (Under review)

D. In preparation

Wenxi Yu, Michael McNitt-Gray, Jin Woo Song, Jonathan G. Goldin, Hua Zhou, and Grace Hyun J. Kim. "Evaluating the robustness of several high-performing deep learning-based models for idiopathic pulmonary fibrosis (IPF) diagnosis within an interstitial lung disease (ILD) population under different CT imaging protocols." (In preparation)

CHAPTER 1

Introduction

Idiopathic pulmonary fibrosis (IPF) is a specific form of chronic, progressive, irreversible, and usually lethal lung disease of unknown causes [KPS11]. IPF is reported to have an estimated median survival time of 3 to 5 years from the time of first diagnosis [KPS11]. In clinical settings, making a correct and reliable IPF diagnosis is critical to ensure patients with other causes of pulmonary fibrosis be treated appropriately and patients with IPF be assessed for novel therapies and lung transplantation.

According to the official clinical guideline [RRM18], computed tomography (CT) has become an integral part of the diagnosis of IPF. Radiological patterns of usual interstitial phenomena (UIP) are the hallmark of IPF [RRM18]. Specifically, several CT features are frequently observed in UIP patterns, including honeycombing, subpleural reticulation, and traction bronchiectasis in a lower lobe subpleural distribution [RRM18]. Despite the existence of these guidelines, the evaluation of these radiological patterns is a difficult task and largely subject to inter-observer variability [WCS16, WL20].

To this end, this dissertation aims to develop a deep learning-based automated diagnosis system to distinguish IPF from non-IPF among subjects with interstitial lung disease (ILD) based on axial chest CT scans. The clinical meanings of this research area are to (1) reduce inter-observer variability in the IPF diagnosis task, (2) enable timely and reliable IPF diagnosis, and (3) ensure patients with different causes of pulmonary fibrosis be treated appropriately.

In the past few years, several machine learning and deep learning approaches have been developed to provide diagnostic support for IPF. For example, Walsh et al. [WCS18] developed a deep learning system that can automatically classify segmented lung slices into

three radiological patterns: UIP, possible UIP, and inconsistent with UIP. Later, Christe et al. developed a pipeline for UIP diagnosis which involves lung segmentation and voxel-level tissue characterization, such as ground glass opacity, honeycombing, etc [CPD19]. The development and maintenance of these techniques usually involve extensive collaborative efforts from radiologists, imaging analysts, software engineers, data scientists, etc.

These two aforementioned UIP diagnosis models are reported to perform on par with the radiologists [WCS18, CPD19]. At the same time, they reflect two common challenges that researchers frequently encounter when applying deep learning methods in medical imaging applications:

1. Limited availability of fine-level annotations from imaging analysts or radiologists. Building an automated diagnosis tool usually requires labels from domain-specific experts as ground truth. Roughly speaking, for a given task, depending on the type of ground truth labels available, we can describe the labels as *coarse-level* (high-level) or *fine-level* labels. Take our purpose of building IPF diagnosis tool as an example, scan-level labels of whether each CT scan is collected from one IPF or non-IPF subject are coarse-level labels; on the other hand, radiologists' labels of lung abnormalities at a voxel-level or region of interest (RoI) level are fine-scale labels. For the aforementioned UIP diagnosis paper [CPD19], the process of acquiring fine-level (i.e. voxel-level) disease labels is necessary for the model construction, which is usually labor-intensive and time-consuming.

Taking the time and resource limitation into account, this dissertation aims to build a diagnosis model with only coarse-level labels, but not fine-scale labels. When only coarse-level labels are provided, this task is a weakly supervised task with inexact supervision. There are three types of weakly supervised tasks: *inexact supervision*, where only coarse-level labels are provided; *incomplete supervision*, where the ground truth labels are only provided to a proportion of the total samples; *inaccurate supervision*, where the labels may be erroneous [Zho18]. For our IPF diagnosis task, where only scan-level information is available, this belongs to the category of inexact supervision. In particular, we only have clinical information of whether the CT scan is collected from an IPF or non-IPF subject, without other fine-scale (such as pixel-level) information, such as whether lung abnormalities

exist and the locations of where they are in the CT scan.

2. Lack of explainability. Deep learning models are often criticized for being unexplainable ("black box nature"), causing doubts and suspicions among healthcare professionals. Walsh et al. mentioned that "Developing better methods for visualizing the inner workings of deep neural networks will be important if this technology is to be integrated into clinical practice." [WCS18]

Taking these two challenges into account, our work aims to build an *efficient, explainable, and domain knowledge-assisted* IPF diagnosis model. For the first challenge of limited information, we propose to bring in population-level domain knowledge, which is easier to acquire as opposed to pixel-level labels, to assist the IPF diagnosis task. In this dissertation, population-level domain knowledge was acquired using two well-developed techniques: one is a machine learning model that can predict whether a CT voxel suggests progression or not, for IPF subjects; the other one is an automated algorithm to quantify the extent of fibrotic patterns for segmented lungs.

For the second challenge of lacking of explainability, project I and project II use post-hoc explanations and trainable attention mechanisms to shed light on how deep learning models function, respectively.

We begin the introduction by discussing the history and background of data science in Section 1.1. Medical imaging and the current advancement of deep learning approaches are discussed in Section 1.2. Since the major focus of this dissertation is about using CT scans for IPF diagnosis, Section 1.3 and Section 1.4 introduces CT and IPF, respectively. Lastly, the aim and novelty are provided in Section 1.5.

1.1 Data Science

At a high level, data science is an interdisciplinary field that supports and guides the extraction of generalizable knowledge from data [PF13].

This term data science can be traced back to a few decades ago. In 2001, Dr. William

S. Cleveland published a paper "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." [Cle01] The author devised a plan to "to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called *data science*."

In recent years, data science has gained popularity and many institutions offer data science as a major. Why do people use this new term *data science*, other than statistics, which has been used for centuries? The increase in the amount of data available itself may not be a sufficient answer. Dhar [Dha13] offered two key differences in the recent development of data science and traditional statistics: (1) data science concentrates on the increasingly heterogeneous and unstructured data. For example, for an online advertising company, the available information collected from text, image, and video, complicates the decision making process. (2) data science emphasizes the predictive power of models. That is, the model performance on the data that will be collected in the future is an essential consideration of data science.

Our work lies in the field of data science because we use medical imaging data, which is especially high-dimensional and heterogeneous, to extract useful disease-specific information, that can generalize to broad clinical applications. Notably, these two aforementioned key characteristics are commonly observed in the field of medical imaging, including this dissertation, since (1) imaging data collected at multiple centers with multiple scanner machines, are, by all means, heterogeneous; (2) the predictive power of the model, i.e. whether certain knowledge extracted from a set of subjects could be extended to other subjects, who share similar characteristics of the sets of subjects in the training data, is very critical for clinical evaluation and furthermore clinical deployment.

1.2 Machine learning in medical imaging

Medical imaging is a series of techniques that use images to capture the interior structure and composition of the body, for the purpose of diagnosing, evaluating disease progression or treatment efficacy. There are several medical imaging modalities, including computed

tomography (CT), magnetic resonance imaging (MRI), ultrasound, etc.

1.2.1 Imaging biomarkers

The flourishing development of medical imaging in the past several decades has created numerous opportunities and challenges for the field of developing quantitative imaging systems. According to the definition from the U.S. Food and Drug Administration (FDA), biomarkers are defined as "a characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions [FDA19]".

A quantitative imaging biomarker is defined as an image-based quantifiable characteristic that can be used for disease diagnosis, prognosis, or measuring a clinical response to a certain intervention. As the need for medical imaging increases, quantitative imaging systems prospered. Recognizing the urgent need for reliable and reproducible quantification of biomedical imaging data, the Radiological Society of North America organized the Quantitative Imaging Biomarkers Alliance (QIBA) in 2007. The aim of QIBA is to "improve the value and practicality of quantitative imaging biomarkers by reducing variability across devices, sites, patients and time [RSN21]". In other words, a successful imaging biomarker needs to be defined under certain standardized technical and clinical contexts to ensure a consistent and reliable measurement [MGA15].

The development of imaging biomarkers usually requires domain-specific expert knowledge. For example, CT texture features within certain regions of interest, such as entropy and uniformity of a tumor in patients with metastatic renal cell cancer, reflect the tumor heterogeneity and are found to be associated with disease progression [GGN11]. The development of these texture features is based on the knowledge that heterogeneity is a well-recognized sign of malignancy and poor prognosis in tumor studies [GGN11, DYL12].

1.2.2 Deep learning applications in medical imaging

Deep learning is a subset of machine learning methods that are based on neural networks to extract useful information from data. It has the general principle of learning *multiple levels of composition* and the word "deep" refers to the use of multiple layers [GBC16].

With the increasing availability of imaging data and computing power, the development of deep learning-based methods in medical imaging domains has thrived in the past decades [LJC17, SWS17]. Deep learning requires limited image processing and incorporates the feature engineering step in a machine-trainable manner. Therefore, the burden of extracting meaningful imaging features shifted from human experts to computers, providing great opportunities for researchers with limited clinical knowledge to contribute to the development of meaningful clinical biomarkers [SWS17].

However, this is not to say that deep learning is magic that solves every medical imaging problem at no cost. Admittedly, the successful development of deep learning approaches requires extensive architecture design, hyperparameter tuning, and monitoring, which are usually based on trial and error. Deep learning approaches also suffer from the critics of lacking explainability and unknown generalizability to unseen domains [ZBL18]. For example, Lehman et al. reviewed the diagnostic accuracy of an FDA-approved and commonly used computer-aided detection (CAD) for mammography that assisted radiologists to detect subtle cancers on a number of 324k women [LWB15]. Although early studies supported the high sensitivity of CAD in laboratory-based environments [JNS99] and the cost of applying CAD was approximately over $400 million per year in US health care expenditures (as of 2015), CAD did not improve diagnostic accuracy [LWB15].

1.2.3 Concepts in deep learning

We clarify several frequently used concepts in the deep learning domain in this section.

2D versus 3D convolutions: An illustrative figure of 2D and 3D convolutions is shown in Figure 1.1. Convolution operations are essentially elementwise multiplication, using the

same set of parameters sliding over the entire region of the input. Suppose we take a 3D image tensor $X_{in} \in \mathbb{R}^{H_{in} \times W_{in} \times D_{in}}$ as input, for 2D convolutions, we use a series of 2D kernels of size $H_K \times W_K$. A filter is defined as a collection of kernels. By convention, the number of kernels for each filter must match with the depth dimension of the input feature maps (both are D_{in} shown in Figure 1.1). Then, for one fixed filter, each kernel is applied to one slice (or channel) of the input feature map (dimension: $H_{in} \times D_{in}$) and the output is later summed up together across D_{in} number of kernels. After adding the bias term, this filter produces one channel of the output feature map. The number of channels in the output feature maps (C) is decided by the number of filters.

For 3D convolutions, three-dimensional kernels ($H_K \times W_K \times D_K$) are used. Each filter is applied to the input feature maps independently, producing one channel of the output feature map. Similar to 2D convolutions, the number of channels in the output feature maps (C) is decided based on the number of filters.

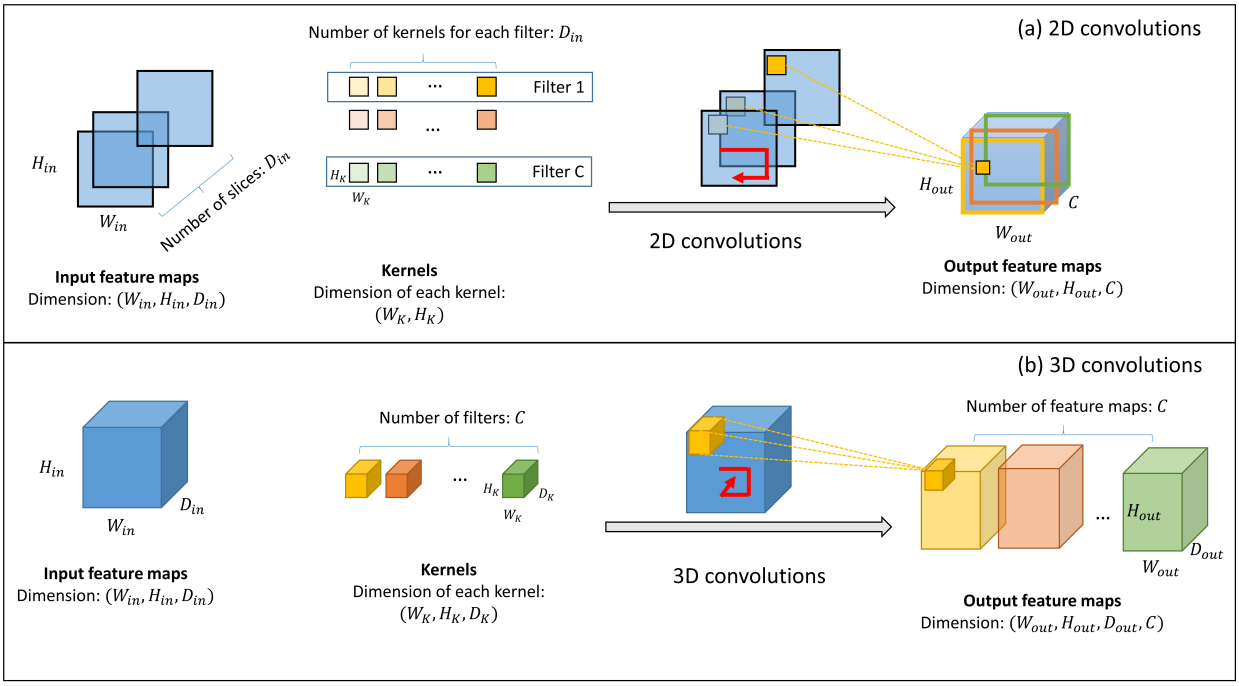


Figure 1.1: An illustrative figure of 2D and 3D convolutions.

In summary, 2D and 3D convolutions differ in terms of how convolutions operate. For 2D

convolutions, the entire input volume is first separated into isolated slices and each kernel operates on one specific slice. Thus, there is no direct parameter sharing across the depth dimension. With respect to 3D convolutions, the entire input feature maps are fed into the system and the same parameters (same kernels) are applied across the depth dimension.

Weights: Parameter within a neural network. For example, for a convolutional layer, kernels and the bias terms are weights.

Layer: A layer is a high-level building block in deep learning, which involves one type of operation. Commonly used layers include *convolutional layer* that applies the convolution operation to the input and passes the results to the next layer, *pooling layer* which reduces the dimension size of the intermediate feature maps while preserving the features with locally large values, *dropout layer* which randomly sets a proportion of deep learning weights to zero during training for the purpose of preventing overfitting, etc.

Feature maps: The output of certain layers, which is a certain representation of the input image.

1.2.4 Explainable artificial intelligence (AI)

Explainable AI is a system that can explain how and why a decision has been made in human-understandable representations. Explainable AI may be beneficial for the following reasons:

Reduce confounding effects. Without building explainable AI, the system may exploit confounding factors, such as imaging protocols, hospital information, etc., to achieve a good model performance. Explainable AI may serve as an important step for model diagnostics and warn the researchers when these things happen. For example, Zech et al. [ZBL18] found that CNN learned to detect a metal token that a technician placed on the patient when the X-ray image was taken, using an activation heatmap. Without realizing this, the system may perform extremely well by leveraging confounding factors using the training cases, but it is prone to fail to generalize to other studies.

Build trust. In medical domains, explainable AI is a critical step for building trust

among clinicians, patients, and healthcare professionals [WCS18].

Manage responsibilities. Explainable AI is important for safety-critical applications, such as self-driving cars or medical domains. For example, in 2018, a self-driving Uber car killed a pedestrian in Arizona [Gua18]. This tragedy reminds the public that blindly believing in deep learning systems is dangerous. Specifically, having explainable models may help us understand the decision process of the model, prevent similar tragedies to happen again, and manage responsibilities among stakeholders.

Consequently, building an IPF diagnosis model that is both accurate and explainable is the main goal of this dissertation. In project I, we implemented a well-developed gradient-based class activation mapping (Grad-CAM, [SCD17]) to visualize the important regions for disease diagnosis. This post-hoc method provides case-specific and class-dominant explanations after the model is trained. In project II, we built an explainable model with guided information to encourage the network to focus on specific regions. The goal of enhancing explainability was integrated into the training of the model, in an end-to-end manner.

1.3 Computed tomography (CT) and image processing

1.3.1 Introduction of CT

Computed tomography (CT) is a commonly used medical imaging device invented and developed in the 1970s. It uses a single X-ray source and multiple radiation detectors which rotate around the object. The word "tomography" is a combination of two Greek words: *tomos*(slice) and *graphein*(draw). Nowadays, it is widely used for preventive or diagnostic clinical purposes in the head, neck, lungs, and many other domains.

Typically, certain image reconstruction methods are needed to reconstruct the object from its raw X-ray projections and produce multiple cross-sectional images of the object as the output of the CT scan. X-rays are attenuated to different extents when passing through different components of the objects. As a result, the reconstructed CT scan is a volume of pixels, where each pixel represents the attenuation value μ, which shows different

radiodensity at varying locations of the object. Normally, CT values are defined by a linear transformation of the attenuation values μ, which are calibrated with reference to water:

$$\text{CT value} = \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}} * 1000, \tag{1.1}$$

where μ_{water} is the attenuation coefficient of water.

CT values are unitless numbers. In honor of the inventor of CT, Godfrey Hounsfield, the unit of CT values is called the Hounsfield unit (HU). CT values typically range from -1000 to 3000 HU. By definition, CT values of water are zero, air is -1000 HU, normal lungs usually lie in the range of -900 HU to -400 HU, and abnormal lungs usually fall in the range of -1000 HU to -200 HU. For better visibility, clinicians usually select different ranges of HU to view different tissues. For example, a window of [-1000 HU, 250 HU] is usually used to view lung parenchyma, which is called as "lung window". By setting a lung window of [-1000 HU, 250 HU], any CT values greater than 250 HU are set to be 250 HU; any CT values below -1000 HU (if any) are set as -1000 HU.

1.3.2 Introduction of DICOM images

Digital Imaging and Communications in Medicine (DICOM) is a commonly used imaging standard for storing and transferring CT images. A DICOM file is composed of both a header file that contains the information about the CT scan (such as slice thickness, dose level, etc.) and a 2D array that stores data for one image slice. Normally, a CT scan is composed of multiple CT slices and each CT slice is stored as one DICOM file. According to the DICOM protocol, image data is saved as grayscale values (usually ranges from 0 to 255). Grayscale values can be transformed into CT values (unit: HU) by the following linear transformation:

$$\text{CT value} = \text{Gray value} * \text{Slope} + \text{Intercept}, \tag{1.2}$$

where the slope and intercept can be extracted from the DICOM header file under tags *Rescale Slope* and *Rescale Intercept*.

1.3.3 Image processing in this dissertation

Starting with CT images of DICOM format, we applied a series of image processing steps as follows. The accompanying figures after each processing step, using one non-IPF ILD scan as an example, are provided in the Supplementary Figure A.1. In this thesis, we refer the axial (cross-sectional) plane as XY plane and the Z-dimension represents craniocaudal direction from apex to base of the lungs.

Step 1: Converted grayscale values to Hounsfield Units and created a lung window of (-1250 HU, 250 HU) Hounsfield units (HU) for better visibility of lung parenchyma. Specifically, HU values below -1250 HU (or above 250 HU) were set to -1250 HU (or 250 HU).

Step 2: Aligned patient positions to be supine. We checked the DICOM image header "ImageOrientationPatient": if the CT was scanned under the prone position, we rotated the image 180 degrees to match supine positions.

Step 3: In project II, we added a step of isotropic resampling to a uniform cube of size 1mm $\times$ 1mm $\times$ 1mm using cubic spline interpolation. Pixel spacings represent the row-wise and column-wise physical distances between the center of each pixel along the axial plane, which can be checked from DICOM header files "PixelSpacing". Z-dimension spacing, which represents the distance between each adjacent CT slice along the Z-dimension, was calculated from the DICOM header file "ImagePositionPatient". If step 3 is added, then non-volumetric scans, where the Z-spacing is not consistent across the entire scan, are excluded.

Step 4: Automatically crop each CT slice based on the presence of the patient's body by canny edge detector using Python library of scikit-image. We note that after automated cropping, the image dimensions are varied for each CT slice.

Step 5: To make each CT slice with uniform image dimension, we center-cut or pad the cropped image to the same dimension ($256 \times 256 \times 128$).

Step 6: We rescaled the image values x_i at image location i to a range of [0, 1] on a scan

level via:

$$\hat{x}_i = \frac{x_i - \min_i x_i}{\max_i x_i - \min_i x_i}, \tag{1.3}$$

where $\hat{x}_i$ is the rescaled image values of x_i.

Step 7: In project II, to boost sample size, we further resampled a certain number of CT samples ("bootstrapped samples") along the Z-dimension from apex to base. For each bootstrap sample, along the Z-dimension, we randomly sampled 64 out of in total 128 slices. Along the axial dimension, each slice was resized from 256×256 from 128×128 using cubic spline interpolation. We named the number of CT samples per scan as M; sensitivity analysis results with a selection of M were shown in Figure 3.4.

1.4 Idiopathic pulmonary fibrosis (IPF)

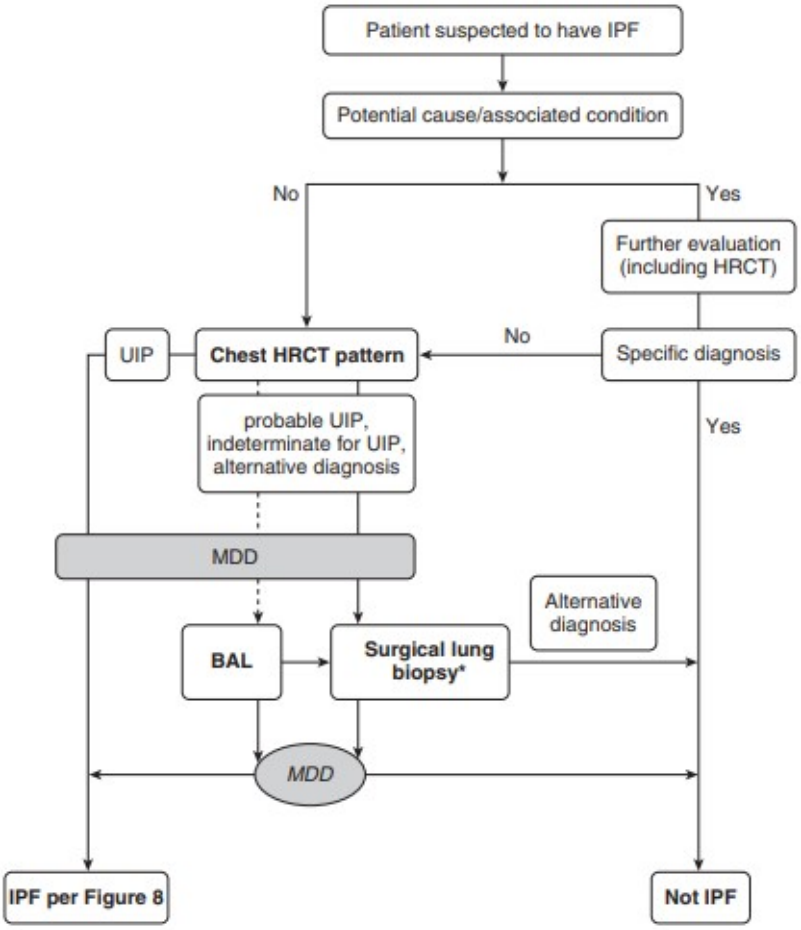
Idiopathic pulmonary fibrosis (IPF) is defined as a specific form of chronic, progressive fibrosing interstitial pneumonia of unknown causes. IPF is limited to the lungs and usually occurs in older adults [RCE11]. It is a rare disease with irreversible and unpredictable progression and survival [RCE11]. The prevalence estimates of IPF in the USA varied between 14 and 27.9 cases per 100,000 in the population [NCR12]. The median survival time ranges from 2 to 5 years, but some patients live much longer [RCE11, NCR12, RRM18].

IPF is associated with histopathologic and/or radiologic pattern of usual interstitial pneumonia (UIP) [RRM18]. Chest CT images are used to determine the presence of the UIP pattern. UIP pattern is associated with some common CT representations, including honeycombing, ground glass opacity, reticular pattern with peripheral traction bronchiectasis or bronchiolectasis, etc [RRM18]. Notably, these CT features usually occur in the subpleural and basal areas.

The diagnosis of IPF involves the collaboration of multi-disciplinary discussion (MDD) from specialists: clinicians, radiologists, and pathologists. The up-to-date clinical practice guideline for IPF, published in 2018 [RRM18], provides a detailed explanation and flowchart regarding the overall diagnostic workflow (see Figure 1.2). In more detail, patients suspected

to have IPF should undergo an in-depth evaluation of potential causes or associated conditions, such as hypersensitivity pneumonitis, connective tissue disease, etc. If there is no potential cause identified, the chest HRCT patterns of the patient is evaluated during an MDD. The patient is diagnosed with IPF if certain combinations of HRCT patterns and histopathological patterns (if applicable) are present.

According to the guideline, CT assessment has become a cornerstone in the diagnosis of IPF, for subjects with unknown clinical causes or associated conditions. However, using CT evaluation for IPF diagnosis is a difficult task and subject to inter-observer variability, even for experienced radiologists [WCS16, WLR19]. Developing an automated diagnosis of IPF using CT can be helpful for a prototype of this task or a pre-screening tool.



MDD, multidisciplinary discussion; UIP, usual interstitial pneumonia; BAL, bronchoalveolar lavage, which is a useful adjunct to lung biopsy.

Figure 1.2: Diagnostic workflow of IPF [RRM18].

Additionally, as shown in Figure 1.2, in some cases where a definite diagnosis of IPF could not be made, surgical lung biopsy is suggested [RRM18]. However, surgical lung biopsy is also known to be associated with an increasing risk of in-hospitalization or mortality [HFM16]. In this context, investigating automated CT evaluation for IPF diagnosis may potentially reduce the need for lung biopsy in the long run.

In recent years, several anti-fibrotic treatments have been found to reduce the decline in lung function in patients with IPF [ANT05, NAB11]. The successful development of these anti-fibrotic treatments further necessitates the urgent needs of developing automated and

reliable IPF diagnosis tools. This is because IPF diagnosis is a critical step for designing the inclusion and exclusion criteria when conducting clinical trial designs.

To summarize, there are three potential clinical meanings of this work: (1) it facilitates automatic diagnosis of IPF that saves time and reduces inter-observer variability; (2) it enables early diagnosis and treatment, which may lead to early anti-fibrotic treatment and increase the likelihood of a slow disease progression; and (3) it potentially reduces the need for lung biopsy in the diagnostic process. The latter is an important consideration since lung biopsy is associated with increased in-hospital mortality.

1.4.1 Population-level domain knowledge (DK)

IPF is a disease of a highly progressive and unpredictable nature. The heterogeneous rate of progression hampers the process of efficient drug development. Developing reliable imaging biomarkers are indispensable for assessing the disease severity and evaluating the efficacy of anti-fibrotic drugs. Compared with other commonly accepted clinical outcomes in IPF studies, such as forced vital capacity (FVC), quantitative imaging biomarkers are more rapid, objective, reproducible, traceable, and are less prone to missing data.

UCLA Computer Vision and Imaging Biomarkers (CVIB) group has been concentrating in the field of interstitial lung disease, including IPF, for decades. Given the available resources from well-developed imaging biomarker tools, including voxel-wise progression prediction and quantification information for IPF subjects, we can acquire DK from previous studies. By leveraging DK at a population level, we hope to provide more knowledge/guidance to the constructed IPF diagnosis models.

Progressive trends across the lungs (IPF progression, 1D information) and disease quantification maps (IPF quantification, 3D information), both on a population level, were incorporated as the DK for the development of project I and project II, respectively. For project I, IPF progressive trends were utilized to judiciously sample CT slices that capture the IPF disease information; for project II, disease quantification maps were included to encourage the model to focus on the regions of interest. We will discuss the acquirement of DK in the

next section.

1. IPF Progression: Progressive trends from apex to base of the lungs

Previous studies used quantum particle swarm optimization incorporated with a resampling technique and a random forest method to predict the pixel-level IPF progression status (i.e. whether the pixel of the segmented CT lung image suggests progressive or not progressive) [SWG19]. CT scans from a total number of of $N = 122$ patients with IPF underwent an automated lung segmentation pipeline and the methodology was applied to predict voxel-level disease progression [SWG19]. Using the predictive results on a CT slice-level, we deduced (1) the number of segmented lung area voxels; (2) among these segmented lung area voxels, the percentage of voxels that are classified as progressive.

We observe that, on a population-level, CT slices that contain more percentage of progressive voxels usually appear in the bottom of the lungs (more details are provided in Figure 2.3). This is consistent with the radiological findings that IPF characteristics are usually predominant in the lower lungs [RRM18]. This IPF progression information across the lung positions can provide guidance for a better sampling strategy, which will be discussed in project I.

2. IPF Quantification: Disease severity maps

Quantitative lung fibrosis (QLF) is a texture-based scores of disease extent calculated from a classification model, based on chest CT scans [KTC10]. QLF score is calculated by the percentage of voxels that are classified as *fibrotic reticulation* patterns [KBC15]. QLF can be calculated as follows: CT images underwent a series of processing steps, including (1) denoise CT, (2) sample voxels within the 4-by-4 grid from the segmented lung boundary, (3) calculate texture features from each grid sampled pixel or voxel, (4) run a support vector machine classifier to classify each sample voxel based on its texture features, (5) count the proportion of voxels which were classified as abnormalities among total samples to get the QLF score. In recent years, QLF scores have been clinically applied to multiple clinical trials for subjects with interstitial lung disease. These biomarkers can provide objective surrogate measures for treatment efficacy evaluation [KBE11, LGT20] and can provide prediction of

clinical progression in subjects with IPF [KWB20].

In this study, we acquired the voxel-wise predictions using the aforementioned technique from a total number of $N = 102$ subjects with IPF. After resizing the CT scans to a uniform dimension, we calculated the marginal probability of getting lung fibrosis (LF, fibrotic reticulation) and other lung fibrosis (OLF, pulmonary fibrosis with similar texture features of vascularity from textural images out of the decomposed images) for these $N = 102$ subjects, for each CT voxel location. By definition, the sum of LF and OLF is the extent of QLF scores. This marginal probability map can serve as population-level guidance on where the disease patterns usually locate, especially for subjects with IPF. In project II, this disease map information was incorporated in the training process of the IPF diagnosis model.

1.5 Aims and novelty

The ultimate goal of this line of research is to develop an automated and reliable tool that can distinguish subjects with IPF from non-IPF among subjects with ILD, using axial chest CT scans.

To this end, the dissertation includes two projects that aim to tackle this problem from different perspectives:

For project I, we built a 2D deep learning-based IPF diagnosis model. We incorporated an optimal design criterion to train the diagnostic model in an end-to-end manner. IPF progression trends across the lung position were used to judiciously sample CT slices.

For project II, we constructed an explainable 3D deep learning-based IPF diagnosis model. Attention models and population-level disease severity maps were included to encourage the network to focus on specific regions of interest.

We summarize the novelty and its corresponding clinical context in this work from these four perspectives:

1. **Domain knowledge-assisted:** We bring in clinical knowledge (i.e. IPF progression and IPF quantification) to this IPF diagnosis task to provide extra information/guidance to

the model.

2. **End-to-end training:** For both project I and project II, domain knowledge is incorporated into the training of the deep learning models in an end-to-end manner. The developed IPF diagnosis models can be easily implemented in a clinical workflow, with no extra lung segmentation needed.

3. **Explainable:** Provide visual explanations on how models make the classification decisions using deep learning-extracted features.

4. **Innovative clinical paradigm:** Current clinical practice relies on pre-defined visual patterns (such as honeycombing, ground glass opacity, etc.) to distinguish IPF. On the other hand, this work is a preliminary and innovative attempt to change the current paradigm of IPF diagnosis by obviating the need for examining visual patterns. Notably, this is a preliminary attempt and further clinical studies are needed.

Clinically, providing automatic IPF diagnosis support is timely and meaningful because the proposed method (1) facilitates automated IPF diagnosis and reduces inter- and intra-reader disagreement; (2) enables early anti-fibrotic treatment and so may prolong patient's survival time; (3) decreases the likelihood of requiring of lung biopsy in the long run and its attendant's risks.

The rest of the dissertation is organized as follows. Chapter 2 and chapter 3 describe project I and project II, respectively. Chapter 4 includes a series of robustness tests which used paired CT images to evaluate the robustness of models constructed from project I and II.

CHAPTER 2

Project I: A domain knowledge-assisted 2D-CNN network

2.1 Background

In recent years, there have been growing research interests in developing automated diagnostic support for patients with interstitial lung disease, using machine learning and deep learning approaches [ACE16, WCS18, CPD19, AKK20].

Image patch-level tissue characterization: Anthimopoulos et al. developed 2D-CNN architectures that took image patches (size 32×32) sampled from human-contoured lung regions as input and produced a label of ground glass opacity, reticulation, honeycombing, etc [ACE16].

Scan-level UIP pattern determination: Patient-level UIP diagnosis has recently gained much attention to provide diagnostic support for patients with fibrotic lung disease. This workflow classifies patients into three categories based on CT scans: UIP, possible UIP or inconsistent with UIP. Walsh et al. developed deep learning tools which showed comparable performance when patients were diagnosed by radiologists [WCS18]. Similarly, Christe et al. developed a pipeline for the automatic classification of CT images to certain UIP patterns [CPD19]. The diagnostic pipeline involves lung segmentation and voxel-level tissue characterization, such as ground glass opacity, honeycombing, etc.

For these aforementioned research efforts, the development and maintenance of these techniques usually involve extensive time and effort. This includes building automated lung segmentation tools, reviewing lung segmentation results, labeling tissue characterization, and

UIP pattern determination. Taking these time and resource considerations into account, our work aims to concentrate on the following perspectives:

1. *Efficiency:* Build an efficient and automated system that does not require lung segmentation or pixel-level tissue labels, by leveraging information provided by domain knowledge. Population-level domain knowledge, which was obtained from previous studies, is more time-efficient to acquire compared with having new images labeled.

2. *Explainability:* Although it has been widely accepted that enhancing explainability of deep learning models is critical for clinical practice [WCS18], current work in this area does not emphasize the explainability of models. With this goal in mind, our work aims to account for the classification result (project I) and enhance (project II) explainability for building IPF diagnosis models.

In conclusion, using CT scans to automatically diagnose IPF is limited so far and we believe our proposed methods from both project I and II can have a potential impact on patient-level classification of IPF, from efficiency and explainability point of view.

2.2 Materials and methods

2.2.1 Datatsets

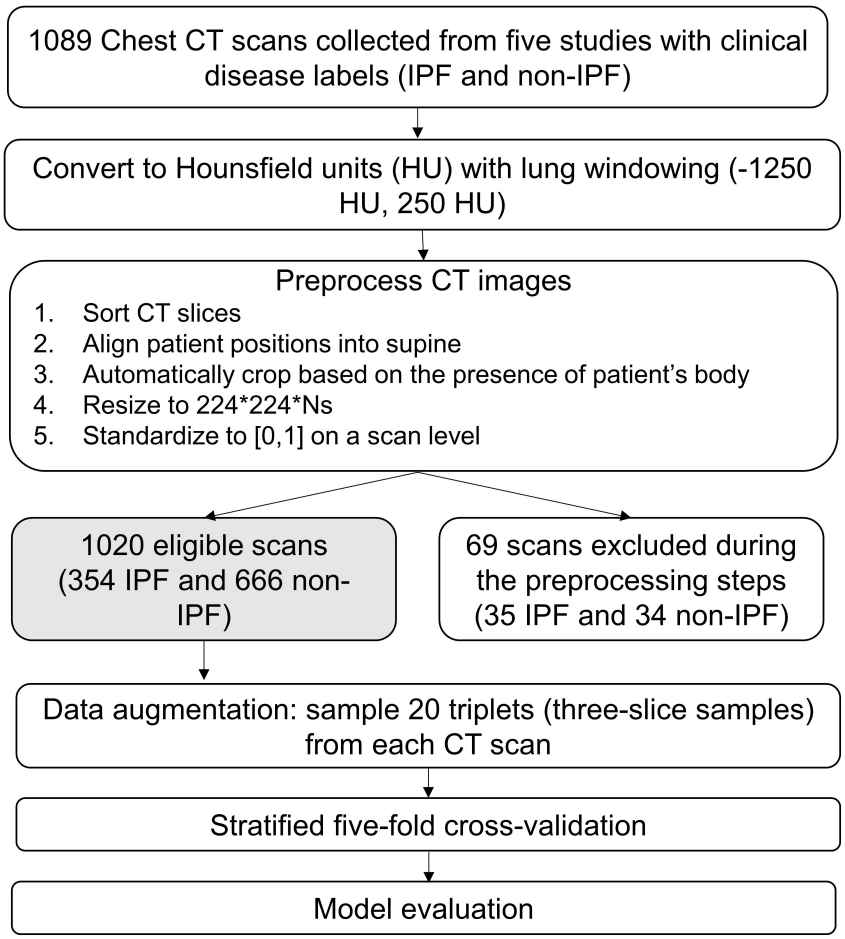
Axial lung CT scans were retrospectively acquired from five multi-center studies, including two IPF studies and three non-IPF studies. The inclusion criterion is that each patient has been clinically diagnosed as interstitial lung diseases. CT scans with IPF diagnosis were confirmed by multidisciplinary clinical teams [RCE11, RRM18]. CT images of IPF patients were collected from December 2004 to July 2016; CT images of non-IPF patients were collected from May 1997 to May 2018. For each patient, only the first available total lung capacity (TLC) scans are used for the algorithm development and testing. In total, there are 1089 patients, including 389 IPF and 700 non-IPF patients, collectively obtained from the five multi-center studies. CT images were acquired under different CT scanners and protocols, which are summarized in the Supplementary Table B.1.

Table 2.1: Basic clinical information of the CT scans.

Study	Type	Disease diagnosis	Number of subjects	Number of CT slices per visit (mean ± SE)
1	IPF	IPF	245	359 ± 106
2	IPF	IPF	144	280 ± 46
3	Non-IPF	RAILD, SjS-ILD, SSc-ILD and HP	449	53 ± 25
4	Non-IPF	Myositis ILD	81	253 ± 75
5	Non-IPF	SSc-ILD	170	106 ± 83

Note: SE, standard error. RAILD, rheumatoid arthritis-associated ILD; SjS-ILD, Sjögren's syndrome-associated ILD; SSc-ILD, Systemic sclerosis-associated ILD; HP, hypersensitivity pneumonitis.

Figure 2.1 shows the data flow of image preprocessing and model construction. Table 2.1 summarizes the disease diagnosis, the number of subjects, and the number of CT slices per visit for the five cohorts with study 1 and 2 involving IPF patients, and study 3, 4, and 5 involving non-IPF ILD patients. CT scans from study 1 and 2 were confirmed as IPF with the IPF diagnostic criteria [RCE11, RRM18]. CT scans from study 3, 4, and 5 were clinically confirmed as other ILD diseases. We note that some scans (13.3%, $N = 60$) from study 3 are non-volumetric scans, where the spacing between each adjacent CT slice along the z-dimension is not consistent. As a result, the average number of CT slices in study 3 is fewer than that of other studies.



Ns: the number of CT slices for each scan, which varies for each scan.

Figure 2.1: Data flow of image preparation and model construction for project I.

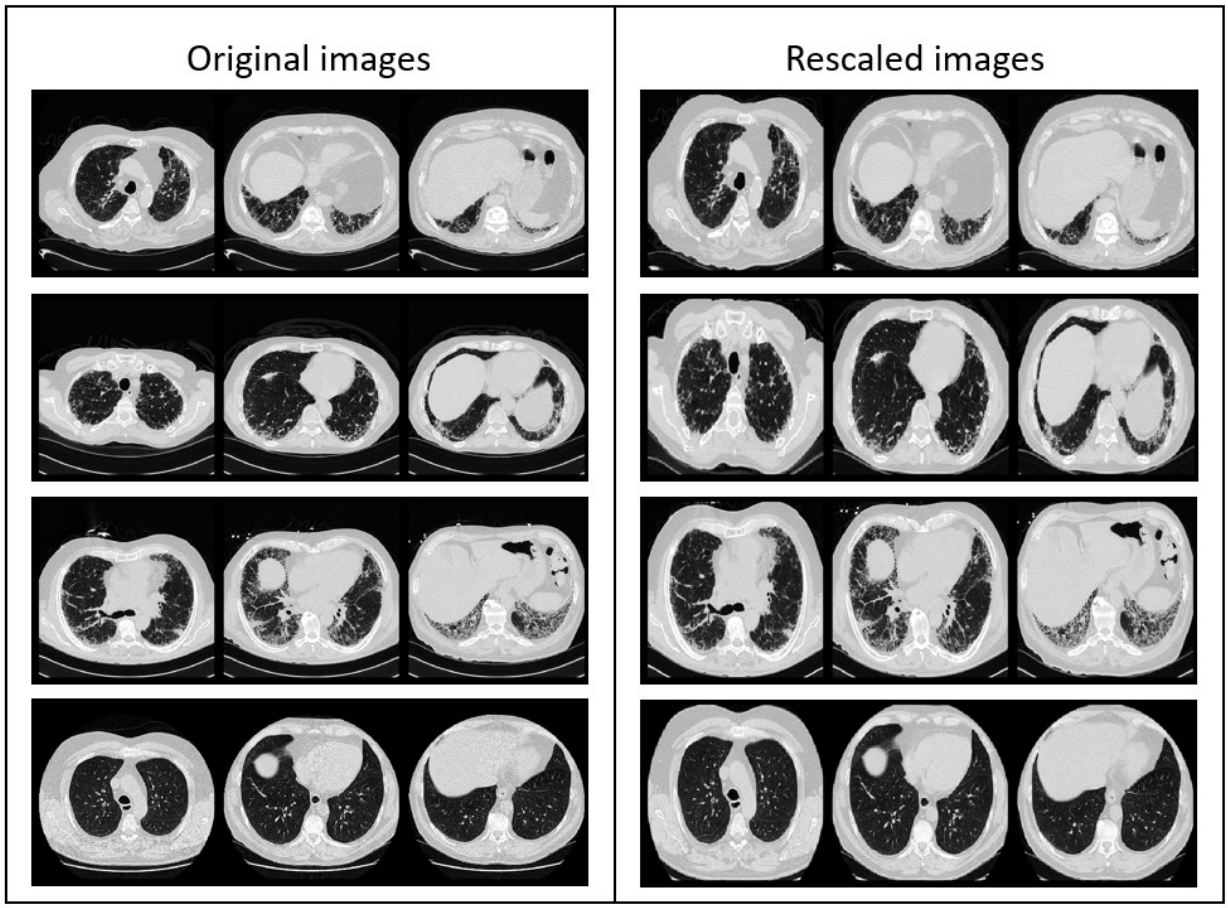
2.2.2 Problem statement

Our main research problem is a binary classification task to determine whether a CT scan is collected from an IPF subject or not. The model input is the axial lung CT images of one CT scan, which are usually of dimension 512 $\times$ 512 $\times$ Number of CT slices. Here 512 is the image resolution and the number of slices usually varies from different CT scans. The output is a binary label $y_i \in \{0, 1\}$ indicating whether the CT scan is from a subject i with IPF or not, $i = 1, 2, .., N$. Further clinical information, such as gender and age, cannot be retrieved due to the anonymization process, and thus is not provided for the automatic diagnosis system.

To reduce dimension size and boost sample size, we use three CT slices, which is refered to as *triplets* in this chapter, as one training and testing unit. We use $i = 1, ..., N$ as the subject index, $j = 1, ..., M$ as the sample (triplet) index, and M is the number of triplets sampled from one CT scan. $X_{ij} \in \mathbb{R}^{224\times224\times3}$ is a three-dimensional tensor of the processed CT triplet from subject i and triplet j; $y_{ij} = y_i, \text{for all } j = 1, ..., M$, is the clinical ground truth of whether the CT scan i is collected from an IPF subject ($y_i = 1$) or non-IPF ILD ($y_i = 0$). The patient-level diagnosis result is decided based on the majority voting of the results from all triplets.

In clinical settings, the classification task needs to be carried out in a timely manner with limited training samples and computational storage. Due to the weak supervision nature of this task (i.e. one ground truth label per CT scan) and the relatively limited number of images available, we propose to use two-dimensional convolutional neural network (2D-CNN) models, rather than 3D-CNN, for this work. 2D-CNN models are commonly used for other medical-related tasks [LYM19, ZHL19].

Dimensionality reduction is necessary before implementing the 2D-CNN models. The input of these models are usually composed of three dimensions: height, width, and depth. The height-width plane is the axial plane for the CT image and the depth plane corresponds to the three RGB channels. We propose to reduce the input dimension to $224 \times 224 \times 3$ by the incorporation of DK and optimal design theory, where 224×224 is the axial CT plane and 3 corresponds to three RGB channels for natural imaging tasks. Thus, for each training and testing sample, only three lung CT slices are used as model inputs. We refer the three CT slices as a *triplet* throughout the rest of the chapter.



Notes: The top row is one IPF patient with radiological diagnosis of UIP pattern; the second row is one IPF patient with possible UIP diagnosis; the third row is a non-IPF patient with possible UIP pattern; and the bottom row is one non-IPF patient.

Figure 2.2: Four representative CT triplets of original CT images and rescaled images.

For illustration, Figure 2.2 shows four representative triplets in terms of their original and rescaled images, with different clinical diagnoses. After preprocessing, we automatically remove the information that is outside of the body. Each CT slice is rescaled to a uniform dimension of 224×224, which is the commonly used as the default size of CNN architectures, to normalize patients with different sizes along the anteroposterior and lateral dimensions. Additionally, for prone CT scans, we rotate the scans 180 degrees to align scans with different patient positions. More details of the preprocessing steps are described in Section 1.3.3.

It is well-known that deep learning models usually require a large amount of training

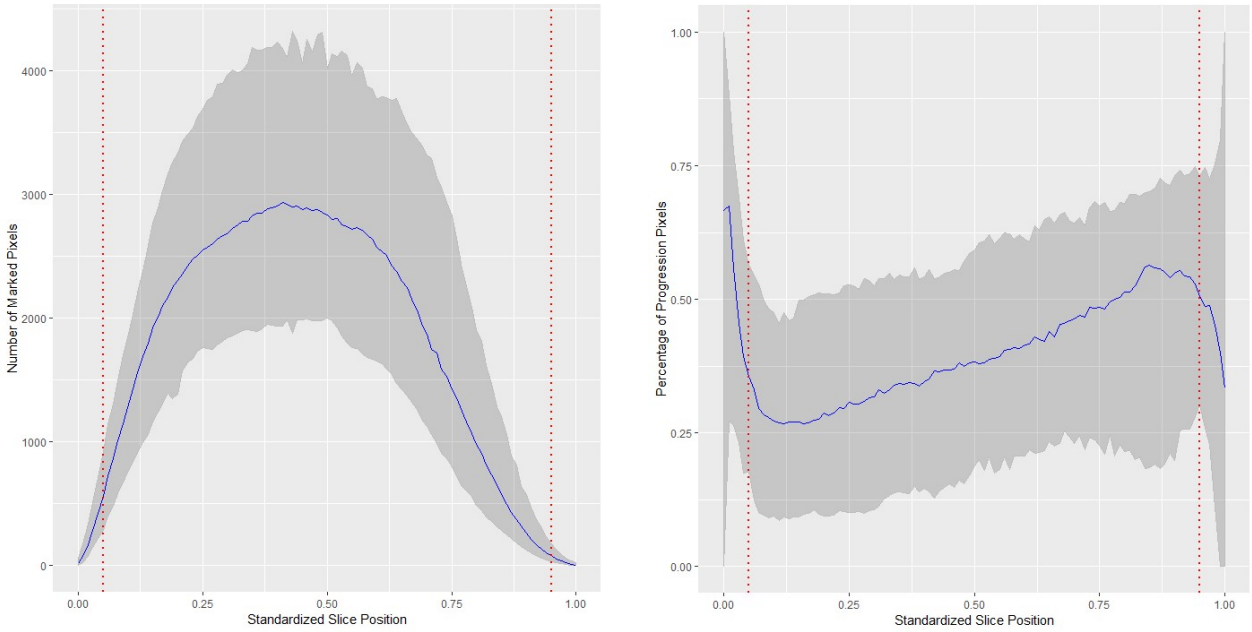
data. Specifically, a rule of thumb is that around $5,000$ labeled cases per category is needed to build a supervised deep learning model with acceptable model performance as of 2015 [GBC16]. Acquiring thousands of labeled medical image is hard; accordingly, for each scan, we randomly sample a user-selected number M of triplets to enrich the number of training and testing samples. In our study, we select $M = 20$. At the same time, we include some sensitivity analysis experiments by setting an adaptive number for M based on the number of CT slices for each scan, with more details provided in Section 2.2.8 (scenario 1).

2.2.3 Domain knowledge (DK)

We leverage DK in the selection of triplet locations using a statistical optimality design criterion and the training of the classification model in an end-to-end manner.

Specifically, we utilize the population-level disease trends of IPF in our classification task. Previous studies used quantum particle swarm optimization incorporated with a resampling technique and a random forest method to predict the pixel-level IPF progression status (i.e. whether the pixel of the segmented CT lung image suggests progressive or not progressive) [SWG19]. Intuitively, CT slices that contain more progressive pixels have more disease patterns of IPF and thus could be useful information in the classification task. Therefore, we assign higher weights for triplets which have well-represented IPF progressive trends, and vice versa. The weights for each triplet are then evaluated using an optimal design criterion.

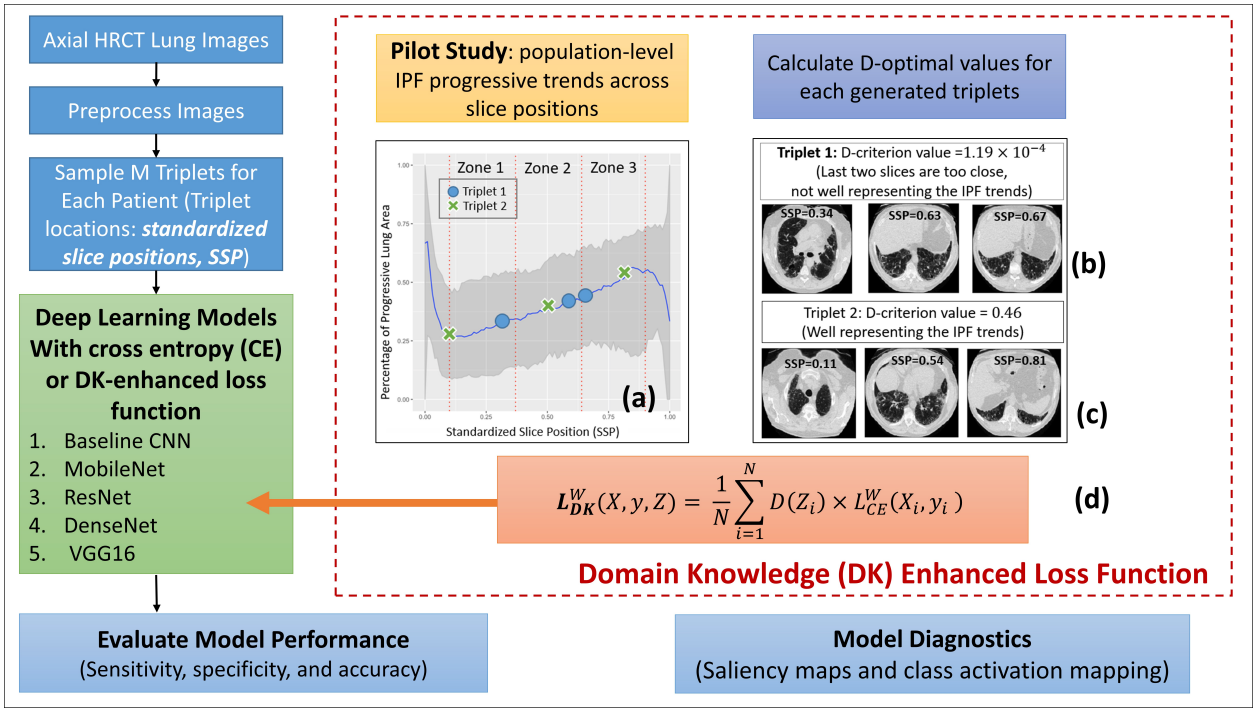
Before discussing technical details, we first define standardized slice position (SSP) to align patient visits with a varying number of CT slices. We define $SSP_i = \frac{i^{th} \text{ CT slice number} - 1}{\text{Number of CT slices} - 1}$. For example, suppose one CT scan contains 400 CT slices, then, for a specific slice number $i = 20$, its corresponding $SSP_i = \frac{20-1}{400-1} = 0.05$. By definition, SSP ranges from 0 to 1, where 0 is the first CT slice at the very first slice that contains the lung and 1 is the last CT slice at the very bottom of the lung.



(a) (b)

Notes: The blue line represents the median curve on a population level and the gray area represents the range of 2.5th percentile and 97.5th percentile. The red dotted lines represent SSP=0.05 and SSP=0.95.

Figure 2.3: The number of segmented lung area voxels (a) and the percentage of progressive voxels (b) across standard slice positions (SSP).



SSP: standardized slice position, DK: domain knowledge with optimization, CE: cross entropy without optimization in selecting slices.

Figure 2.4: Flowchart of the study design for project I.

Based on the predictive results [SWG19], we plot the number of segmented lung area pixels and the percentage of progressive lung area versus SSP based on the population level, see Figure 2.3 (a) and (b), respectively. The blue line represents the median curve on a population level and the gray area represents the range of 2.5^{th} percentile and 97.5^{th} percentile.

We observe that except for the boundaries (i.e. the apex and base of the lungs), which are defined by the top and bottom 10%, the percentage of progressive lung areas gradually increases as the slice moves towards the base of the lungs. This is consistent with previous findings for UIP patterns, which are indicative of IPF and usually reside in the base of lung parenchyma. We note that, at the boundaries (the first and last few CT slices), the number of segmented lung area voxels are much smaller than that of other areas (shown in Figure 2.3(a)). Also, there is a high level of noise effect due to the proton refection near scapula. Based on these two reasons, the prediction results at the boundaries are unstable with wide percentiles for the percentage of progressive lung areas. We therefore remove the boundaries for future analysis. Figure 2.4 (a) shows four vertical orange dotted lines, which are the SSP locations at 0.1, 0.37, 0.64, and 0.9. They are obtained by removing the top and bottom 10% to avoid the boundary effects, and then evenly dividing the rest of the lung positions into three zones, indicated as zone 1, 2, and 3 in the figure. Specifically, zone 1, 2, and 3 represent SSP locations from 0.1 to 0.37, from 0.37 to 0.64, and from 0.64 to 0.9, respectively, and they capture the upper, middle, and lower of the lungs respectively.

For each triplet, we sample one slice from each zone. We test the model performance with and without DK-enhanced loss function in Figure 2.4. Without DK, we treat each triplet identically and assign the same weights for all triplets. With DK, we assign greater weights to triplets that are more representative of the population level IPF progressive trends; see for example, triplet 2 shown in Figure 2.4 (c) for calculating the loss function. Thus, these triplets play an important role in estimating parameters in the IPF diagnostic model when the entire process is conducted in an end-to-end manner. We provide the detailed steps on how to calculate the D-criterion value of triplet 1, shown in Figure 2.4 (b), in the Supplementary B.3.

2.2.4 D-optimal design

Model-based optimal design theory has numerous and useful applications in medical research, engineering and many other disciplines [BW05, BW09]. When we have a statistical model to describe the relationship between the mean response variable and covariates, optimal design theory provides guidance on how to judiciously design an experiment to optimize the criterion. One common criterion is that model parameters be estimated as accurately as possible with minimal cost. Such an objective is attained by a D-optimal and described in more details below. For our project, a D-optimal design helps us determine the weights to be used in each triplet to assess the overall trends of the population-level IPF progressive curve using information from prior studies (see Figure 2.4 (a)) via a DK-enhanced loss function shown as $D(Z_i)$ in the formula (d) in Figure 2.4. Additional background information on optimal designs can be found in Berger and Wong [BW05], and the following design monographs [BW09, Puk06, Fed13].

We now provide some fundamentals on constructing D-optimal designs. Suppose we have N independent responses from an assumed statistical model given by $y_i = f(x_i)^T\beta + \epsilon_i$, $i = 1, ..., N$. Here y_i is the univariate response variable from subject i, $f(x_i)$ is a design vector of dimension $p \times 1$, β is the unknown parameter of dimension $p \times 1$ and the error term ϵ_i is normally distributed with mean 0 and constant variance. For example, we may have two covariates age and gender in our study and the regression function $f(x_i) = (1, \text{age}_i, \text{gender}_i)^T$ has $p = 3$ parameters.

If the interest is to estimate the three parameters in the model, two common design criteria are D-optimality and A-optimality, and if interest is to estimate the entire response surface, G-optimality is frequently used [BW05]. Here D, A, and G stand for the determinant (Det), average variance and global criterion, respectively and the resulting optimal designs have different properties. The D-optimality criterion is the most popular for estimating model parameters and mathematically, it is defined by $\text{Det}[\text{Cov}(\hat{\beta})]$. A design that achieves the smallest D-criterion value among all designs is D-optimal and such a design estimates the model parameters with the smallest volume of the confidence ellipsoid for β.

For nonlinear models, the criterion depends on the unknown parameters that we want to estimate and they have to be replaced by an initial set of estimates for the model parameters before the D-optimality criterion can be optimized. The resulting designs are, strictly speaking, locally D-optimal designs because they depend on the initial set of model parameters estimates.

Our response variable is the population trends of the percentage of progressive lung area over SSP and we estimated it using data acquired from the pilot study for $N = 122$ subjects with IPF [SWG19]. We used the generalized linear model (GLM) with a logit link function since the response variable, the percentage of progressive pixels, is not normally distributed.

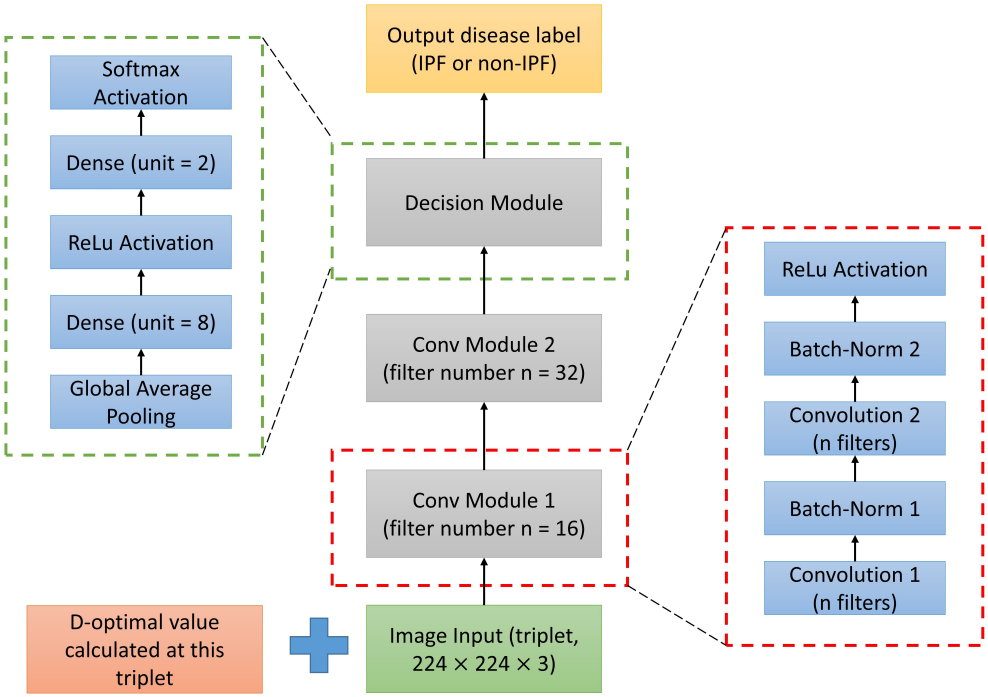
We used data and fitted several what we thought are plausible models: they include polynomial models of degrees 3 and 4 and more flexible models like fractional polynomials. The latter class models the mean response as a polynomial but additionally allows for fractional powers in each nominal. Fractional polynomials were proposed by Royston et al. [RA94, RAS99, RS08, AR01] where they showed via many examples that fractional polynomials can fit univariate response variables in the biomedical sciences much better than polynomials. They further recommended that for practical applications, it suffices to consider a set consisting of positive and nonnegative powers only. For this reason, we also used fractional polynomials to estimate the median population level disease progression. Akaike information criterion (AIC) and visual examination were used as criteria for model selection [SIK86]. Both criteria suggest that FP is the best model that describes the median population trends of IPF progression among all the models we have considered. Details on the model comparisons and estimated parameters are in the Supplementary B.2.

In a nutshell, for each randomly sampled triplet, we evaluate its D-criterion value based on the determinant of the information matrix. Triplets with a larger D-criterion value better represent the overall population level IPF progressive trends. Supplementary B.3 and B.4 contain further discussion on the D-optimal design under a generalized linear model setting and the visualization of D-criterion values.

2.2.5 Two-dimensional convolutional neural network (2D-CNN)

Before implementing 2D-CNN models, we normalized each CT scan if the scan did not meet the study-level criteria. Four main study-level criteria are: (a) align patient's position into supine, (b) center a patient position, (c) automatically remove the location of table information, and (d) rescale to a uniform image size. If a CT scan was deviated from the general platform, we normalized the images prior to the algorithm development. As a result, the processed image has the uniform property of creating a consistent lung windowing based on Hounsfield units, aligning patients' positions, automatically cropping the scans based on the presence of the body by canny edge detector using Python library scikit-image, 33 resizing to a uniform scale of 224 $\times$ 224 by cubic spline interpolation, and standardizing to a scale of zero to one. Traditional 2D-CNNs are designed for processing RGB images (three channels), which are usually of size 224 $\times$ 224 $\times$ 3. We use each triplet as one training or testing sample, where three CT slices correspond to three RGB channels.

Four state-of-the-art 2D-CNN structures are implemented for this disease classification task, which are MobileNet [HZC17], VGG16 [SZ14], ResNet-50 [HZR16], and DenseNet-121 [HLV17].



Conv: convolution.

Figure 2.5: Baseline CNN architecture.

To compare, a baseline CNN model is also designed with two convolutional modules and one decision module. The architecture of the baseline CNN model is provided in Figure 2.5.

For all of the aforementioned models (baseline CNN, MobileNet, VGG16, ResNet-50, and DenseNet-121), we run 40 epochs using batch size of 10. We use Adam optimizer with learning rate 0.0001 for all scenarios. These hyper-parameters are selected based on exploratory attempts. Model parameters are pre-trained by ImageNet [DDS09] and updated using medical images for this task. All models are implemented using Keras.

2.2.6 DK-enhanced training of 2D-CNN

We add a dense layer at the last layer of the CNN for all models, producing two CNN scores (IPF and non-IPF) for each input triplet. The softmax function is applied afterwards to normalize the CNN scores from two real numbers into two probabilities that sum up to 1. The two probabilities are the probabilities of the patient being classified into one of two

classes: IPF ($\hat{p}_{ij}$) or non-IPF ($1-\hat{p}_{ij}$) based on their specific input triplet j from subject i. Let s_{ij0} and s_{ij1} be the CNN scores after the last dense layers for triplet j from subject i being classified as non-IPF or IPF, respectively. Softmax function is used to calculate the predicted probability of being classified as IPF:

$$\hat{p}_{ij} = \frac{\exp(s_{ij1})}{\exp(s_{ij0}) + \exp(s_{ij1})}. \tag{2.1}$$

Without leveraging DK, categorical cross entropy is used as the loss function. The categorical cross entropy evaluated with deep learning model weights W at triplet j from subject i and is presented below:

$$L_{CE}^{W}(X_{ij}, y_i) = -[y_i \log(\hat{p}_{ij}) + (1-y_i)\log(1-\hat{p}_{ij})]. \tag{2.2}$$

Let X_{ij} be the CT input triplet j from subject i, let $X = (X_{11}, ..., X_{NM})$ be the set of all triplets and let $y = (y_1, ..., y_N)$, where y_i is the label of ground truth for subject i with $y_i = 1$ if the subject i is an IPF patient and $y_i = 0$ if subject i is an non-IPF patient. The overall categorical cross entropy is calculated by averaging the categorical cross entropy across all NM triplets:

$$L_{CE}^{W}(X, y) = \frac{1}{NM} \sum_{j=1}^{M} \sum_{i=1}^{N} L_{CE}^{W}(X_{ij}, y_i), \tag{2.3}$$

where N and M are the total number of patients and the number of sampled triplets from each patient respectively (N=1089 and M=20 in our research).

With DK, we designed a DK-enhanced loss function, where we weigh each triplet by its D-criterion value $D(Z_{ij})$ and $Z_{ij} = (z_{ij1}, z_{ij2}, z_{ij3})$ is a 3×1 vector representing the SSP for triplet j from subject i, and $Z = (Z_{11}, ..., Z_{NM})$ is the set of SSPs for all NM triplets. The DK-enhanced loss function is

$$L_{DK}^{W}(X, y, Z) = \frac{1}{NM} \sum_{j=1}^{M} \sum_{i=1}^{N} D(Z_{ij}) L_{CE}^{W}(X_{ij}, y_i), \tag{2.4}$$

Two sample proportion tests between DK and CE were conducted for the overall sen-

sitivity, specificity, and accuracy on all five models (baseline CNN, MobileNet, VGG16, ResNet-50, and DenseNet-121), respectively. We set the significant level to be 0.05. To account for multiple hypothesis testing, we used the Bonferroni correction to set the significance cutoff for each statistical test at 0.05/3=0.017, where 3 is the number of tests for each model, i.e. the overall sensitivity, specificity, and accuracy [Sha95].

2.2.7 Explainability evaluation: Grad-CAM

Gradient-weighted class activation mapping (Grad-CAM) is a commonly used technique in deep learning to produce visual explanations for model decisions [SCD17]. This method can provide class-discriminative and case-specific explanations for a broad range of deep learning models. Grad-CAM is a post-hoc explanation method and therefore can only be used to evaluate models, without impacting the training process of models.

Using 2D-CNN as an example, Grad-CAM contains the following procedures [SCD17]: (1) calculate the gradients of the score for each class c, y^c, with respect to a certain two-dimensional intermediate feature map $A^k \in \mathbb{R}^{u \times v}$: $\frac{\partial y^c}{\partial A^k}$; (2) global average these gradients to calculate a weight for each feature map A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{2.5}$$

where Z is a normalizing factor to make the weights sum up to 1;

(3) the final output is obtained by calculating the weighted average of the feature maps A^k and applying an ReLU transformation:

$$L_{\text{Grad-CAM}}^c = ReLU(\sum_k \alpha_k^c A^k), \tag{2.6}$$

where ReLU is an activation function with $ReLU(x) = max(0, x)$.

2.2.8 Sensitivity analysis

Sensitivity analysis is defined as a method to determine the quality of a model by evaluating the extent to which results are impacted by changing model assumptions, methods, or certain model inputs. We design three scenarios to assess whether altering one of the preprocessing steps may lead to a different model performance, including sampling different number of triplets for each scan (scenario 1), adding an image interpolation step which CT voxels are resampled into an isotropic dimension of $1mm \times 1mm \times 1mm$ (scenario 2), and sampling triplets only from lower zones (scenario 3).

Under scenario 1, instead of sampling a fixed number of triplets per scan, we sample a varying number of triplets from each scan. That is to say, the number of triplets is decided based on the number of CT slices from each CT scan. This tests if the number of triplets should vary in scans which contain different numbers of CT slices. We empirically set $M_i = 0.1 \times \text{Number of CT slices}_i$, where M_i is the number of sampled triplets for this CT scan i. For example, if one CT scan contains 250 CT slices, we set M_i=25 for this CT scan i, i.e. sample 25 triplets from this scan. The DK-enhanced loss function under **scenario 1** is

$$L^W_{DK,S_1}(X, y, Z) = \frac{1}{\sum_{i=1}^{N} M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} D(Z_{ij}) L^W_{CE}(X_{ij}, y_i). \tag{2.7}$$

Under scenario 2, in order to mitigate the possible confounding effects caused by varying slice thicknesses and pixel spacing, we resample all CT scans to a uniform isotropic cube of volume $1 \times 1 \times 1mm^3$ by cubic spline interpolation. In this step, we exclude scans which have inconsistent spacing along the z-dimension across all CT slices (non-volumetric scans, N=68, 6.2%). This step aims to align scans with different pixel spacing and slice thicknesses. The DK-enhanced loss function under **scenario 2** is

$$L^W_{DK,S_2}(X, y, Z) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} D(Z_{ij}) L^W_{CE}(X'_{ij}, y_i), \tag{2.8}$$

where X'_{ij} is the CT input triplet j sampled from subject i (X_{ij}) after isotropic resampling.

Regarding scenario 3, since IPF-related radiological features usually occur in the lower lungs, it is instructive to add one experiment to use triplets only collected from lower lungs (i.e. zone 3 in Figure 2.4 (a)). The DK-enhanced loss function with respect to **scenario 3** is

$$L^W_{DK,S_3}(\tilde{X}, y, \tilde{Z}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} D(\tilde{Z}_{ij}) L^W_{CE}(\tilde{X}_{ij}, y_i), \tag{2.9}$$

where $\tilde{Z}_{ij} = (\tilde{z}_{ij1}, \tilde{z}_{ij2}, \tilde{z}_{ij3})^T$ is the 3×1 standardized slice position for triplet j from subject i which are sampled from zone 3 only, i.e. $\tilde{z}_{ijk} \in (0.64, 0.9], k = 1, 2, 3$ for all i, j. And $\tilde{X}_{ij}$ is the CT input triplet collected based on the standardized slice position $\tilde{Z}_{ij}$.

2.3 Results

In this section, we summarize the main results and the sensitivity analysis results in Section 2.3.1 and 2.3.2, respectively.

2.3.1 Main results

We pooled CT images from all five cohorts (two IPF studies and three non-IPF studies) together for the training and testing of the model. We performed a stratified five-fold cross-validation, a commonly used technique to separate training and testing sets 41, where the proportion of IPF versus non-IPF is fixed across all folds. During cross-validation, these five folds were separated at the patient level, therefore, no triplets from the same patient are evaluated in both training and testing samples. During the testing phase, M triplets were sampled from each scan following the manner as discussed, producing M predictive results (IPF versus non-IPF) for each scan. The final predictive result for each scan was decided based on majority vote of all M triplets. We set $M = 20$ for our task. We use sensitivity, specificity, and accuracy as statistical measures. Sensitivity is defined as the number of scans which are correctly classified as IPF divided by the total number of IPF scans. Specificity is defined as the number of scans which are correctly classified as non-IPF divided by the

total number of non-IPF ILD scans. Accuracy measures the proportion of CT scans that are correctly classified.

Table 2.2 summarizes the study-wise and overall model performance using five models (Baseline CNN, MobileNet, VGG16, ResNet-50, and DenseNet-121) under two loss functions, i.e. cross entropy loss (CE) and DK-enhanced loss function (DK). Note that study 1 and study 2 include IPF patients, which is referred to as positives in this research, with sensitivity information only. Similarly, study 3, 4, and 5 contain non-IPF ILD patients, which is defined as negatives, with specificity information only. For baseline CNN model, using DK significantly increases the overall sensitivity ($P < 0.001$), but decreases the overall specificity ($P < 0.01$). There is no significant difference between DK and CE for other methods under this scenario.

2.3.2 Sensitivity analysis results

The complete results of scenario 1 (selecting a varying number of triplets per scan), 2 (adding an isotropic resampling step), and 3 (sampling from lower zones only) are provided in the Supplementary Table B.3, Table B.4, and Table B.5, respectively. For each of the scenario, we calculate the absolute difference in terms of the overall model sensitivity, specificity, and accuracy between the main results (Table 2.2) and that of each scenario. We calculate the median and interquartile range (IQR) across all ten models for each metric, under each scenario.

Under scenario 1, the median ($\pm$ IQR) for the overall model sensitivity, specificity, and accuracy between the main results and that of scenario 1 across all ten model architectures is 0.04 ($\pm$ 0.04), 0.01 ($\pm$ 0.03), and 0.02 ($\pm$ 0.03), respectively.

Under scenario 2, the median ($\pm$ IQR) for the overall model sensitivity, specificity, and accuracy between the main results and that of scenario 2 across all models are 0.01 ($\pm$ 0.03), 0.01 ($\pm$ 0.01), and 0.01 ($\pm$ 0.02), respectively.

Under scenario 3, the median ($\pm$ IQR) for the overall model sensitivity, specificity, and accuracy between the main results and that of scenario 3 across ten models is 0.03 ($\pm$ 0.03),

Table 2.2: Study-wise model performance and overall model performance.

	Sensitivity (IPF patients)		Specificity (Non-IPF ILD patients)			Overall model performance		
Model (Loss function)	Study 1	Study 2	Study 3	Study 4	Study 5	Sensitivity	Specificity	Accuracy
Baseline CNN (CE)	0.77 (0.38)	0.68 (0.39)	0.96 (0.04)	0.94 (0.09)	0.98 (0.02)	0.74 (0.38)	**0.97 (0.03)**	0.89 (0.12)
Baseline CNN (DK)	0.89 (0.13)	0.81 (0.20)	0.91 (0.07)	0.88 (0.19)	0.96 (0.03)	**0.86 (0.15)**	0.94 (0.05)	0.91 (0.04)
MobileNet (CE)	0.97 (0.01)	0.96 (0.07)	1 (0)	0.96 (0.04)	0.99 (0.02)	0.97 (0.02)	0.98 (0)	0.98 (0.01)
MobileNet (DK)	0.98 (0.02)	0.94 (0.06)	1 (0)	0.96 (0.04)	0.98 (0.01)	0.96 (0.02)	0.98 (0.01)	0.97 (0.01)
VGG16 (CE)	0.96 (0.03)	0.87 (0.07)	0.99 (0.02)	0.95 (0.06)	0.99 (0.01)	0.93 (0.04)	0.98 (0.01)	0.96 (0.01)
VGG16 (DK)	0.95 (0.04)	0.86 (0.09)	0.99 (0.02)	0.95 (0.06)	0.99 (0.01)	0.92 (0.05)	0.98 (0.01)	0.96 (0.01)
ResNet-50 (CE)	0.96 (0.02)	0.92 (0.05)	0.98 (0.05)	0.97 (0.03)	0.99 (0.01)	0.95 (0.02)	0.98 (0.01)	0.97 (0.01)
ResNet-50 (DK)	0.96 (0.02)	0.90 (0.09)	1 (0)	0.96 (0.05)	0.99 (0.01)	0.94 (0.03)	0.98 (0.01)	0.97 (0.01)
DenseNet-121 (CE)	0.97 (0.02)	0.98 (0.02)	1 (0)	0.97 (0.04)	0.98 (0)	0.97 (0.01)	0.98 (0.01)	0.98 (0)
DenseNet-121 (DK)	0.96 (0.04)	0.94 (0.06)	1 (0)	0.97 (0.04)	0.99 (0)	0.95 (0.02)	0.99 (0.01)	0.97 (0)

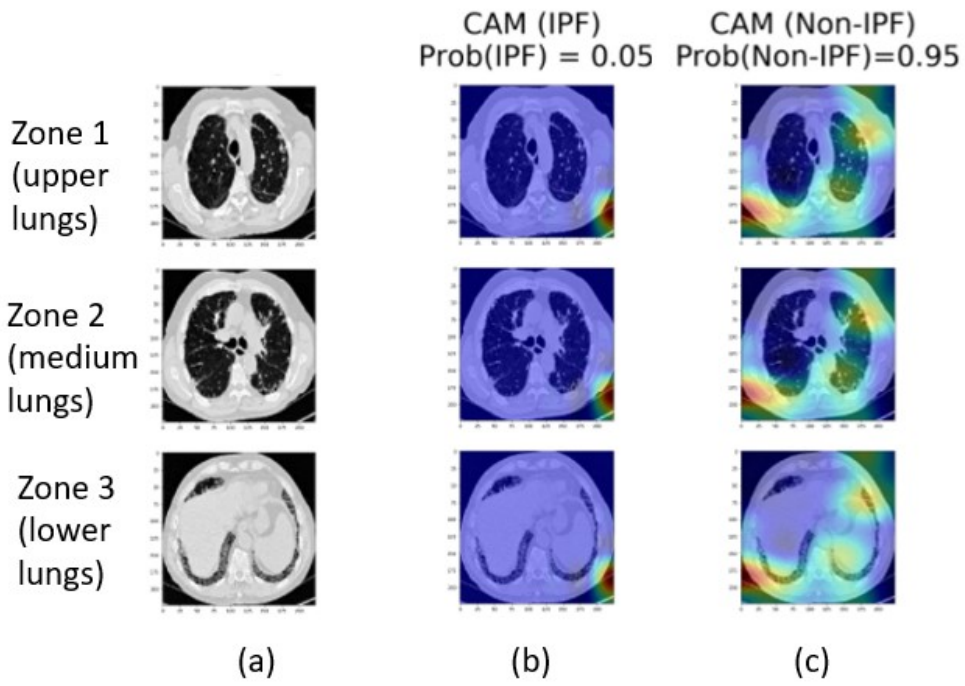
Note: Mean and standard deviations shown in brackets are calculated across the results from each testing fold. CE: cross entropy loss without domain knowledge-enhanced loss function; DK: domain knowledge-enhanced loss function. Statistically significant results ($P < 0.017$) are highlighted in bold font. The significance cutoff 0.017 is decided by Bonferroni correction for multiple testing, which is dividing the pre-specified significance level 0.05 by the number of tests (3, including the overall sensitivity, specificity, and accuracy) for each model.

0.01 ($\pm$ 0.01), and 0.02 ($\pm$ 0.01), respectively.

2.3.3 Explainability evaluation: Grad-CAM

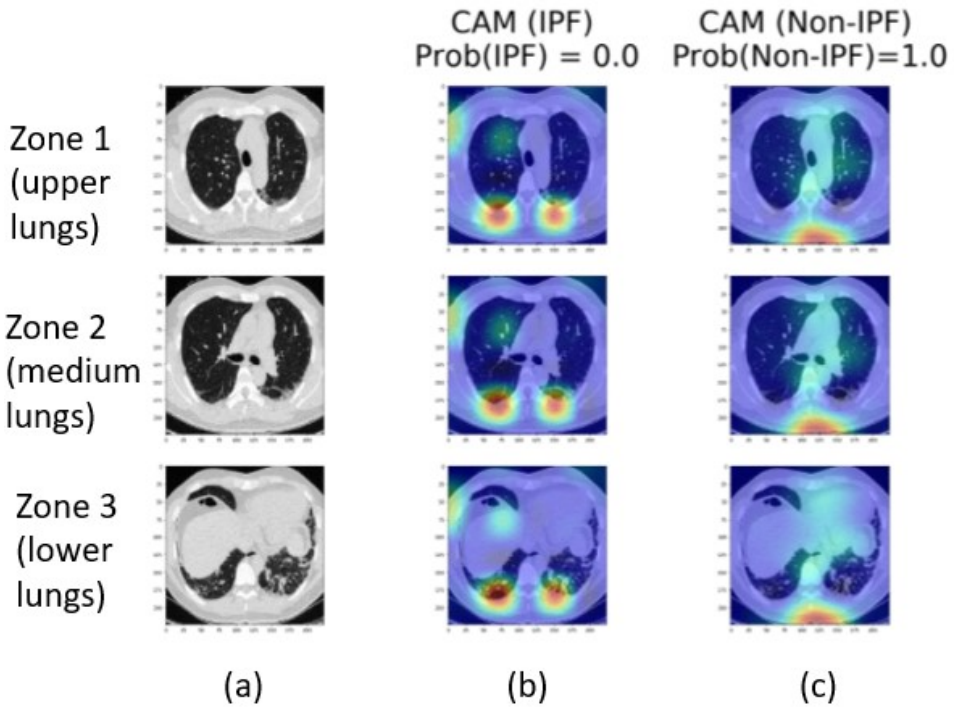
Figure 2.6 and 2.7 are the Grad-CAM results for one IPF subject non-IPF subject with a clinical diagnosis of myositis-related ILD, respectively. We used one ResNet-50 model with categorical cross entropy loss as an example. These plots are the testing cases (not training cases) for the selected model. By definition, Grad-CAM plots are case-specific and class-specific. Therefore, for each triplet, we have its corresponding Grad-CAM plots with respect to the IPF class and non-IPF class, respectively. Greater values in Figure 2.6 and 2.7, as shown in red, are the important regions in this triplet for classifying as each class.

Since the three CT slices are combined and treated as one input for the model, by the definition of Grad-CAM, we can only acquire one 2D image as the output for a given triplet, per class. For ResNet-50 models, the dimension of the 2D Grad-CAM plot is 7×7, based on the default dimension changes of the feature maps. For plotting purposes, we rescaled the 2D Grad-CAM plot as the same size as the processed CT slice (i.e. 224×224) and superimposed the *same* Grad-CAM plot to these three CT slices.



Notes: This triplet is incorrectly classified as Non-IPF by the ResNet-50 model (i.e. Prob(Non-IPF)$>$ 0.50).

Figure 2.6: Grad-CAM plots for one IPF subject. Processed CT triplets, Grad-CAM plots for the IPF class, and Grad-CAM plots for the non-IPF class are plotted in column (a), (b), and (c).



Notes: This triplet is correctly classified as Non-IPF by the ResNet-50 model (i.e. Prob(Non-IPF)> 0.50).

Figure 2.7: Grad-CAM plots for one Non-IPF subject. Processed CT triplets, Grad-CAM plots for the IPF class, and Grad-CAM plots for the non-IPF class are plotted in column (a), (b), and (c).

2.4 Discussions and conclusions

We developed a deep learning-based model for IPF diagnosis: (1) from a clinical perspective, by incorporating DK regarding the disease pattern distribution of IPF; (2) from a methodological perspective, by including optimal design methods in building a loss function. Methodologically, to the best of our knowledge, this is the first work that leverages the merits of optimal design in the training of deep learning methods in an end-to-end manner. Clinically, providing automatic IPF diagnosis support is timely and meaningful because the proposed method (1) facilitates automated IPF diagnosis and reduces inter- and intra-reader disagreement; (2) enables early anti-fibrotic treatment and so may prolong patient's survival time; (3) decreases the likelihood of requiring of lung biopsy in the long run and its

attendant's risks.

In medical imaging domain, as contrary to natural imaging, well-labeled and high-quality images are time-consuming and expensive to acquire. Therefore, several researchers aim to tackle the limited sample size problem in medical imaging by utilizing DK [CLX18, PMW19]. Unlike previous work, we now focus on the population-level information acquired from the previous studies and utilize both DK and optimal design guidelines in the training process of the deep learning models.

Each of the earlier studies used in this research contains either IPF patients in study 1 and study 2 or non-IPF patients in study 3, study 4 and study 5, and one may argue that the diagnosis model captures confounding effects (or batch effects) rather than IPF-related CT features. Admittedly, this is one limitation of this work due to the availability of imaging data and the nature of retrospective data collection. However, we note that each study is conducted at multiple sites with different protocols and a variety of experimental conditions that likely involve CT scanners, slice thickness, reconstruction kernel, and patient positions, see Supplementary B.1 for an expanded list of potential confounders. This heterogeneous experimental setup contributes to a fair model that concentrates on the underlying CT features of IPF rather than picking up other confounding factors.

In addition, to address this concern of confounding effects, we have added multiple model generalizability experiments (see Supplementary B.6 for more details). By setting aside one study as the holdout test set at one time, we evaluate the generalizability of the constructed model to unseen domains (i.e. institutions and clinical diagnoses) using MobileNet. The results suggest that, most experiments can successfully classify more than 90% of patients in the holdout study (accuracies greater than 90%). This suggests that most experiments are able to generalize well to unseen domains. Notably, there is a certain level of decrease in overall model accuracy compared to results provided in the Table 2, when using one study as the holdout study at a time. For example, for six out of eight generalizability experiments, we observe a 1%-4% degradation in model accuracy; for two out of eight experiments, we observe a 25%-26% decrease in model accuracy, which we provide some explanations in the Supplementary B.6. This degradation in performance may due to the fact that the number

of training and testing samples are fewer since we set one study aside as the holdout set. At the same time, this lack of generalizability is not surprising as such findings are frequently reported in many areas of research when deep learning models are applied to unseen domains [ZBL18]. This provides a warning that when deploying the developed model to scans collected from other institutions or ILD patients with different clinical diagnoses, some decrease in model performance is to be expected. Many domain adaptation and domain generalization techniques have been developed to tackle this problem, but they are out of the scope for this dissertation [KBL17, DOC18, DCK19, MMK18, SWU18].

In summary, we have, for the first time, incorporated the population-level DK (i.e. IPF progression trends across the lung position acquired from pilot studies) with ideas of optimal design methodology into the training of deep learning models. Specifically, we sample 20 triplets from each CT scan to augment the number of training data and boost model performance. These triplets were randomly sampled with one from each zone (the top, middle, and bottom of the lungs). Intuitively, these 20 triplets should not be treated identically, as these randomly sampled CT slices might not be fully representative and reflect the disease characteristics fairly. Some triplets might contain three slices which are adjacent to each other, and thus contain less disease information.

To this end, we estimated the population-level disease trends across lung positions from previous studies and evaluated the importance of each triplet by its D-optimality value. The triplet with a larger value is "a better design" for estimating the parameters of the population-level trends, and consequently, it is believed to be more representative of the overall disease trends. We then design the DK-enhanced loss function, where the D-criterion value of each triplet is used as a weight to evaluate the importance of each triplet. This process is incorporated into the training of the deep learning models in an end-to-end manner. Current experiments show that incorporating DK in the training of deep learning models increases the overall accuracy from 0.89 to 0.91 for the baseline CNN model. However, this increase in the overall accuracy using DK is not observed for other well-known model architectures, including MobileNet, VGG16, ResNet50, and DenseNet-121. This may occur due to the existence of ceiling effect, since other well-developed deep learning architectures

have already achieved a satisfactory model performance with overall accuracy greater than 0.95. We also expect the proposed methodology is generally applicable to tackle other similar problems in the medical arena as well, even though our work here only concerns IPF diagnosis.

Sensitivity analysis experiments suggest that (1) selecting a flexible number of triplets per scan, (2) isotropic resampling each scan to a constant size of $1mm^3$ cube, and (3) sampling triplets only from lower zones may change the overall model sensitivity, specificity, accuracy in a reasonable range. For example, there is no notable increase in the overall model performance by adding one step of isotropic resampling, for our experiment. This may in part due to the ceiling effect.

Our future work includes exploring the constructed model on prospective studies, where IPF and non-IPF ILD patients are collected under the same imaging protocols. This is a more accurate reflection of the clinical applicability of the developed model, as contrary to using five-fold cross validation without independent studies.

In conclusion, we develop an efficient IPF diagnosis model using DK (i.e. population-level disease information) and optimal design theory. This study shows satisfactory performance using various well-known deep learning models in the task of IPF diagnosis using CT images. To the best of our knowledge, this is the first work that (1) leverages population DK with optimal design criterion to train deep learning models in an end-to-end fashion; (2) focuses on patient-level IPF diagnosis solely based on CT images.

CHAPTER 3

Project II: A Two-stage Multi-scale Guided Attention Model

3.1 Background

Generally speaking, the successful application of deep learning systems in clinical practice usually relies on these three prerequisites: (1) the availability of well-labeled fine-scale data, which are usually at a voxel, regions of interest (RoI), or image slice level; (2) the extent of explainability on where and how the deep learning-based system makes the decision; and (3) the ability to generalize well to a new dataset [LYM19].

To address these aforementioned concerns, we propose a two-stage, attention-based model that is generally applicable to weakly supervised tasks, where only CT scan labels are available, to enhance the explainability and generalizability.

Specifically, this scan-level IPF diagnosis task falls into the category of *inexact supervision* in the field of weakly supervised learning, where some level of high-level supervision (in our case, CT scan-level ground truth) is provided, but not as desired, such as having every region of interest or CT slice labeled with IPF-related features [Zho18]. Contrary to the expensive fine-scale labels, population-level domain knowledge is easier and less labor-intensive to acquire. In this project, population-level domain knowledge was acquired using a well-developed and automated algorithm to characterize IPF prognosis over the entire lung. We used attention models to encourage the constructed diagnosis system to focus on the regions guided by the population-level domain knowledge.

Attention mechanisms (or attention models), which originated from natural language

processing, have gained substantial interest in multiple research areas, such as computer vision, medical imaging, speech recognition, etc [CMP19]. The intuition behind attention models is similar to human behaviours that different input elements have a varying extent of relevance and contribution for a specific task. For example, for an image classification task, our visual recognition system tends to selectively concentrate on several important regions while paying less attention to other irrelevant regions, such as background. This is what attention models are designed for: to encourage models to capture the relevance of input elements in spite of the distance between each element and concentrate on the important regions for this task. Accordingly, attention models are reported to have several advantages. Firstly, they can capture long range dependencies between each input element, such as two words in a sentence that are semantically related but are distant from each other [WGG18]. Secondly, attention models are one way to explain which region of the input image the network's decision depends on and can enhance the explainability of deep learning-based systems [LWP18]. Thirdly, attention models encourage the network to focus on the task-specific regions and therefore strengthen model generalizability to a new dataset [JLL18].

Attention mechanisms have recently become popular in the medical imaging domain to solve the research question of segmentation [SOS19, LDT20, SD19], classification [YKK19], detection [ZDG19], and so on. In this work, guided attention modules of multiple scales are implemented to encourage the deep learning-based system to focus on the areas of interests, which are lung parenchyma, especially the peripheral lung areas, based on the provided population-level domain knowledge acquired from prior studies.

To summarize, in this work, we propose a two-stage automated IPF diagnosis model: at stage one, the multi-scale guided attention (MSGA) is a trainable, end-to-end IPF diagnosis model that leverages the provided domain knowledge at two resolution scales; at stage two, we further construct a random forest (RF) model that takes the MSGA output and produces the final patient-level diagnosis results. Our contributions are (1) developing an IPF diagnosis model that only uses scan-level weak supervision; (2) incorporating population-level domain knowledge into the training of IPF diagnosis model in an end-to-end manner;

(3) enhancing the explainability of deep learning systems at various layers by introducing multi-scale attention mechanisms; (4) further boosting model performance by adding an RF classifier.

3.1.1 Attention models

1. Origination: natural language processing

Attention models, originated from natural language processing, have gained research interests in recent years. Attention models have been known for its abilities to capture long-range dependencies, regardless of the distance between two input words in a sentence.

For a language translation task, attention models can adaptively capture the relative alignment ("attention scores") between words. Take a sentence as an example: she is reading a fascinating book. In this case, the word "reading" should have a strong association (higher attention scores) with the word "book", regardless of the distance between these two words.

Luong et al. developed a language translation model which for each current word, they calculated its *alignment* between all (global attention) or some (local attention) of the previous words [LPM15]. A schematic of the global attention model is shown in Figure 3.1. At each time step t, the alignment of the current word and previous words is shown as a_t, which is then used to weigh all of the previous words, producing a context vector c_t. The context vector is concatenated with the current word vector h_t to produce a hidden state $\tilde{h_t}$ for further word predictions. By this design, every previous word has a potentially varying impact on the current word prediction, which is consistent with how people perceive languages.

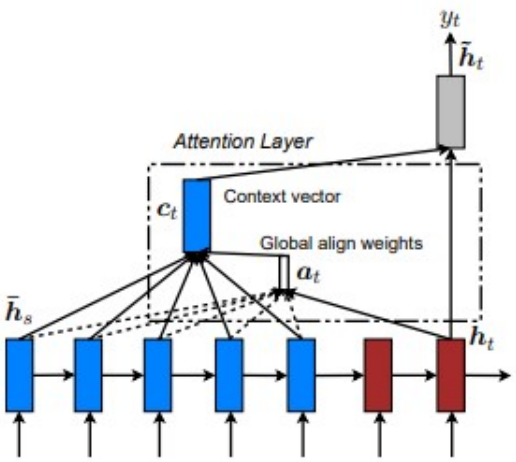


Figure 2: **Global attentional model** – at each time step t, the model infers a *variable-length* alignment weight vector $\boldsymbol{a}_t$ based on the current target state $\boldsymbol{h}_t$ and all source states $\bar{\boldsymbol{h}}_s$. A global context vector $\boldsymbol{c}_t$ is then computed as the weighted average, according to $\boldsymbol{a}_t$, over all the source states.

Figure 3.1: A schematic of the global attention model from [LPM15].

Not limited to natural language processing, attention models have gained popularity in medical domains recently, largely due to its ability to model long-range dependencies, utilize parameters efficiently, and highlight salient regions that are important for the task.

Attention models fall under two main categories - unguided (without external guidance) and guided (guided by external domain knowledge). The majority of the current work focuses on building *unguided* attention mechanisms within different layers of the constructed networks, without providing external guidance of domain knowledge. For example, researchers usually used the coarse features extracted at later layers to guide the training under an attention model, without providing external guidance [SOS19, LDT20]. Recent work on *guided* attention models include using region-level coarse annotation [YKK19] or binary maps of some RoIs [YKH19] to guide the model in an end-to-end training fashion. In this work, we design an attention model under the guidance of *population-level* domain knowledge, which is less labor-intensive to acquire, compared to the previous work [YKH19] [YKK19].

2. Definitions

We discuss several commonly-used terms in the field of attention models here.

Guided versus unguided attention: as mentioned before, this refers to whether external guidance or information is provided to the attention model. If external guidance, such as case-level contours or population-level information, is provided, then the model belongs to the category of guided attention model. Specifically, for project II, we aim to provide the model with population-level domain knowledge maps, then our method lies into the category of guided attention.

Soft versus hard attention: this definition is based on whether some information is selectively and completely ignored (hard attention) or information is reweighted but never removed (soft attention). Currently, the vast majority of the research is based on soft attention networks. This is because hard attention is non-differentiable, making gradients-based deep learning framework failed [MDS18]. Therefore, future discussions are based on soft attention models and we use soft attention models in our project II.

Inter-attention versus intra-attention: this defines whether the attention scores are calculated within input (intra-attention) or with other information (inter-attention). Intra-attention, also called as self-attention, models the relationship between each input (i.e. image location or sentence sequence) with all positions at the input [VSP17, ZGM19]. On the contrary, inter-attention models the alignment between the current intermediate feature maps with other information, such as global features which are extracted from other layers [JLL18] or previous words in the sentence [LPM15]. We note that the method we used in project II falls into the category of intra-attention because we model the relationship within fixed intermediate feature maps.

3. Advantages

Attention models have prospered in computer vision due to the following strengths.

- Attention models can characterize **long-range dependencies** and utilize model parameters more efficiently. Without attention gates, common convolutions operate on local receptive fields, which are fixed-size local areas; it takes several convolutional layers to propagate long-range dependencies. Attention models, which capture the

dependencies between each local region with all other positions or a provided global feature, can effectively model any existing long-range relationships across the image with the need for fewer model parameters [WGG18].

- Attention mechanisms can contribute to the development of **explainable AI** and serve as an effective tool for model diagnostics. If the attention coefficients focus on suspicious regions for medical imaging tasks, such as a metal token in the previous example [ZBL18], this warns the researchers that certain preprocessing steps are needed.

- Some researchers found that attention models demonstrated superior **generalizability** to unseen datasets [JLL18]. This may due to the fact that attention models capture the regions of interest, suppress irrelevant background information, and thus are able to generalize well.

3.1.2 IPF quantitative index: kurtosis

Attention maps can provide visual guidance to enhance the explainability of the models.

However, the visual examination of attention maps requires expert knowledge to examine whether (1) attention maps can capture the lung parenchyma; (2) the highlighted regions correspond to the important regions for IPF diagnosis. These evaluations are subjective and are usually based on the examination of a limited number of cases. Therefore, evaluating attention maps based on a few randomly-selected or human-picked cases can introduce bias in terms of both case selection bias and human evaluation bias.

To solve this problem, we aim to develop a quantitative and reproducible measure to evaluate attention maps objectively without human intervention. This can benefit the researchers from the following perspectives: (1) provide insights on the discrepancy between IPF and non-IPF using the estimated attention maps; (2) automatically evaluate attention maps with limited resources.

Previous researchers found that for IPF subjects, histogram features of CT values in the segmented lung regions, especially kurtosis, are associated with physiological abnormalities

[BLB03]. Kurtosis measures the tailedness in the distribution of CT values among the lung regions.

It has been found that the histogram of CT values in a normal lung or mild IPF has longer tails (higher kurtosis), whereas the histogram of an IPF subject has shorter tails (lower kurtosis). This may due to the existence of lung abnormalities, such as honeycombing and reticulation. Figure 3.2 shows an illustrative figure of the two subjects with mild and severe IPF, where kurtosis is 5 and 0.41, respectively [KBC15].

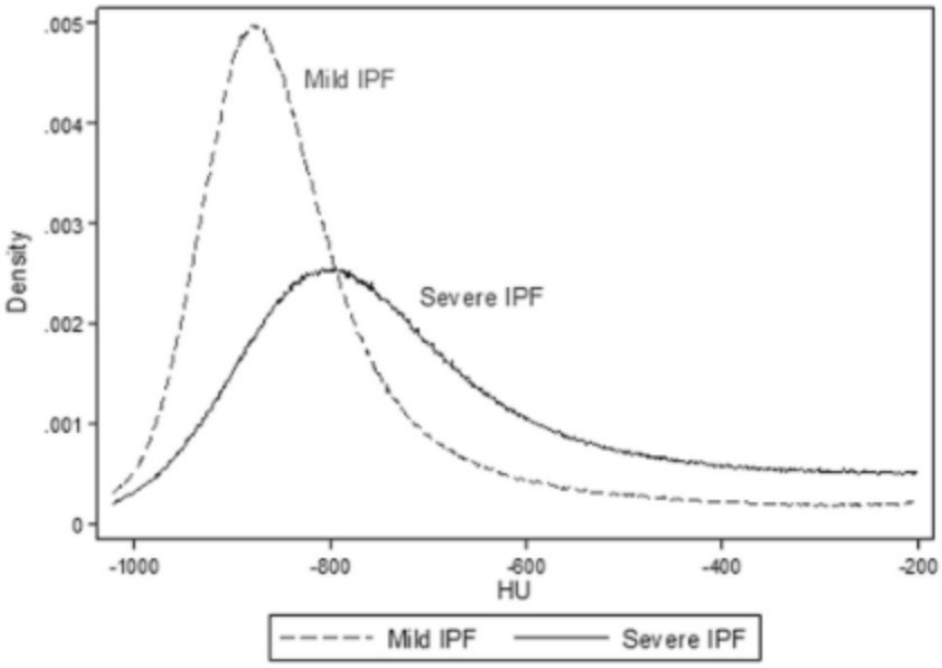


Figure 3.2: Histograms of CT values from two patients with a mild (kurtosis=5) and severe IPF (kurtosis=0.41). [KBC15]

Inspired by this work, we plan to use kurtosis as a quantitative measure to evaluate model explainability. In more detail, we calculate the kurtosis of intermediate feature maps produced by the attention gates (instead of CT images) and examine whether there is a statistical difference between IPF and non-IPF subjects.

3.2 Materials and methods

3.2.1 Datasets

A total number of 878 volumetric non-contrast high-resolution CT (HRCT) scans were retrospectively collected from two IPF (N=349, 39.7%) and three non-IPF ILD cohorts (N=529, 60.3%). CT images of IPF patients were collected from December 2004 to July 2016; CT images of non-IPF patients were collected from May 1997 to May 2018. For each subject, only the first total scans (performed at total lung capacity) were used for model construction and testing.

We randomly split these CT scans into two subsets while preserving the proportion of IPF to non-IPF subjects: the training and validation set (N=702, 80.0%, IPF%=39.7%) and the testing set (N=176, 20.0%, IPF%=39.8%), as illustrated in Figure 3.3. The training and validation sets were used for model training and hyperparameter selection for MSGA+RF; the testing set was employed as a holdout set to examine the final model performance.

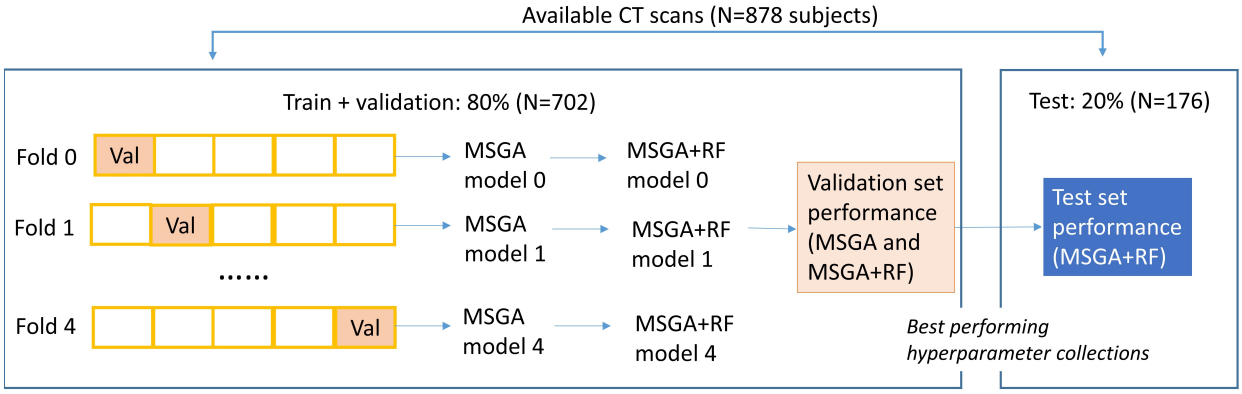


Figure 3.3: The overall separation of the dataset. Val: validation, which is the subset that is used to evaluate the model performance at a specific fold.

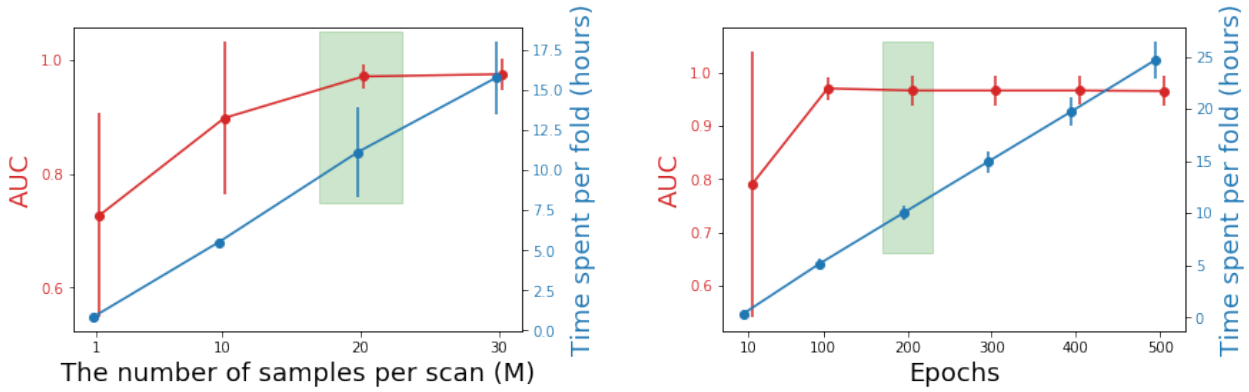
A stratified five-fold cross validation was employed to the training and validation set to explore the changes in the results with respect to different hyperparameter selections. For each fold, the entire training and validation set were separated into five subsets while fixing the proportion of IPF subjects at each subset: four subsets were used to construct

the model and one subset was used to evaluate the model performance (shown as "Val" in Figure 3.3). For each hyperparameter selection, we constructed five MSGA models, leaving one fold of data as the validation set as a time; after the training of MSGA was complete, we then built an RF model for each MSGA model using the training cases in that fold only. The mean and standard deviations across five folds for both MSGA and MSGA+RF were reported as validation set performance. Based on the validation set performance, we further selected best performing hyperparameter combinations as our final model(s) to apply to the test set. Test set performance was reported as the mean and standard deviation across five folds constructed using the selected best performing hyperparameters.

3.2.2 Image processing

HRCT scans underwent an in-house image preprocessing pipeline, including (1) creating a lung window of (-1250 HU, 250 HU) Hounsfield units (HU) for better visibility of lung parenchyma, (2) aligning patients' positions to be supine, (3) automatically cropping the scans based on the presence of patient's body by canny edge detector using Python library scikit-image [WSN14], (4) adding a step of isotropic resampling to a uniform cube of size $1mm \times 1mm \times 1mm$, (5) resizing to a uniform scale by cubic spline interpolation, (6) and standardizing to a range of [0,1] on a scan level. After preprocessing, each CT scan was resized to a standardized dimension $256 \times 256 \times 128$. To boost sample size and reduce the data dimension, we further resampled a fixed number (M) of 3D-volumes, with dimension $128 \times 128 \times 64$ from each scan. Each sample is treated as a unit for the training and validation step for MSGA. We use subject index i and sample index $j = 1, \ldots, M$; for example, X_{ij} is the j^{th} sampled CT volume from subject i.

We evaluate the model performance and computational time with a varying resampling size M, including $M = 1, 10, 20, 30$, and this analysis is reported in Figure 3.4 (a). $M = 20$ is chosen in our experiment due to the satisfactory model performance within a limited computational time.



(a) (b)

Notes: The hyperparameters that we selected, M=20 and epochs=200, are highlighted as green rectangles.

Figure 3.4: Average AUC scores (± standard errors) and average time spent per fold (± standard errors, in hours) with different values of the number of samples per scan (a) and epochs (b) using five-fold cross validation.

3.2.3 Problem statement

The main research question is a supervised binary classification task to determine whether the CT scan represents a subject with IPF or non-IPF, among subjects diagnosed with ILD. The input of the MSGA+RF system contains three components: $\{(X_1, ..., X_N), (y_1, ..., y_N), \widetilde{DK}\}$ and the expected output contains two parts: $\{(\hat{y}_1, ..., \hat{y_N}), (\hat{\beta}_{11}, ..., \hat{\beta}_{NM}).\}$.

Specifically, X_i is the patient-level CT scan collected from subject i; $y_i \in \{0, 1\}$ is the ground truth indicating whether the subject i is clinically diagnosed as IPF ($y_i = 1$) or non-IPF ILD ($y_i = 0$); N is the number of subjects in the study; $\widetilde{DK}$ is a standardized quantitative measure of population-level domain knowledge collected from previous studies, indicating which regions are usually critical for this task. $\hat{y}_i$ is the predicted label for scan i and $\hat{y}_i \in \{0, 1\}$.

Here $\hat{\beta}_{ij}$ is the estimated attention maps for scan i and sample j, highlighting the regions that are meaningful for this task. In this work, we implemented two attention modules at a high and medium resolution level, then $\hat{\beta}_{ij} = (\hat{\beta}^h_{ij}, \hat{\beta}^m_{ij})$, where $\hat{\beta}^h_{ij}$ and $\hat{\beta}^m_{ij}$ is the estimated attention map at a high- and medium- resolution for subject i and sample j, respectively. In this study, since only the first CT scan is used for each subject, index i is both subject

index and scan index.

We provide some dimension information. X_i is usually of dimension $512 \times 512 \times$ Number of CT slices, where 512 is the number of voxels in the x- and y-dimension for each CT slice. Standardized domain knowledge ($\widetilde{DK}$) is a multidimensional array of dimension $256 \times 256 \times 128$, which is downsampled to a high- and medium- resolutions, as represented by $\widetilde{DK}^h$ and $\widetilde{DK}^m$, which are dimension $64 \times 64 \times 32$ and $16 \times 16 \times 8$, respectively. For the estimated attention maps, the dimensions of $\hat{\beta}_{ij}^h$ and $\hat{\beta}_{ij}^m$ are $64 \times 64 \times 32$ and $16 \times 16 \times 8$, respectively. The image dimension is represented as $H \times W \times D$ throughout this chapter, where the depth dimension D is the dimension along the patient's body from apex to base and height-width $(H - W)$ plane is the axial plane of each CT slice. The dimension of intermediate features generated by 3D-convolutions is $H \times W \times D \times C$ where C-dimension is the channel dimension. The initial design of channels was inspired by the three RGB channels for 2D images but was extended to a broader definition afterwards.

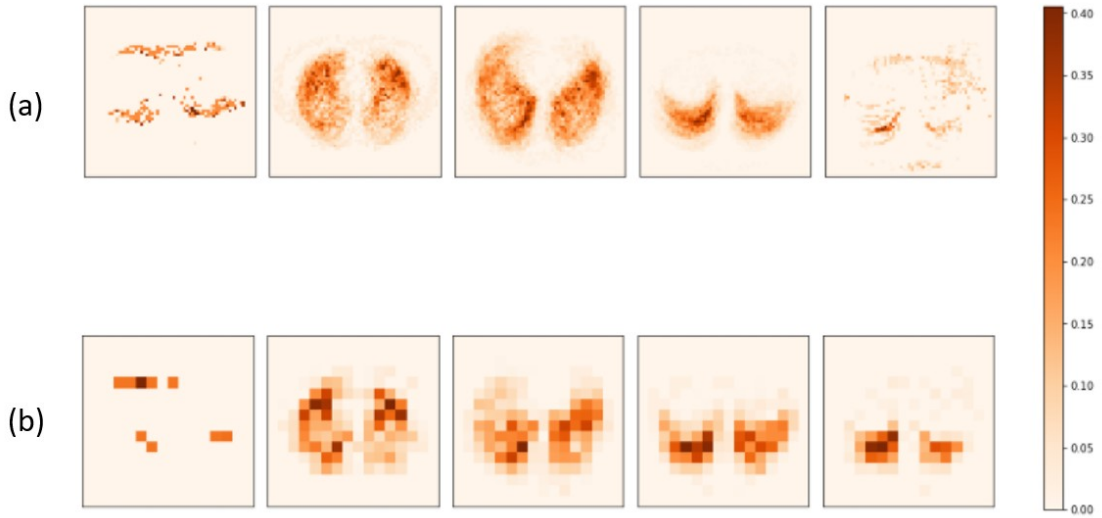
3.2.4 Population-level Domain knowledge

Explainability: In the past ten years, quantitative CT imaging biomarkers have been developed and evaluated as clinical outcome measures among patients with ILD [KTC10]. These developed measures are sensitive to localized changes and can be used as domain knowledge to guide the training of the IPF diagnosis model.

Kim et al. developed an automated algorithm to evaluate the voxel-level disease prognosis based on CT scans [KTC10]. The algorithm can be applied on denoised and segmented CT scans and predict whether each pixel is an indication of lung fibrosis (LF, pulmonary fibrosis) or other lung fibrosis (OLF, pulmonary fibrosis with similar texture features of vascularity). When extending the software constructed using RoI-level information on a whole lung level, the researchers added one additional step of classify if the voxels are pseudo vessels: that is, if the mean feature of the decomposed texture image is greater or equal to zero, the voxels are classified as pseudo vessels (OLF) [Hyu07]; if the mean feature of the decomposed texture image is less than zero, the voxels are classified as LF. It follows that the sum of LF and

OLF is the extent of quantitative lung fibrosis (QLF) scores.

Based on a prior study conducted on 102 eligible IPF subjects, with subjects $t = 1, ..., T = 102$, we estimated whether each voxel was predicted as LF or OLF based on the developed algorithm [KTC10]. We define $LF_v^t = 1$ or $OLF_v^t = 1$ if the scan for subject t at voxel location v is predicted as LF or OLF, respectively; We have $LF_v^t = 0$ or $OLF_v^t = 0$ if the scan for subject t at voxel location v is not predicted as LF or OLF, respectively. At each location v, we summed over all T subjects by $LF_v = \sum_{t=1}^{T} LF_v^t$ and then standardize to a scale of [0,1]: $\widetilde{LF}_v = \frac{LF_v}{\max_v LF_v}$. Standardized other lung fibrosis ($\widetilde{OLF}_v$) can be estimated similarly.



Notes: subplots (a) are produced at the 3%, 28%, 53%, 78%, 97% position along the depth D-axis; Subplots (b) are produced at the 13%, 38%, 63%, 75%, 88% position along the D-axis.

Figure 3.5: Population-level domain knowledge at high (a) and medium (b) resolutions.

We defined the domain knowledge ($\widetilde{DK}_v$) as the maximum of $\widetilde{LF}_v$ and ($\widetilde{OLF}_v$) for each fixed location v: $\widetilde{DK}_v = \max(\widetilde{LF}_v, \widetilde{OLF}_v)$, where $\widetilde{DK}_v$ ranges from [0,1] by definition. Domain knowledge ($\widetilde{DK}$) is later downsampled to two resolution scales: 64×64×32 and 16×16×8, as shown in Figure 3.5. Higher intensity values (more orange) in Figure 3.5 represent a greater value of ($\widetilde{DK}_v$), which concentrates on the RoI for this IPF diagnosis task. Lung areas, especially peripheral lungs, are highlighted in Figure 3.5, which is in

agreement with IPF-related CT features. In the future sections, we will discuss how domain knowledge is incorporated as an integral part of the loss function during training to encourage the model to focus on IPF disease patterns.

3.2.5 Attention gates

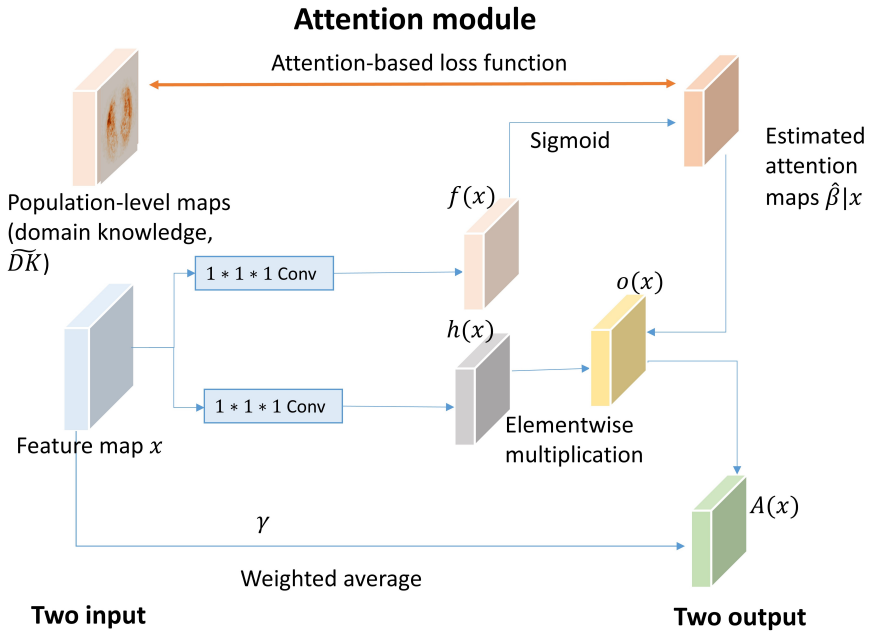


Figure 3.6: Attention modules for project II.

We provide a schematic of the proposed guided attention gates in Figure 3.6. Under the current setting, the attention gates take intermediate feature maps x and population-level domain knowledge (DK) as input and produce two outputs, including (1) a feature map that is of same dimension as the input: $A(x)$ and (2) an estimated attention map $\hat{\beta}|x$. For simplicity, $\hat{\beta}|x$ is represented as $\hat{\beta}$ throughout the manuscript. Theoretically, attention gates can be incorporated in any layer of any existing CNN architecture. In this work, we focus on the attention gates that are suitable for 3D-CNN architectures, which generate intermediate feature maps of four dimensions, including height, width, depth and channels. Attention gates can be applied for 2D-CNN architectures, which contain feature maps of three dimensions, including height, width, and channels, with minor revisions.

Suppose the attention gates are implemented at the l^{th} layer and takes the intermediate feature maps x^l that are generated at the previous layer, i.e. $(l-1)^{th}$ layer, as input. For 3D-CNN architectures, x^l is a four-dimensional tensor with $x^l \in \mathbb{R}^{H^l \times W^l \times D^l \times C^l}$, where H^l, W^l, D^l, C^l are the height, weight, depth, the number of channels at the l^{th} layer, respectively. For simplicity, we omit the subject index i and sample index j throughout this section. The intermediate feature maps x^l are first transformed into two feature spaces $f(x)^l$ and $h(x)^l$ using 1×1×1 convolutions:

$$f(x)^l = x^l \times W_f^l, \tag{3.1}$$

$$h(x)^l = x^l \times W_h^l, \tag{3.2}$$

where $W_f^l \in \mathbb{R}^{C^l}$, $f(x)^l \in \mathbb{R}^{H^l \times W^l \times D^l}$, $W_h^l \in \mathbb{R}^{C^l \times C^l}$, $h(x)^l \in \mathbb{R}^{H^l \times W^l \times D^l \times C^l}$.

A sigmoid function is applied to the feature space $f(x)^l$ to calculate the attention scores (i.e. estimated attention maps) at layer l at a three-dimensional voxel location $v = (v^{H^l}, v^{W^l}, v^{D^l})$, $\hat{\beta}_v^l$, where

$$\hat{\beta}_v^l = \frac{1}{1 + \exp(-f(x)_v^l)}. \tag{3.3}$$

$\hat{\beta}_v^l$ is a scalar, and $v^{H^l} \in \mathbb{R}^{H^l}$, $v^{W^l} \in \mathbb{R}^{W^l}$, $v^{D^l} \in \mathbb{R}^{D^l}$.

The dimension of $\hat{\beta}^l$ is decided by the choice of layers l where the attention module is implemented in. In our example, let the model layers where the attention modules are incorporated be $l = h$ and $l = m$, which represent the high and medium attention respectively. Based on our design, $\hat{\beta}^h$ is a three dimensional tensor with $\hat{\beta}^h \in \mathbb{R}^{H^h \times W^h \times D^h} = \mathbb{R}^{64 \times 64 \times 32}$ and $\hat{\beta}^m \in \mathbb{R}^{H^m \times W^m \times D^m} = \mathbb{R}^{16 \times 16 \times 8}$.

We further calculate the element-wise multiplication of $h(x)^l$ and the estimated attention maps $\hat{\beta}^l$ across each channel:

$$o(x)_c^l = \hat{\beta}^l \odot h(x)_c^l, \tag{3.4}$$

where $o(x)_c^l$ is the c^{th} channel of the intermediate feature maps $o(x)^l$, $o(x)_c^l \in \mathbb{R}^{H^l \times W^l \times D^l}$;

$h(x)_c^l$ is the c^{th} channel of $h(x)^l$, $h(x)_c^l \in \mathbb{R}^{H^l \times W^l \times D^l}$, and $\odot$ is the elementwise multiplication operation.

The final output of the attention gate ($A(x)^l$) is a weighted average of the input intermediate feature maps x and $o(x)$:

$$A(x)^l = \gamma^l \times o(x)^l + (1 - \gamma^l) \times x^l, \tag{3.5}$$

where γ^l is a trainable scalar parameter initialized at zero.

3.2.6 Multi-scale guided attention (MSGA)

Loss function: We use the voxel-wise mean absolute error as the attention-based loss to measure the similarity between the estimated map of each sample ($\hat{\beta}_{ij}^l$) with the provided population-level maps ($\widetilde{DK}^l$):

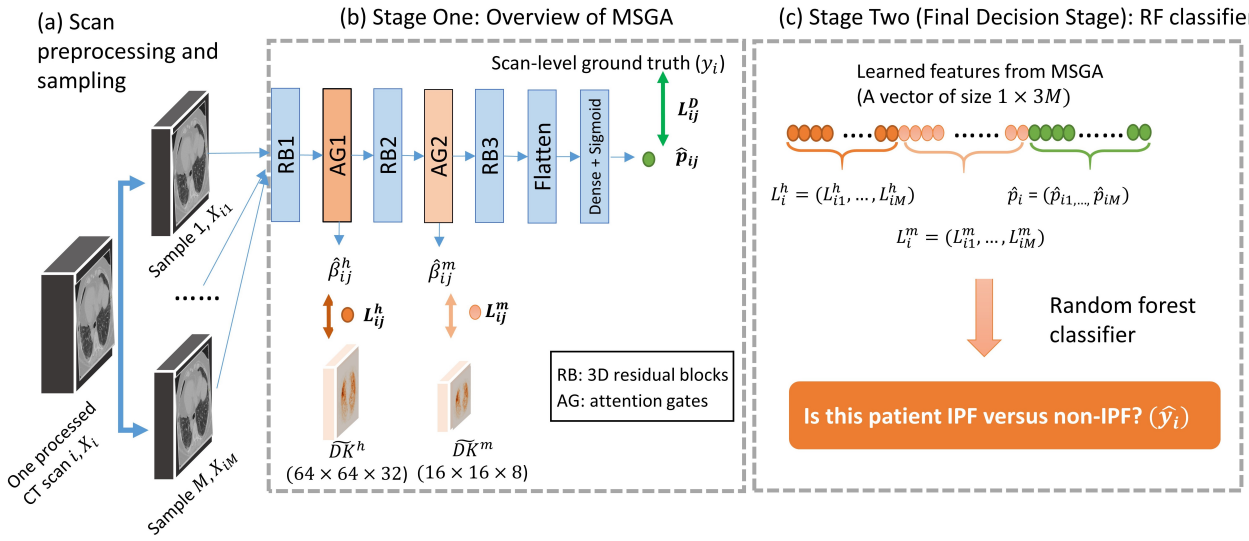
$$L_{ij}^l = avg(|\hat{\beta}_{ij}^l - \widetilde{DK}^l|), \tag{3.6}$$

where $\hat{\beta}_{ij}^l$ is the estimated attention maps for subject i and sample j at layer l, $\widetilde{DK}^l$ is the rescaled domain knowledge map at layer l that has the same dimension as $\hat{\beta}_{ij}^l$, and $avg(x)$ is the grand average of all elements from a tensor x.

During training, attention-based loss function is calculated by averaging all of the samples:

$$L^l = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} L_{ij}^l}{NM}. \tag{3.7}$$

In this work, we introduced two attention modules at a high- and medium- resolution scales; therefore, attention-based loss (L^l) is incorporated into the overall loss function under two forms: L^h and L^m, where h and m represent high and medium.



Notes: Firstly, a total number of M CT samples are generated from one processed CT scan i, X_i. The samples are presented as X_{ij}, where $j = 1, ..., M$. MSGA takes each sample X_{ij} as input and produces: the predicted probability of being IPF at the last layer ($\hat{p}_{ij}$) the estimated loss function at a high- (L_{ij}^h) and medium- (L_{ij}^m), and the estimated attention maps at a high- ($\hat{\beta}_{ij}^h$) and medium- ($\hat{\beta}_{ij}^m$) resolutions. At the final decision stage, RF takes the output from MSGA from all M samples and produces a patient-level diagnosis. RB: 3D residual blocks; AG: attention gates.

Figure 3.7: Schematic of the overall system for project II.

Explainability: The overall schematic diagram of MSGA is provided in Figure 3.7 (b). 3D-residual blocks are used as building blocks for our model, which is shown as RB1, RB2, and RB3 in Figure 3.7 (b). Detailed implementations of 3D-residual blocks, including layer name, hyperparameters, and output size, are provided in the Table 3.1.

For each scan i, we first produce M number of 3D samples for each scan, indexed by $j = 1, ..., M$. The system includes three types of input: the processed CT scans (X_i), the population-level domain knowledge maps at two resolution scales ($\widetilde{DK}^h$ and $\widetilde{DK}^m$), and the patient-level clinical ground truth (y_i). MSGA takes each sample as a training or testing unit and produces three types of output for each input sample: the sample-level predicted score of being IPF ($\hat{p}_{ij}$)the learned attention map at different resolution scales ($\hat{\beta}_{ij}^h$ and $\hat{\beta}_{ij}^m$) and the estimated attention-based loss values at two resolution scales (L_{ij}^h and L_{ij}^m). The attention gates are incorporated into the training of the IPF diagnosis model in an end-to-end manner,

Table 3.1: Model implementation details of MSGA, including layer name, hyperparameters, and output size.

		Hyperparameters			**Output Size**		
Layer	**Layer Name**	**RB1**	**RB2**	**RB3**	**RB1**	**RB2**	**RB3**
1	3D CONV	(5,5,5)@16	(5,5,5)@64	(5,5,5)@8	(32,64,64,16)	(8,16,16,64)	(2,4,4,8)
2	Batch normalization				(32,64,64,16)	(8,16,16,64)	(2,4,4,8)
3	3D CONV	(1,1,1)@8	(1,1,1)@16	(1,1,1)@4	(32,64,64,8)	(8,16,16,16)	(2,4,4,4)
4	Dropout	0.6	0.6	0.6	(32,64,64,8)	(8,16,16,16)	(2,4,4,4)
5	Batch normalization, ReLU activation				(32,64,64,8)	(8,16,16,16)	(2,4,4,4)
6	3D CONV	(3,3,3)@32	(1,3,3)@32	NA	(32,64,64,32)	(8,16,16,32)	NA
7	3D CONV	(1,1,1)@8	(1,1,1)@16	NA	(32,64,64,8)	(8,16,16,16)	NA
8	ADD: layer 5 + layer 7			NA	(32,64,64,8)	(8,16,16,16)	NA
9	Dropout	0.6	0.6	NA	(32,64,64,8)	(8,16,16,16)	NA
10	Batch normalization, ReLU activation			NA	(32,64,64,8)	(8,16,16,16)	NA

Note: The output size is represented as (H, D, W, C), which are the height, depth, width, and the number of filters of the intermediate feature maps, respectively. The hyperparameters in the 3D convolution layers (3D CONV) are presented in terms of (kernel size) @ the number of filters. The hyperparameter in the dropout layer is the dropout rate. RB1, RB2, and RB3 correspond to the three 3D-residual blocks of MSGA.

at two resolution scales, shown as AG1 and AG2.

Binary cross entropy loss is used for the IPF diagnosis task:

$$L^D = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}[y_i \log(\hat{p}_{ij}) + (1-y_i)\log(1-\hat{p}_{ij})], \tag{3.8}$$

where $y_i = 0, 1$ if the subject i is clinically diagnosed as non-IPF or IPF respectively, and $\hat{p}_{ij}$ is the predicted probability of subject i, sample j being IPF at the last layer of MSGA.

The overall loss function of the system is composed of a weighted average of two attention-based losses and one diagnosis-based loss:

$$L = L^D + \lambda^h L^h + \lambda^m L^m, \tag{3.9}$$

where L^D is the binary cross entropy for IPF diagnosis, L^h is the attention-based loss at a high resolution, L^m is the attention-based loss at a medium resolution. λ^h and λ^m are the relative task importance for the high- and medium- resolution attention model, respectively, with $\lambda^h \geqslant 0$ and $\lambda^m \geqslant 0$. We note that when setting $\lambda^h = 0$ and $\lambda^m = 0$, this represents a

scenario where both attention modules are unguided with population-level maps.

3.2.7 Random forests (RF)

Enhanced Improvement: Random forest (RF) is a popular supervised machine learning approach where the model output is decided based on majority voting of multiple decision trees [Bre01]. For a classification task, such as patient-level IPF diagnosis in our example, RF outputs the mode of the classes (IPF versus non-IPF) predicted by individual decision trees. It has been widely used in medical fields due to its high accuracies, robustness to outliers, explainable nature, and possibility to parallel processing [LWV14]. RF is chosen as the final stage classifier for this research since (1) it is easy to implement and computationally fast; (2) it can handle correlated variables, for example, in our case, the estimated attention loss from 20 samples; and (3) it is a relatively interpretable algorithm where the variable importance can be used to empirically understand the model decision process.

The intuition of adding RF in the final decision stage is that other than the predicted probability of IPF generated in the last layer of MSGA, we observe that the estimated attention-based loss (L_i^h and L_i^m) may also play a role in distinguishing between IPF and non-IPF. We provide a figure (Supplementary Figure C.1), which shows the distribution of the estimated attention loss values is visually different for IPF and non-IPF subjects. Therefore, for each CT scan i, we leverage these three types of information acquired from all samples, including the estimated high- ($L_i^h = (L_{i1}^h, ..., L_{iM}^h)$) and medium- ($L_i^m = (L_{i1}^m, ..., L_{iM}^m)$) resolution attention loss and the predicted probability of being IPF ($\hat{p}_i = (\hat{p}_{i1}, ..., \hat{p}_{iM})$), to build an RF model that classifies whether a given CT scan is from an IPF subject or a non-IPF ILD subject. For each scan, the designed MSGA produces a vector of size $1 \times M$ for L_i^h, L_i^m, $\hat{p}_i$, respectively, representing the estimated high-, medium attention-based loss function and the predicted IPF score from the M samples. This is later combined into a vector of size $1 \times 3M$, in our case, 1×60, as the input for the RF model, as shown in Figure 3.7.

After the training process of the MSGA is completed, we continue to build an RF-based

classifier for each hyperparameter selection (λ^h and λ^m) and for each fold. At each fold, we construct an RF using training samples only, not including validation or testing samples. For simplicity, we fix the hyperparameters during the training of RF for each model: RF classifier was consistently configured to use 90 decision trees with a maximum depth of 4.

3.2.8 Overall proposed method: MSGA+RF

We propose a two-stage model for scan-level IPF diagnosis.

Stage one (MSGA): for each CT scan i, MSGA provides (1) two estimated attention maps at a high- and medium- resolutions and (2) three outputs, including the loss function for high- (L_i^h) and medium- (L_i^m) attention gates, and the binary cross entropy loss for IPF diagnosis (L_i^D). The training process of stage one is end-to-end.

Stage two (RF): after finalizing the MSGA model, we move to the second stage. For each CT scan, RF takes the features produced by MSGA as input and produces the final probability of being IPF for each scan.

3.2.9 Explainability measures

Both qualitative and quantitative measures are utilized to examine the explainability of the developed models.

For qualitative measures, we plot the case-specific estimated attention maps $\hat{\beta}|x$ from Figure 3.6.

Certain quantitative measure is needed to evaluate the validity of explainable models. The intermediate attention gated output, $o(x)$ from Figure 3.6, is an elementwise attention weighted output. We chose $o(x)$ to visualize the discrepancy between subjects with IPF and non-IPF, on a population-level.

After MSGA models were constructed, we further calculated the average $o(x)$ map across the *validation* samples for IPF ($y = 1$) and non-IPF ($y = 0$) population for a certain channel c and model fold f (for a five-fold cross-validation, $f = 0, 1, 2, 3, 4$), shown as $\widetilde{o(x)}^y_{c,f}$. As shown

in Figure 3.3, the validation samples of the five folds of MSGA models are not overlapping.

We have

$$\widetilde{o(x)}^{y}_{c,f} = \frac{1}{M \times N_y^{val}} \sum_{j=1}^{M} \sum_{i=1}^{N_y^{val}} o(x)^{h,y}_{ij,c,f}, \tag{3.10}$$

where N_y^{val} is the number of validation samples for subjects with IPF ($y = 1$) or non-IPF ($y = 0$); $o(x)^{h,y}_{ij,c,f}$ is the observed $o(x)$ map for subject i (with clinical diagnosis y) and sample j at the high-resolution attention gate (layer h), under channel c and fold f. $o(x)^h$ is a four-dimensional tensor with eight channels, $o(x)^h \in \mathbb{R}^{64\times64\times32\times8}$, where channel is shown as the last dimension. For a specific channel, we have $o(x)^{h,y}_{ij,c,f} \in \mathbb{R}^{64\times64\times32}$. For example, for MSGA model 0 that was constructed from fold 0, $N^{val}_{y=1} = 56$ and $N^{val}_{y=0} = 85$.

We used the formula below for calculating kurtosis. For a sample of n values ($x_1, ..., x_n$):

$$\text{kurtosis}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3. \tag{3.11}$$

Furthermore, we calculated the kurtosis of the vectorized $\widetilde{o(x)}^{y}_{c,f}$:

$$\mu_{c,f,y} = \text{kurtosis}(\text{vec}(\widetilde{o(x)}^{y}_{c,f})), \tag{3.12}$$

where $\text{vec}(x)$ is the vectorization of a tensor x.

To analyze whether IPF and non-IPF subjects have a different kurtosis value of the marginal $o(x)$ map, we built a linear model as follows:

$$\mu_{c,f,y} = \beta + \alpha_{c,f} + \gamma_c + l_f + \delta_y + \epsilon_{c,f,y}, \tag{3.13}$$

where $\mu_{c,f,y}$ is the kurtosis of the vectorized $o(x)$ among validation samples for channel $c = 1, 2, ..., 8$, fold $f = 0, 1, 2, 3, 4$, and disease group y ($y = 1$ represents IPF and $y = 0$ denotes non-IPF). Both channels and folds are treated as categorical variables. β is an intercept term, $\alpha_{c,f}$ is the interaction effect of channel c and fold f, γ_c is the channel effect, l_f is the fold effect, δ_y is the disease diagnosis term that we are interested in, and $\epsilon_{c,f,y}$ is the

error term with constant variance $\epsilon_{c,f,y} \sim N(0, \sigma^2)$.

We conducted hypothesis testing for the regression coefficient for the disease type. Specifically, given all other factors constant, we tested if being IPF decreases the kurtosis of $o(x)$ map, compared with non-IPF (one-side test). That is, we set Non-IPF group ($y = 0$) as the reference group, i.e. $\delta_{y=0} = 0$, and test the following hypothesis:

$$H_0 : \delta_{y=1} \geqslant 0; H_1 : \delta_{y=1} < 0. \tag{3.14}$$

3.3 Experiments and results

3.3.1 Model implementation details

For model training, we used Adam optimizer with an initial learning rate of 10^{-4}, followed by an exponential decay after 20 epochs of decay rate 0.05. The batch size was set to be 5 and the model trained after 200 epochs was saved for evaluation. The hardware of Tesla V100-SXM2-32GB and GeForce RTX 2080 Ti and Keras framework were used. Sensitivity analysis of epoch numbers are included in Figure 3.4 (b). Model performance on the validation set increases as epochs increase but tends to stabilize after 100 epochs. At the same time, computational time increases linearly with more epochs. Therefore, we selected epochs=200 to acquire satisfactory model performance and save computational time.

3.3.2 Search ranges for the relative task importance in the loss function

The overall loss function for training MSGA contains three individual loss functions (i.e. learning three tasks), including one IPF diagnosis loss and two attention-based loss functions, see formula 3.9 for more details. The relative task importance for these three loss functions is controlled by two pre-specified hyperparameters: λ^h and λ^m. In this section, we provide some empirical evidence on how to determine the search ranges of these two hyperparameters. In this dissertation, three things should be taken into account when deciding the hyperparameter search ranges: 1. achieve satisfactory IPF diagnosis performance

(accuracy); 2. achieve satisfactory attention maps estimation (explainability); 3. consider the observed loss value ranges for three loss functions.

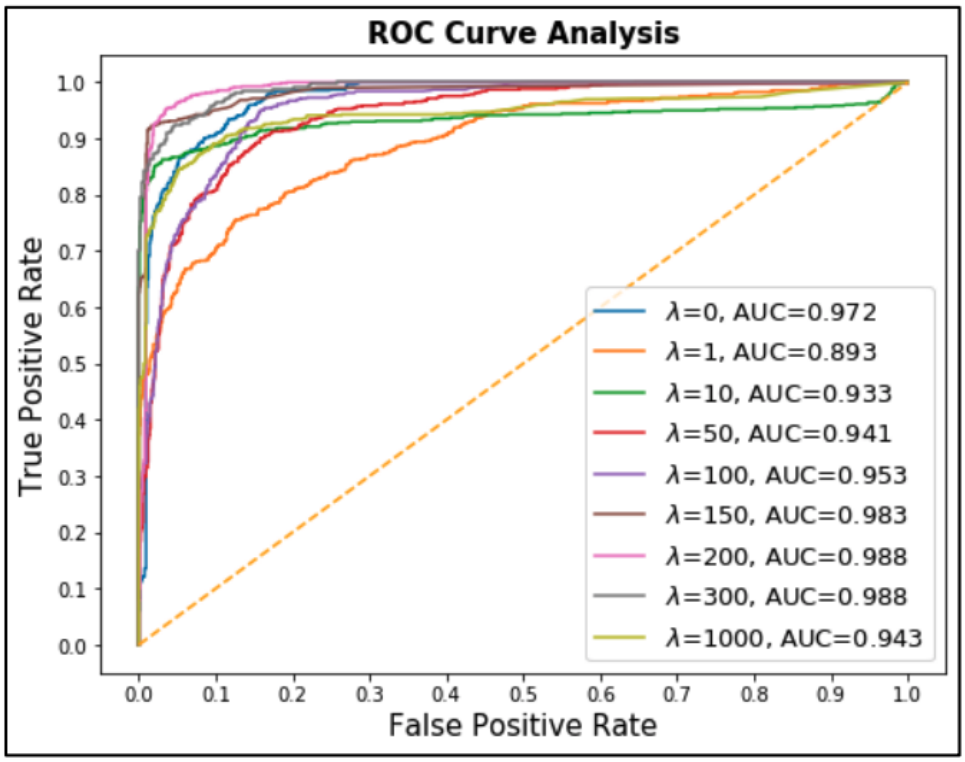
1. Accuracy: achieve satisfactory IPF diagnosis performance

To observe the potential impact of adding an attention module on the performance of IPF diagnosis, we first designed an *one-resolution* guided attention model with only one attention gate [YZC21]. Under this scenario, only the high-resolution attention gate is incorporated. Similar to MSGA, the overall loss function of one-resolution guided attention model is composed of both binary cross entropy loss (L^D) for IPF diagnosis and attention-based loss function (L^A):

$$L = L^D + \lambda L^A, \tag{3.15}$$

where λ is the relative task importance and $\lambda \geqslant 0$. Here only the high-resolution attention gate, which is similar to AG1 in Figure 3.7, is included in the overall design. The hyperparameter λ is analogous to λ^h in Formula 3.9.

Receiver operating characteristics (ROC) curves for the IPF diagnosis model on the test set (70 IPF and 106 non-IPF ILD subjects) with different selections of relative task importance are reported in Figure 3.8. This suggests that when no guided attention is included ($\lambda = 0$), IPF diagnosis model can achieve satisfactory AUC performance (AUC=0.972); when adding more emphasis on the attention modules, i.e. higher values of λ, AUC performance can increase up to 0.988; when λ boosts up to 1000, AUC performance decreases to 0.943. This implies that when putting excessive emphasis on estimating attention modules (in this case, $\lambda = 1000$), model accuracy can be hampered.



Notes: Under the one attention module scenario, only high-resolution attention gate is included. Therefore, λ is similar to λ^h in the MSGA setting (see Formula 3.9 for more details).

Figure 3.8: ROC curves for one attention module under different selection of relative task importance (λ) [YZC21].

2. Explainability: achieve satisfactory attention maps estimation

We provide the estimated attention maps for one-resolution attention module under different selections of relative task importance (λ) in Figure 3.9, using one randomly selected IPF subject as an example. By visual examination, a moderate value of relative task importance ($\lambda = 100$) can achieve good explainability by capturing lung parenchyma and concentrating on the peripheral regions, which is comparable to that of $\lambda = 1000$.

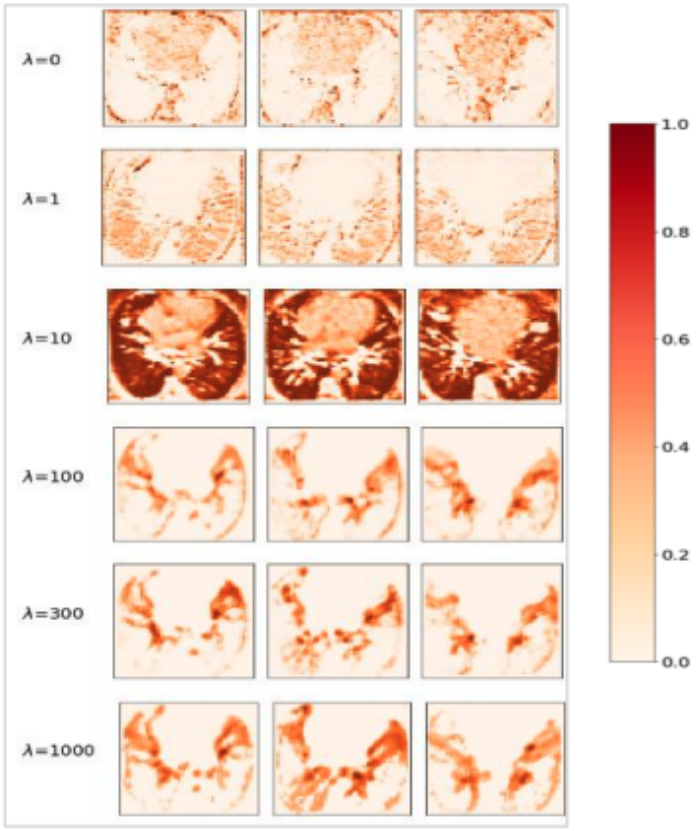


Figure 3.9: Estimated attention maps for one attention module using an randomly selected IPF subject under different selection of relative task importance (λ) [YZC21].

3. Consider the observed loss function ranges

The aforementioned discussions on one attention module provide a general guideline on choosing the hyperparameter λ. Specifically, a reasonable selection of λ, such as $\lambda = 100$, can achieve good model accuracy and explainability. We further extend the discussion to two-scale attention module, i.e. MSGA, and examine how the empirical loss function values change with different selections of hyperparameters.

Under the setting of MSGA, we report the observed ranges of three loss functions (L^D, L^h and L^m) under three hyperparameter selections ($a, \lambda^h = 1, \lambda^m = 1$, $b, \lambda^h = 1, \lambda^m = 100$, and $c, \lambda^h = 100, \lambda^m = 1$) in Table 3.2. We summarize the 2.5^{th} percentile and 97.5^{th} percentile of three loss functions in the training and validation samples, respectively, for the last 50 epochs. Only the last 50 epochs (epoch number 151-200) are reported since the values of the

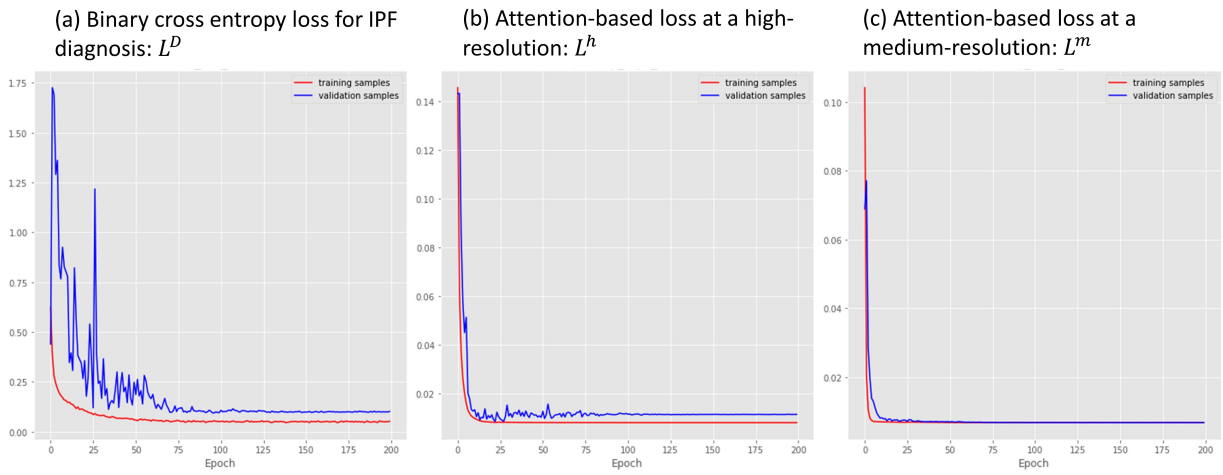
Table 3.2: Ranges of 95$^{\text{th}}$ percentile in the observed loss function values under three hyper-parameter selections for both training and validation samples.

		Training samples in the last 50 epochs (2.5$^{\text{th}}$ percentile, 97.5$^{\text{th}}$ percentile)	**Validation** samples in the last 50 epochs (2.5$^{\text{th}}$ percentile, 97.5$^{\text{th}}$ percentile)
a, λ^h=1, λ^m=1	L^D	(0.046, 0.054)	(0.097, 0.103)
	L^h	(0.008, 0.008)	(0.011, 0.011)
	L^m	(0.007, 0.007)	(0.007, 0.007)
b, λ^h=1, λ^m=100	L^D	(0.106, 0.116)	(0.127, 0.132)
	L^h	(0.007, 0.007)	(0.008, 0.008)
	L^m	(0.003, 0.003)	(0.008, 0.008)
c, λ^h=100, λ^m=1	L^D	(0.070, 0.080)	(0.134, 0.146)
	L^h	(0.006, 0.006)	(0.006, 0.006)
	L^m	(0.006, 0.006)	(0.020, 0.021)

Notes: Only one fold of the constructed model is presented in this table. L^D is the binary cross entropy loss for IPF diagnosis, L^h and L^m are the attention-based loss functions at a high- and medium-resolution, respectively. λ^h and λ^m are relative task importance for estimating high- and medium-resolution attentions, respectively.

loss functions are decreasing and not stabilized during the first 150 epochs, which is common in training machine learning methods. We provide the curves of three loss functions for one equal weight scenario ($\lambda^h = 1, \lambda^m = 1$) in Figure 3.10. As shown in this figure, for all three loss functions, the observed values from both training samples and validation samples decrease drastically at the beginning of the training process (first 100 epochs) and stabilize at a later stage (around last 50 epochs).

Empirically, we find that under the equal weight scenario, the observed loss functions of L^h and L^m across training samples are reasonably close in the last 50 epochs: L^h is close to 0.008 and L^m is close to 0.007. When we increase λ^m to emphasize more on estimating the medium resolution attentions ($\lambda^h = 1, \lambda^m = 100$ in Table 3.2), the observed medium-resolution loss decreases from 0.007 to 0.003.



Notes: Red and blue curves represent training and validation samples, respectively. Only one model fold is plotted.

Figure 3.10: Loss function curves for binary cross entropy loss (a) and attention-based loss (b and c) over 200 epochs under an equal weight scenario ($\lambda^h = 1$ and $\lambda^m = 1$).

In summary, taking the accuracy, explainability, and loss value ranges into consideration, we decide to conduct a grid search to select the relative task importance for MSGA: a series of λ^h and λ^m are tested, including 0, 1, 10, 50, 100, 200. The upper bound of 200 is selected since model accuracy can be hampered with excessive emphasis on attention modules, as shown in the one-resolution attention model (Figure 3.8). The lower bound of 1 is selected since the explainability performance can be hindered with a smaller value of λ (Figure 3.9). We observe that the estimated attention loss functions for high- and medium- resolutions are similar, under an equal weight scenario (Table 3.2).

3.3.3 MSGA model performance (Validation set performance)

In this section, we report the performance of MSGA using AUC from an ROC analysis reporting the classifier output at a sample-level.

Table 3.3 summarized the AUC values of MSGA with mean and standard errors (SE) across folds under the validation set, with different selections of hyperparameters (λ^h and λ^m). Both λ^h and λ^m are selected from a range of values: 0, 1, 10, 50, 100, 200. Similar work

Table 3.3: AUC mean and standard deviation values of **MSGA performance** on validation set for various λ^h and λ^m (task importance) parameters.

		λ^m					
		0	1	10	50	100	200
λ^h	200	0.87 (0.14)	**0.98 (0.02)**	0.88 (0.21)	0.89 (0.18)	0.87 (0.21)	0.97 (0.02)
	100	0.85 (0.20)	0.96 (0.04)	0.86 (0.20)	0.90 (0.10)	0.84 (0.21)	0.97 (0.03)
	50	0.83 (0.20)	0.88 (0.09)	0.89 (0.22)	0.84 (0.22)	0.97 (0.01)	**0.98 (0.02)**
	10	0.87 (0.21)	0.92 (0.09)	0.84 (0.17)	0.85 (0.21)	**0.99 (0.01)**	0.81 (0.23)
	1	0.87 (0.18)	0.84 (0.21)	0.95 (0.07)	0.89 (0.08)	0.89 (0.12)	0.76 (0.23)
	0	0.93 (0.07)	0.93 (0.07)	0.93 (0.09)	0.86 (0.15)	0.94 (0.04)	0.85 (0.21)

Note: λ^h and λ^m are the relative task importance parameters in the overall loss function, representing high- and medium- resolution attention, respectively. Three top performing combinations ($\lambda^h = 200$ and $\lambda^m = 1$; $\lambda^h = 50$ and $\lambda^m = 200$; $\lambda^h = 10$ and $\lambda^m = 100$) are in bold font.

which optimizes a multi-objective loss function utilizes hyperparameters within this range [LWP18, YKK19].

As shown in Table 3.3, without including guided attention by attention-based loss function ($\lambda^h = 0$ and $\lambda^m = 0$), the IPF diagnosis model reached an AUC value of $AUC \pm SE = 0.93 \pm 0.07$. Only incorporating guided high- ($\lambda^h > 0$ and $\lambda^m = 0$) or medium-resolution attention ($\lambda^h = 0$ and $\lambda^m > 0$) decreased the performance of IPF diagnosis, compared to without guided attention in the loss function ($\lambda^h = 0$ and $\lambda^m = 0$).

Our proposal, which included both high- and medium- resolution attentions, was able to reach the highest AUC value ($AUC \pm SE = 0.99 \pm 0.01$) for all of the experiments, under certain hyperparameter selections ($\lambda^h = 10$ and $\lambda^m = 100$). Notably, model performance is sensitive to the selection of relative task importance. For example, under certain hyperparameter combinations, i.e. $\lambda^h = 1$ and $\lambda^m = 200$, the AUC decreased to AUC$\pm$SE=0.76$\pm$0.23.

Table 3.4: AUC mean and standard deviation values of **MSGA+RF performance** on validation set for various λ^h and λ^m (task importance) parameters.

		λ^m					
		0	1	10	50	100	200
λ^h	200	0.95 (0.04)	**0.98 (0.01)**	0.99 (0.01)	0.97 (0.01)	0.97 (0.04)	0.98 (0.02)
	100	0.97 (0.03)	0.98 (0.02)	0.97 (0.03)	0.95 (0.06)	0.96 (0.04)	0.97 (0.02)
	50	0.97 (0.03)	0.96 (0.03)	0.97 (0.03)	0.94 (0.05)	0.97 (0.02)	**0.98 (0.02)**
	10	0.95 (0.06)	0.98 (0.02)	0.97 (0.03)	0.95 (0.05)	**0.99 (0)**	0.96 (0.02)
	1	0.99 (0.02)	0.98 (0.02)	0.97 (0.05)	0.94 (0.05)	0.97 (0.03)	0.92 (0.08)
	0	0.97 (0.03)	0.98 (0.01)	0.99 (0.01)	0.94 (0.04)	0.95 (0.03)	0.95 (0.06)

Note: λ^h and λ^m are the relative task importance parameters in the overall loss function, representing high- and medium- resolution attention, respectively. Three top performing combinations based on MSGA ($\lambda^h = 200$ and $\lambda^m = 1$; $\lambda^h = 50$ and $\lambda^m = 200$; $\lambda^h = 10$ and $\lambda^m = 100$) are in bold font.

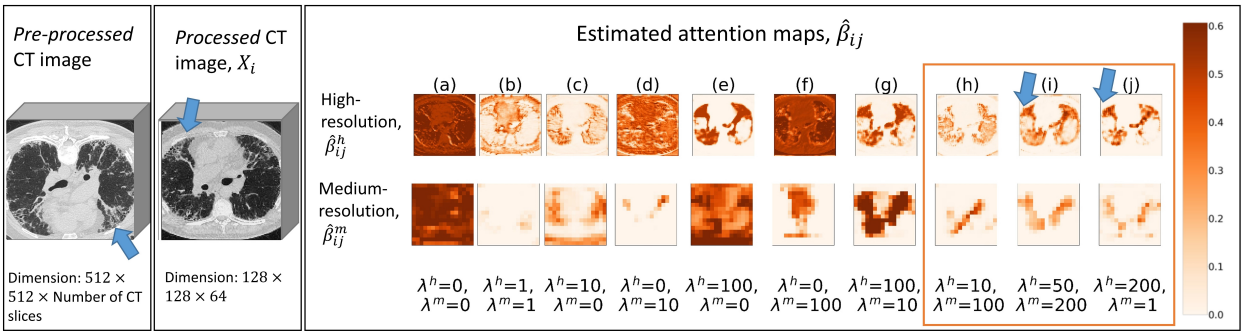
3.3.4 MSGA+RF model performance (Validation set performance)

Table 3.4 summarized the model performance using MSGA+RF with mean and SE across five folds under the validation set, under different selections of hyperparameters (λ^h and λ^m). Top three hyperparameter selections based on MSGA remained one of the best performing hyperparameter groups for MSGA+RF (average $AUC \geqslant 0.98$); therefore, these three models were selected as best performing models and were used as the final models for this task.

3.3.5 Explainability measures

We report the explainability measures using a qualitative approach, by visualizing the estimated attention maps on several cases, and a quantitative approach, by calculating the kurtosis of attention gated output.

1. **Qualitative measures: estimated attention maps**



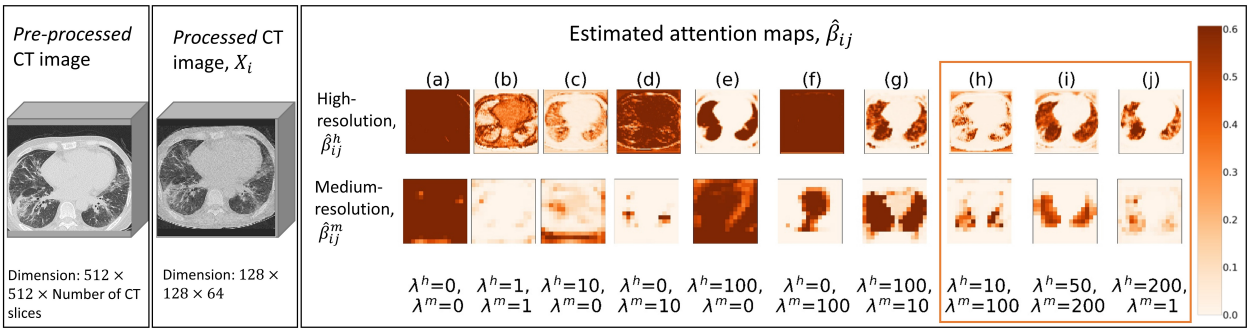
Notes: One representative CT slice (slice number=153; in total 309 slices for this scan) of the pre-processed image is provided. One processed CT image is plotted at D=33 out of 64. The estimated attention maps for high- and medium- resolutions are plotted at D=17 out of 32 and D=5 out of 8, respectively. Key CT features of UIP are highlighted as arrows. The three final models are highlighted as an orange rectangle. The models that used this scan as validation samples were selected for plotting. For all ten hyperparameter collections (λ^h and λ^m), both MSGA and MSGA+RF successfully classify this scan as IPF (true positives).

Figure 3.11: Pre-processed, processed CT image, and the estimated attention maps under ten hyperparameter selections (λ^h and λ^m) for one randomly sampled **IPF subject**.

We explored the model explainability by plotting the estimated attention maps at both high- and medium- resolutions ($\hat{\beta}^h_{ij}, \hat{\beta}^m_{ij}$) using one randomly sampled IPF as an example, shown in Figure 3.11. We also provided one non-IPF ILD subject in Figure 3.12. We note that without guided attention models (Figure 3.11 column a), the observed attention maps are uninformative and lack explainability.

When we provide the guidance from population-level domain knowledge in constructing the overall loss function, the estimated attention maps begin to focus on the lung parenchyma. Specifically, when the relative task importance is low (Figure 3.11 column b), the attention maps begin to concentrate on the lungs, but it is not clear. When we add solely the high-resolution guided attention in the loss function (Figure 3.11 column c and e), visual examinations indicate that high-resolution attention maps can characterize the lungs, while the medium-resolution attention maps are less informative. On the other hand, when only medium-resolution guidance are added (Figure 3.11 column d and f), both high- and medium- resolution attention maps do not concentrate on the lung parenchyma.

Finally, when we provide guidance on both high- and medium- resolution attentions with a considerable relative task importance (Figure 3.11 column g, h, i, and j), the estimated attention maps become instructive, focus on the lung parenchyma, and suppress irrelevant background areas. Under certain hyperparameter collection (Figure 3.11 column i and j), both the estimated attention map and a high- and medium- resolution can focus on peripheral lungs, which are the key regions for making a correct IPF diagnosis. These highlighted areas are critical for this task of IPF diagnosis and are incorporated into the training of the deep learning systems.



Notes: One representative CT slice (slice number=38; in total 62 slices for this scan) of the pre-processed image is provided. One processed CT image is plotted at D=48. The estimated attention map is plotted for high- and medium- resolutions at D=25 out of 32 and D=6 out of 8, respectively. The models that used this scan as validation samples were selected for plotting. For all ten hyperparameter collections (λ^h and λ^m), both MSGA and MSGA+RF successfully classify this scan as non-IPF ILD (true negatives).

Figure 3.12: Pre-processed, processed CT image, and the estimated attention maps under ten hyperparameter selections (λ^h and λ^m) for one randomly sampled **non-IPF subject**.

2. Quantitative measures: kurtosis results

Using one hyperparameter combination $\lambda^h = 200, \lambda^m = 1$ and one specific fold (fold=0) as an example, the calculated kurtosis of the vectorized marginal $o(x)$ among validation samples is reported as below in Table 3.5. There are four types of attention maps voxels: all voxels from $o(x)$ for IPF subjects ("IPF all", Number of voxels=$64 \times 64 \times 32 = 131,072$), all voxels from $o(x)$ among non-IPF subjects ("Non-IPF all", Number of voxles=$131,072$), voxels from peripheral lungs from IPF subjects ("IPF peri", Number of voxels varies based

Table 3.5: Kurtosis results using one model at one fold ($\lambda^h = 200, \lambda^m = 1,$ fold 0) as an example.

Channel	Voxel types				Compare results	
	IPF all	IPF peri	Non-IPF all	Non-IPF peri	IPF all < Non-IPF all	IPF peri < Non-IPF peri
0	0.72	-0.05	2.70	1.03	**True**	**True**
1	-0.50	-0.80	0.15	-0.60	**True**	**True**
2	2.29	0.88	9.59	5.23	**True**	**True**
3	10.14	5.99	16.00	9.52	**True**	**True**
4	1.49	0.47	4.50	2.23	**True**	**True**
5	0.77	-0.05	-0.16	-1.00	False	False
6	6.20	5.78	12.30	7.83	**True**	**True**
7	-0.41	-0.96	0.50	-0.60	**True**	**True**

Table 3.6: Hypothesis testing for the covariates of clinical diagnosis ($\delta_{y=1}$) in influencing kurtosis of $o(x)$

Model	Voxel type	Estimates (SE)	P value
$\lambda^h = 50, \lambda^m = 200$	all voxels	-0.62 (0.15)	$<$ **0.001**
	peripheral	-0.48 (0.11)	$<$ **0.001**
$\lambda^h = 200, \lambda^m = 1$	all voxels	-0.90 (0.44)	**0.02**
	peripheral	-0.47 (0.27)	**0.05**
$\lambda^h = 10, \lambda^m = 100$	all voxels	0.44 (0.48)	0.82
	peripheral	0.08 (0.30)	0.61

on lung segmentation results), and voxels from peripheral lungs from Non-IPF subjects ("Non-IPF peri").

For this specific model fold ($\lambda^h = 200, \lambda^m = 1$, fold 0), we observe that for seven out of eight channels, kurtosis of the average of $o(x)$ among IPF subjects is less than that of non-IPF subjects, for both all voxels and peripheral lungs only. We provide a systemic statistical analysis across all models and all folds below.

For each of the top three hyperparameter combination (λ^h and λ^m), we constructed two linear models, one for all voxels and the other for peripheral voxels. The estimated $\hat{\delta}_{y=0}$, its corresponding standard errors (SE), and one-sided P value is reported in Table 3.6. Due to the violation of normality assumption, we also used box-cox transformations and applied log transformations on the shifted kurtosis, as shown in the Table 3.7.

Table 3.7: **After a log transformation on the shifted kurtosis:** Hypothesis testing for the covariates of clinical diagnosis ($\delta_{y=1}$) in influencing kurtosis of $o(x)$

Model	Voxel type	Estimates (SE)	P value
$\lambda^h = 50, \lambda^m = 200$	all voxels	-0.48 (0.13)	$<$ **0.001**
	peripheral	-0.11 (0.09)	0.13
$\lambda^h = 200, \lambda^m = 1$	all voxels	-0.41 (0.16)	**0.007**
	peripheral	0.02 (0.19)	0.55
$\lambda^h = 10, \lambda^m = 100$	all voxels	0.30 (0.19)	0.94
	peripheral	0.28 (0.18)	0.93

3.3.6 Test set performance

Based on the validation set performance and the estimated attention maps, we applied the three best performing models to the holdout test set (N=176). The three best performing models (i.e. (1) $\lambda^h = 200$ and $\lambda^m = 1$; (2) $\lambda^h = 50$ and $\lambda^m = 200$;(3) $\lambda^h = 10$ and $\lambda^m = 100$) received an AUC value of $AUC \pm SE = 0.987 \pm 0.007, 0.975 \pm 0.011, 0.980 \pm 0.018$, respectively on the holdout test set.

3.4 Discussions and conclusions

In this chapter, we presented a two-stage model for automated IPF diagnosis among subjects with ILD based on axial chest high-resolution CT images. The model combines a multi-scale guided attention network, MSGA, for explainability and a random forest, RF, model for enhancing accuracy in the final decision. This network is generally suitable for weakly supervised tasks, with only scan-level labels available. Several advantages can be addressed using MSGA+RF. Firstly, population-level domain knowledge from the prior studies is more accessible, whereas acquiring well-labeled fine-scale medical imaging data is time-consuming and labor-intensive. Guided with population-level domain knowledge at various resolution scales, we can accomplish satisfactory model performance only using the clinical information of IPF diagnosis in subjects with ILD. Secondly, using attention models at various resolution scales increase model explainability, which is a crucial step for building trust in the medical imaging domain.

Explainability: Over the past decade, there has been extensive discussions regarding enhancing the explainability for deep learning-based systems, especially in clinical settings [LGL20]. Building explainable deep learning models can increase model trust and it is a critical step for model diagnostics. Saliency maps [SZ14], class activation mapping [ZKL16] are effective post-hoc methods for visualizing deep learning models; attention mechanisms, on the other hand, can encourage the network focus on specific areas of interests (in our case, lung parenchyma) in a trainable and end-to-end manner.

Accuracy: To boost model performance, a traditional machine learning tend to increase a model accuracy by adding model features in a classifier [HTF09]. We borrowed a similar idea here by adding RF classifiers using the features sets learned from the estimated loss function of learning from MSGA, as the final decision stage. This is necessary since we note that results on the validation set are sensitive to the selection of relative task importance (i.e. λ^h and λ^m). For example, in Table I, among $6 \times 6 = 36$ hyperparameter combinations, 7 out of 36 combinations have a mean AUC less than 0.85 using stratified five-fold cross validation on the validation sets. However, after adding the RF classifier, as results shown in Table II, all of 36 combinations have a mean AUC greater than 0.92. Therefore, in our example, having a two-stage model increases the model robustness against the changes regarding relative task importance. We hypothesize that this phenomenon is due to the fact that RF further leverages the information from the attention-based loss functions to make a more reliable patient-level diagnosis. We provide more detailed explanations using variable importance plots in Supplementary C.1.

A good model accuracy does not guarantee satisfactory model explainability on the validation set, and vice versa. For example, based on our explorations, when $\lambda^h = 0$ and $\lambda^m = 0$, the model lacks explainability, but can perform well on the validation set ($AUC \pm SE = 0.93 \pm 0.07$). It is of research interests to compare the generalizability to new prospective studies between unexplainable models (for example, $\lambda^h = 0$ and $\lambda^m = 0$) and explainable ones.

We designed a two-stage model that combines explainability achieved by a deep learning approach, MSGA, and accuracy by a machine learning technique, RF. Strengthened by the

combined benefit of transparent model decision process and boosted diagnostic performance, the proposed method serves as an important step for clinical applications.

Certain limitations exist in this work: (1) the current MSGA setup requires population-level domain knowledge acquired from previous studies; (2) only volumetric CT scans with consistent slice spacing were included in the training and testing sets, which limited the applicability of this trained model to other non-volumetric CT scans; (3) the selections of relative task importance (λ^h and λ^m) requires extensive computational time and resources; (4) although some research works demonstrated the superior generalizability of attention models to unseen datasets [JLL18], the evaluation of our proposed model to independent datasets is underway and is out of scope of this project.

In this chapter, we have demonstrated that MSGA+RF is one promising method for both enhancing explainability and increasing in the performance of model for the task of automated IPF diagnosis using CT images only. Future work includes examining the trained MSGA+RF on independent test set and prospective studies.

CHAPTER 4

Robustness tests: Evaluate the model robustness under different CT imaging protocals

4.1 Background

There has been a surge of work in enhancing the explainability and interpretability of machine learning methods. The definitions of these terms may be slightly different from multiple papers. In this dissertation, we borrow the concept defined in [MSM18]:

Explanations: "An explanation is the collection of *features* of the interpretable domain, that have contributed for *a given example* to produce a decision" [MSM18]. Explanations are (or can be rescaled to) the same size as the input and can provide certain scores suggesting the extent of contributions for each feature. In this dissertation, Grad-CAM plots in project I and the estimated attention maps in project II both belong to the category of explanations. They both provide a score for each (training or testing) case indicating the level of contributions to the model decision process (i.e. IPF diagnosis).

Interpretations: "An interpretation is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of" [MSM18]. In this dissertation, we focus on *post-hoc interpretations*, which aim to extract information from learned models [Lip18]. Specifically, our method for providing post-hoc interpretations is similar to the subcategory of "explanation by example" discussed in the paper [Lip18], which is a mechanism that can provide other examples which are most similar to the given example, from the model perspective. In this chapter, we describe a method that provides the model with similar images and therefore produces post-hoc explanations on how model diagnosis

results change with a different set of imaging protocols.

Deep learning (DL) has prospered in the field of medical imaging in recent years, among various tasks, including diagnosis, segmentation, detection, etc [AKB19, LYM19, RFB15]. Many state-of-the-art DL algorithms were reported to perform on par with experienced radiologists [LFK19]. Traditional machine learning approaches, including DL methods, usually assume that the training data should be representative of the testing data. However, in clinical practice, it is often unrealistic to have testing cases that follow the same distributions as that of the training scenarios at all times. When deploying a DL-based system into clinical practice, many imaging acquisition factors are subject to change according to the specific sites, protocols, or the preference of specific practitioners, including slice thickness, effective mAs, patient positions, etc.

This phenomenon is known as *dataset shift*, which occurs when the joint distribution between the inputs (in our case, CT images) and the outputs (in our case, the clinical diagnosis of IPF versus non-IPF) differs between the training and testing cases [QSS09]. Due to this reason, many researchers have reported the lack of generalizability in the DL models in clinical settings, which is the decrease in model performance when deploying a well-trained model to an external test set [ZBL18, PAM19, WLZ20]. However, few research efforts have quantitatively analyzed the factors that lead to this decrease in model performance. Recently, Badgeley et al. examined patient data and hospital process features and suggested that these variables were the main source that contributed to the success of a deep leaning model, other than patients' imaging features [BZO19]. Therefore, we extended this line of research to evaluate the robustness of the developed model using available CT scans that were acquired under different sets imaging protocols. We applied statistical methods to analyze the factors that may lead to this inconsistency in the model performance, using the IPF diagnosis task as an example.

Consistent with Project I and II, the major scientific question of this line of research is to distinguish IPF from non-IPF among ILD subjects based on chest CT scans. This project, in particular, focuses on assessing the robustness of several constructed IPF diagnosis models on different CT technical parameters.

In this chapter, we evaluate the robustness of three pretrained and high performing DL-based IPF diagnosis model under different sets of imaging protocols. We included one 2D DL model (project I) that uses ResNet-50 as its backbone structure [YZG21] and two 3D DL models (project II) that use multi-scale guided attention models [YZC21]. These two 3D DL models performed equally well on the validation cases (sensitivity = 0.97 for both models), but was trained with different hyperparameters that controlled the relative task importance for estimating attention maps.

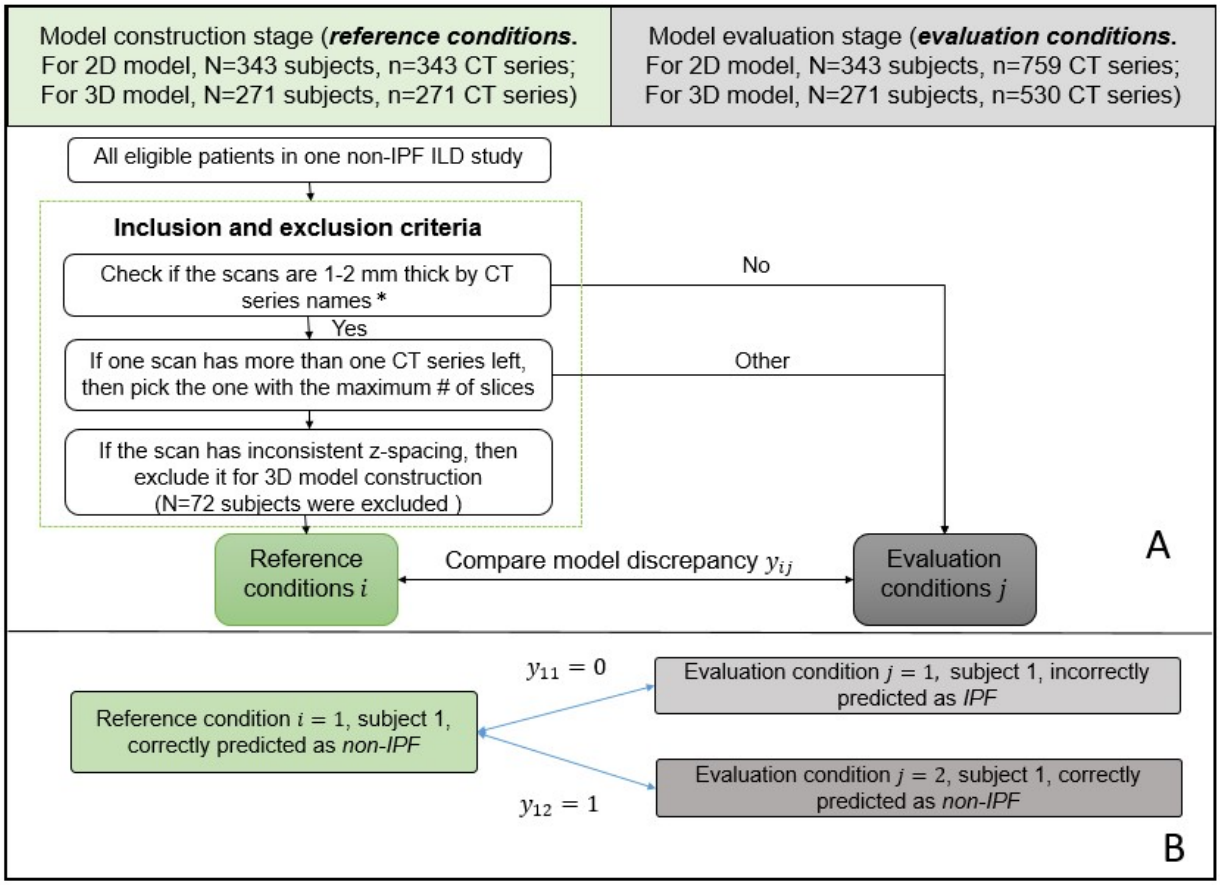
According to the study protocol, we have acquired some CT scans where the same patient was scanned multiple times in different patient positions (i.e. both feet first supine and feet first prone) and the image dataset may have different protocols (lower dose for feet first prone than for feet first supine) and have multiple reconstructions (1 mm and 5 mm for the same scan). Selected imaging protocols include effective tube current-time product (known as "effective mAs", which is related to scanner reported dose and patient absorbed dose), reconstruction kernels, slice thickness, patient positions, manufacture model name, and clinical diagnosis.

To leverage this information, we compare the performance of *paired CT series* acquired on the same patient under the model construction stage (reference conditions) and the model evaluation stage (evaluation conditions). Statistical methods are used to analyze the factors that are associated with this change in model diagnosis results.

4.2 Materials and methods

4.2.1 Datasets

Figure 4.1 provides an overview of the study design. We define the entire study as two phases: the model construction stage and the model evaluation stage, where the latter one is the research focus of the project.



Notes: Model discrepancy measure $y_{ij} = 0$ if the reference condition i and evaluation condition j received conflicting model predicting results; $y_{ij} = 1$ if both CT series received consistent predictive results. *N=8 (2%) CT scans that are 3-5 mm thick were included as reference conditions due to CT series naming styles.

Figure 4.1: Overview of the study design for robustness tests. A, inclusion and exclusion criteria for the construction of reference conditions and evaluation conditions. B, an example of two pairs of CT series constructed using one reference condition and two evaluation conditions collected from one patient.

At the model construction stage, one CT series per patient was used to build the IPF diagnosis model. If the patient had more than one CT scans, then the first available total lung capacity scans was selected. In total, there were 389 IPF patients (defined as "positives" in this project) and 700 non-IPF ILD patients (defined as "negatives") included, which were retrospectively obtained from five multi-center studies. Notably, for one non-IPF ILD cohort, according to the protocol, one subject was scanned and/or reconstructed under multiple

conditions. We thereby devised the inclusion and exclusion criteria, as shown in Figure 4.1 A, to select and use one CT series per subject for the model construction stage and utilize the remaining CT series to evaluate the model robustness, which is referred to as model evaluation stage.

We defined the CT series that were included in the model construction stage and model evaluation stage for this non-IPF ILD cohort as reference conditions and evaluation conditions, respectively. In detail, there are 343 subjects (343 CT series) included in the reference conditions; of these same 343 subjects, there are an additional 759 CT series included in the evaluation conditions. Among these 343 non-IPF ILD subjects, there are four subtypes of ILD subjects, including 122 (35.6%) rheumatoid arthritis (RAILD), 91 (26.5%) hypersensitivity pneumonitis (HP), 80 (23.3%) systemic sclerosis (SSc-ILD), and 50 (14.6%) Sjögren's syndrome (SjS-ILD). For 3D model training and testing, we added one additional criterion to exclude CT scans without consistent z-spacing (N=72 scans were excluded). Due to this reason, the number of paired evaluation CT series in the 3D models (n=530 CT series) is also lower than that of 2D models (n=759 CT series).

To evaluate the model performance of the constructed model under varying CT imaging protocols and determine the possible factors that may cause this discrepancy, we constructed n=759 "paired CT series" from N=343 patients between the reference and evaluation conditions, for 2D models. In more detail, N=64 patients (18.7%) have only one CT pair, N=170 subjects (49.6%) have two CT pairs, N=86 subjects (25.1%) have three CT pairs, N=18 subjects (5.2%) have four CT pairs, and N=5 subjects (1.5%) have five CT pairs. Specifically, for each patient, we constructed paired CT series (denoted as ij) between its reference condition i and each of the evaluation condition $j = 1, 2, \ldots, n_i$, where n_i is the number of evaluation conditions matched with the reference condition i. Suppose one patient has one reference CT series (denoted as $i = 1$) and two evaluation CT series (denoted as $j = 1, 2$), then two paired CT series should be constructed $i = 1, j = 1$ and $i = 1, j = 2$, as shown in Figure 4.1 B. Since the research interest is to determine whether similar predictive performance could be achieved under reference and evaluation conditions, the change of model predictive result for the pair ij is used as the dichotomous outcome measure y_{ij}, where $y_{ij} = 1$ and 0 if this

CT pair ij gets different or identical predictive result for the reference condition i and the evaluation condition j, respectively.

CT scans of one patient visit, including one reference CT series and three evaluation CT series, are provided in Figure 4.2. We also provide the corresponding reconstruction kernel, slice thickness, effective mAs at this slice, the average effective mAs across the scan, and patient positions in Table 4.1. Notably, reference condition (a), evaluation condition (b) and (c) were collected from one scan, but were reconstructed differently according to the slice thickness, leading to a visible difference. Evaluation condition (d) was collected from the same subject on the same date, but with a different patient position and effective mAs. For example, as shown in Figure 4.2, evaluation condition (d), which was scanned under an effective mAs of 104 mAs at this CT slice, contained more noise, especially outside of the lung parenchyma, compared to that of other three CT slices, which were scanned under a dose of 113-114 mAs. Therefore, the distinctions between the reference conditions and evaluation conditions are often visible and we sought to examine whether the constructed DL models can remain robust under these variations.

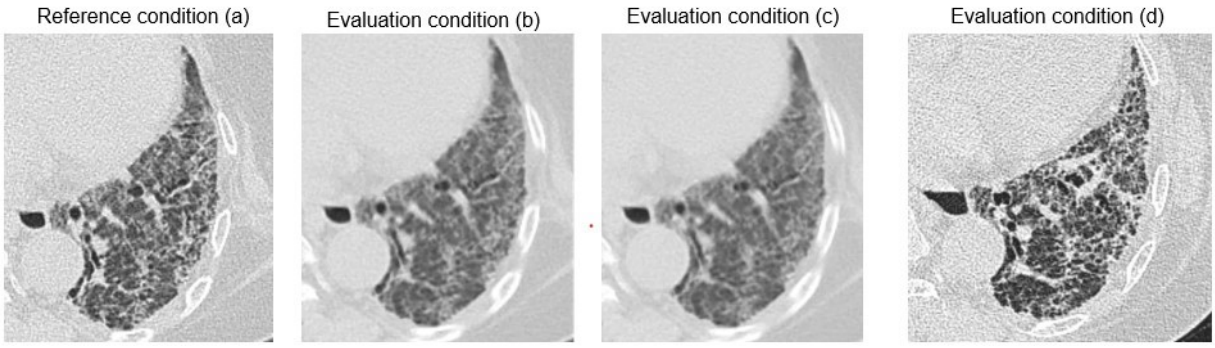


Figure 4.2: CT scans of a typical patient evaluated under four conditions, including a reference condition (a) and three evaluation conditions (b, c, and d). Detailed imaging protocols are provided in Table 4.1.

Table 4.1: CT technical information of CT scans of a typical patient evaluated under four conditions.

	Reference condition (a)	Evaluation condition (b)	Evaluation condition (c)	Evaluation condition (d)
Reconstruction kernel	B60f	B50f	B50f	B60s
Slice thickness (mm)	1	3	5	1
Effective mAs at this slice	113	114	114	97
Average effective mAs for this scan	127	127	127	104
Patient position	Supine	Supine	Supine	Prone

4.2.2 Model construction stage: DL-based algorithms

Three models were developed, including one 2D-based model (project I) and two 3D-based models (project II). For 2D models, ResNet-50 [HZR16] with pre-trained model weights from ImageNet [DDS09] were utilized as the backbone DL-based algorithm for the IPF diagnosis task. For each CT scan, we randomly sampled a fixed number (we empirically chose 20) of triplets and used one triplet as a training or testing sample, where each triplet was composed of three CT slices collected from the top, middle, and bottom of both lungs. We provide some sample triplets images in Figure 2.2 from Project I. The predicted result for each CT scan was decided by the majority voting of all of the sampled triplets. In our case, since we sampled 20 triplets per scan, if more than ten triplets were predicted as IPF, then we classified this scan as an IPF patient.

For 3D models, we used multi-scale guided attention networks with residual building blocks [YZC21]. 3D attention models were utilized since attention models have been known for its ability to capture the regions of interest (in our case, lung parenchyma) and increase model generalizability to unseen domains [JLL18].

For both 2D and 3D models, stratified five-fold cross validation was used during this process, where all subjects were randomly separated into five subsets, while fixing the proportion of IPF versus non-IPF subjects for each subset. At each fold, four subsets of data were used to train the model and the remaining one was used to test the model. The results were reported based on the test fold. The hardware of GeForce RTX 2080 Ti, Tesla

Table 4.2: Summary of the technical and clinical parameters for the reference and evaluation conditions.

	2D models		3D models	
	Reference conditions	**Evaluation conditions**	**Reference conditions**	**Evaluation conditions**
Number of patients (N), number of CT series (n)	N=343 n=343	N=343 n=759	N=271 n=271	N=271 n=530
Manufacturer models	Siemens Sensation 16 (53.4%), Siemens Definition (46.6%)	Siemens Definition (57.2%), Siemens Sensation 16 (42.8%)	Siemens Sensation 16 (53.5%), Siemens Definition (46.5%)	Siemens Definition (55.1%), Siemens Sensation 16 (44.9%)
Slice thickness (mm)	1∼2 (97.7%), 5 (2.0%), 3 (0.3%)	5 (48.9%), 1∼2 (32.7%), 3 (18.4%)	1∼2 (97.4%), 5 (2.2), 3 (0.4)	5 (49.6%), 1∼2 (30.2%), 3 (20.2%)
Reconstruction kernels	B60f and B60s (76.4%), B70f (17.2%), Iterative (4.1%), B50f (2.3%)	B50f (59.4%), B60f and B60s (28.2%), Iterative (6.2%), B70f (6.2%)	B60f and B60s (73.1%), B70f (19.2%), Iterative (5.2%), B50f (2.6%)	B50f (66.6%), B60f and B60s (27.7%), B70f (5.7%)
Patient positions	Supine (69.7%), Prone (30.3%)	Supine (83.4%), Prone (16.6%)	Supine (86.3%), Prone (13.7%)	Supine (77.2%), Prone (22.8%)
Average effective mAs per scan (mean ± standard errors)	93.6 ± 23.7	95.5 ± 25.1	94.4 ± 24.4	93.5 ± 24.1
Clinical disease diagnosis	RAILD (35.6%), HP (26.5%), SSc-ILD (23.3%), SjS-ILD (14.6%)	RAILD (32.0%), HP (29.6%), SSc-ILD (24.0%), SjS-ILD (14.4%)	RAILD (37.3%), HP (26.6%), SSc-ILD (21.4%), SjS-ILD (14.8%)	RAILD (32.6%), HP (29.2%), SSc-ILD (23.2%), SjS-ILD (14.9%)

Note: ILD, interstitial lung disease; RAILD, rheumatoid arthritis-associated ILD; HP, hypersensitivity pneumonitis; SSc-ILD, Systemic sclerosis-associated ILD; SjS-ILD, Sjögren's syndrome-associated ILD.

V100-SXM2-32GB and Keras framework was used.

4.2.3 Technical and clinical parameters

Both reference conditions and evaluation conditions contained a heterogeneous set of technical and clinical parameters, including CT manufacturer model name, slice thicknesses, reconstruction kernels, patient positions, effective mAs, clinical diagnoses, etc. Technical parameters were extracted from the DICOM header files for each CT series. Table 4.2 summarizes the different technical and clinical parameters in the reference and evaluation conditions, for both 2D and 3D models.

4.2.4 Statistical analysis

As noted before, a dichotomous outcome measure y_{ij} is used to evaluate the predictive performance consistency, where $y_{ij} = 1$ and 0 if this CT pair ij gets different or identical predictive result for the reference condition i and the evaluation condition j, respectively.

Generalized linear mixed effects models (GLMM) with a logit link function were utilized to determine the contributing factors that influenced the dichotomous outcome measure y_{ij} using the lme4 package of the R Software [BMB18]. Due to the hierarchical nature of the dataset, i.e. multiple CT pairs may be collected from the same patient, the independent sample assumption is violated. Therefore, GLMM, instead of generalized linear models (GLM), was chosen for this research to account for the hierarchical data structure. Several covariates were included as the fixed effects, including reconstruction kernels, slice thicknesses, patient positions, manufacturer model names, clinical diagnosis, average effective mAs in the evaluation condition, and mean effective mAs between the reference and evaluation condition. The odds of observing conflicting predictive results between reference and evaluation conditions were predicted using a pre-specified baseline category: evaluation conditions with reconstruction kernel at B50f, 1-2 mm CT slices, prone positions, with CT model of Siemens Sensation 16, and a clinical diagnosis of hypersensitivity pneumonitis (HP). We also included a patient level random intercept, which represents the patient-level influence on the paired CT series that is not captured by the fixed effects.

For each CT pair ij, we calculated the differences regarding the mean effective mAs, ΔE_{ij} , which is the difference between the average effective mAs in the reference condition i ($\bar{E}_i$), and the average effective mAs in the evaluation condition j ($\bar{E}_j$):

$$\Delta E_{ij} - \bar{E}_i - \bar{E}_j. \tag{4.1}$$

We standardized ΔE_{ij} to have a mean of zero and standard deviation of one, using

$$\widetilde{\Delta E_{ij}} = \frac{\Delta E_{ij} - \text{mean}_{ij}(\Delta E_{ij})}{\text{SD}_{ij}(\Delta E_{ij})}, \tag{4.2}$$

Table 4.3: Specificity for the reference and evaluation conditions, calculated from all three models.

	2D model	3D model-1	3D model-2
Reference conditions: Sensitivity (# of CT series that are correctly classified as non-IPF / total # of CT series)	0.99 (340/343)	0.97 (263/271)	0.97 (264/271)
Evaluation conditions: Sensitivity (# of CT series that are correctly classified as non-IPF / total # of CT series)	0.90 *** (681/759)	0.94 *** (498/530)	0.84 *** (445/530)

Note: One sample test of proportions were conducted for each model between the reference and evaluation conditions, *** means $P < 0.001$.

where mean(x) and SD(x) calculates the mean and standard deviation for a vector x, respectively.

For 2D models ($n = 759$ CT pairs), the estimated mean and standard deviation of ΔE_{ij} are 0.54 and 18.23, respectively. For 3D models ($n = 530$ CT pairs), the estimated mean and standard deviation of ΔE_{ij} are -2.33 and 14.57, respectively.

4.3 Results

4.3.1 Overall model performance

Since both reference conditions and evaluation conditions were collected from one non-IPF ILD cohort, which were considered as negatives in this example, we only calculated specificity as our statistical measure (without sensitivity). Specificity is defined as the number of CT series that are correctly predicted as non-IPF (true negatives) divided by the total number of non-IPF CT series (negatives). Table 4.3 provides the specificity calculated for both reference and evaluation conditions, for three models.

For all three models, specificity decreased to some extent when applying to the evaluation conditions: 2D model reduced from 0.99 to 0.90; 3D model-1 decreased from 0.97 to 0.94; 3D model-2 declined from 0.97 to 0.84. We conducted one sample test of proportions, for each model, between the reference and evaluation conditions. For all three models, proportion tests suggested that the probability of correctly classify one series was different among the

reference and evaluation conditions (all $P < 0.001$).

4.3.2 Factors influencing predictive results consistency

Table 4.4 provides the GLMM results, including the adjusted odds ratio (OR), 95% confidence intervals (CI), and P values for six types of technical and clinical parameters: reconstruction kernels, slice thickness, patient positions, manufacturer model name, clinical diagnosis, and effective mAs.

Table 4.4 shows that two types of variables are not significant ($P > 0.05$) in contributing to the inconsistent predictive results between the reference and evaluation conditions based on the GLMM analysis, for all three models: reconstruction kernels and patient positions. Specifically, for 3D model-1, there are no significant factors that lead to inconsistent model performance between reference and evaluation conditions (all factors $P > 0.10$). Among the other two models (2D model and 3D model-2), systemic sclerosis-associated ILD (SSc-ILD) group lowers the probability of getting conflicting results between the reference and evaluation conditions, as compared with the reference group (hypersensitivity pneumonitis, HP; P=0.03 for 2D model and P=0.05 for 3D model-2). Additionally, for 3D model-2, slice thickness (3mm) ($P = 0.04$), manufacturer model name of Siemens Definition ($P = 0.05$), Sjögren's syndrome-associated ILD (SjS-ILD, $P = 0.03$), and the standardized difference of mAs between reference and evaluation conditions ($P = 0.02$) are flagged as significant factors that are associated with model discrepancy.

4.4 Discussions and conclusions

DL approaches have long been criticized for their unexplainable and black box nature. It is often questionable what information do DL methods use when making predictions. Previous research demonstrated that DL algorithms may inexplicably leverage patient, scanner, or center information in the diagnosis process [BZO19]. If this is the case, then the developed DL model may work deceptively well in the data distribution that is similar to the training cases by levering irrelevant information, but fail to generalize to other scanners and hospital

Table 4.4: GLMM logistic analysis results.

	2D Model (n=759 CT pairs)		**3D Model-1** (n=530 CT pairs)		**3D Model-2** (n=530 CT pairs)	
	Adjusted OR (95% CI)	P value	Adjusted OR (95% CI)	P value	Adjusted OR (95% CI)	P value
Reconstruction kernels						
B50f (ref)	1.00		1.00		1.00	
B60f or B60s	0.81 (0.09, 7.46)	0.85	2.41 (0.06, 95.06)	0.64	0.42 (0.01, 13.67)	0.62
B70f	0.77 (0.04, 14.62)	0.86	2.28 (0.03, 200.42)	0.72	56(0.72, 4284)	0.07
Iterative	1.27 (0.26, 6.14)	0.77	NA	NA	NA	NA
Slice thickness (mm)						
1-2 (ref)	1.00		1.00		1.00	
3	8.05 (0.71, 91.12)	0.09	8.25 (0.12, 563.23)	0.33	116 (1.25, 10743)	0.04 *
5	5.19 (0.54, 49.56)	0.15	1.67 (0.03, 87.28)	0.80	27 (0.36, 2068)	0.14
Patient positions						
Prone (reference category)	1.00		1.00		1.00	
Supine	2.01 (0.30, 13.59)	0.47	0.12 (0.01, 1.93)	0.14	0.12 (0.01, 1.82)	0.13
Manufacturer model name						
Siemens Sensation 16 (ref)	1.00		1.00		1.00	
Siemens Definition	1.32 (0.38, 4.59)	0.67	1.40 (0.25, 8.00)	0.71	4.44 (1.02, 19.38)	0.05 *
Clinical diagnosis						
HP (ref)	1.00		1.00		1.00	
RAILD	0.93 (0.22, 3.82)	0.91	1.92 (0.28, 13.14)	0.51	0.35 (0.07, 1.64)	0.18
SjS-ILD	0.82 (0.15, 4.62)	0.82	3.00 (0.35, 25.88)	0.32	0.09 (0.01, 0.75)	0.03 *
SSc-ILD	0.13 (0.02, 0.81)	0.03 *	0.45 (0.05, 4.33)	0.49	0.18 (0.03, 1.03)	0.05 *
Effective mAs						
$\widetilde{\Delta E_{ij}}$	1.07 (0.60, 1.89)	0.82	1.05 (0.58, 1.89)	0.88	2.08 (1.11, 3.92)	0.02 *

Note: * means $P < 0.05$. Adjusted odds ratio (OR) and 95% confidence intervals (CIs) for the probability of receiving conflicting results between the reference and evaluation conditions. RAILD: rheumatoid arthritis-associated ILD; HP: hypersensitivity pneumonitis; SSc-ILD: Systemic sclerosis-associated ILD; SjS-ILD: Sjögren's syndrome-associated ILD. The results for iterative reconstruction kernels are listed as NA (Not available) for 3D models since there are no iterative reconstructions for 3D models among evaluation conditions. $\widetilde{\Delta E_{ij}}$: Normalized mean difference in effective mAs between reference and evaluation condition.

centers, etc.

In this article, we discussed a post-hoc evaluation method to assess the impact of a series of technical and clinical parameters on the diagnostic accuracy of several well-trained IPF diagnosis models. We used one 2D DL model and two 3D attention-based DL models, where two 3D models were constructed under with different hyperparameter selections. Attention-based models were utilized due to its ability to concentrate on the specific regions of interest provided by domain-specific knowledge and its capability to handle domain shifts [JLL18]. For all three models, the sensitivity decreased to some extent when applying to the evaluation sets: 2D model achieved a high specificity of 0.99 among reference conditions, but specificity decreased to 0.90 when testing on evaluation conditions (one sample proportion test, $P < 0.001$); 3D model-1 decreased from 0.97 to 0.94 (one sample proportion test, $P < 0.001$); and 3D model -2 dropped from 0.97 to 0.83 ($P < 0.001$).

To further analyze the variables that lead to the model specificity, Table 4.4 shows that there are no clinical and technical factors that are associated with the model diagnostic discrepancy in 3D model-1 when applying to the evaluation conditions ($P > 0.1$ for all factors), indicating that this model can stay relatively robust when testing on the evaluation conditions. The clinical diagnosis of SSc-ILD is a leading significant factor ($P = 0.03$ and $P = 0.05$) that is associated with reducing model discrepancy, for two out of three models. Among these three models, 3D model-2 is the one that flagged the most number of factors, including slice thickness, manufacturer model name, clinical diagnosis, and effective mAs, causing concerns when applying to different test sets.

Several limitations in this study merit considerations, including unmeasured confounding and the retrospective nature of this study. Firstly, there may exist unmeasured confounding in the statistical analysis due to the data anonymization process, such as patient weights. Unmeasured confounding are caused by the failure to include the factors that are associated with the outcome (in our case, model predictive discrepancy) and independent variables (in our case, imaging protocols-based covariates). For example, patient weight/size may be one important unmeasured confounding that is not included in the GLMM. Specifically, modern scanners use an automatic exposure control system called Tube Current Modulation (TCM)

that adapts the scanner output to patient size and the attenuation differences in different parts of the body [MBK06, LGY08]. As a result, larger patients require higher CT scanner output (higher effective tube current time product) that smaller patients using the same CT scanner settings; therefore, in clinical practice, the reported effective mAs is influenced by the patient's size. Due to the lack of patient weight information, our model does not contain patient size information and thus is unable to conclude that whether patient size has an association with the effective mAs level or the model predictive discrepancy. Secondly, due to the retrospective nature of this study, the main purpose of this research is to observe and analyze the possible factors that lead to the model performance decrease, while we cannot systematically design experiments to test out hypotheses. Future work includes conducting prospective studies to holistically evaluate model inconsistency under different variable settings.

Conclusions: Our preliminary findings showed that when applying three high-performing IPF diagnosis models to CT series collected under different imaging protocols, specificity decreased for all three models. We further demonstrated that clinical diagnosis is a key factor that leads to the lack of robustness for two out of three models. Our work indicated that care should be taken when training and deploying DL models into clinical practice.

CHAPTER 5

Discussions and conclusions

We discussed two challenges in this IPF diagnosis task, which are also frequently observed in other deep learning applications in the medical imaging tasks. One challenge is the **weakly supervised nature** of this task where the ground truth labels are coarse (such as scan-level), but not fine-scale levels as desired (such as voxel-level). Weakly supervised tasks are usually more efficient with respect to the data collection process, but pose significant challenges to constructing machine learning models since coarse-scale labels contain limited information. The other challenge is that deep learning models are often criticized for **lacking explainability** since deep learning methods utilize complicated functions which are hard to explain or understand.

Our proposals to address these challenges are to (1) leverage domain-specific knowledge from previous studies, which is more accessible than fine-scale annotations, to guide the learning of the model; (2) implement post-hoc explanation methods and attention mechanisms to enhance the extent of explainability. Specifically, in project I, domain knowledge is used to judiciously weigh different CT slices. Post-hoc visualizations are utilized to tackle the second challenge. With respect to project II, including domain knowledge provides extra guidance to the network and increases model accuracy and explainability.

We first discuss the design differences between project I and project II in Section 5.1. Cautionary notes of the dissertation are provided in Section 5.2.

5.1 Compare project I and project II

We summarize the major design differences of project I and project II in Table 5.1.

Convolutions: Projects I and II employed 2D convolutions and 3D convolutions, respectively. We provided a schematic that shows the differences between these two convolutions in Figure 1.1. Specifically, 2D convolutions capture information from the axial CT plane, whereas 3D convolutions exploit spatial information across the lungs.

Input dimensions: Since project I uses 2D convolutions, by default, it takes three CT slices (i.e. one CT triplet) with dimension $224 \times 224 \times 3$ as one training or testing input to feed into the state-of-art deep learning architectures, including ResNet-50, VGG16, DenseNet-121, and MobileNet. On the contrary, project II leverages 3D convolutions and takes one sampled CT volume of dimension $128 \times 128 \times 64$ as a training or testing input for MSGA.

DK: With respect to the use of domain knowledge, IPF progression and IPF quantification map were used in project I and project II, respectively. For project I, IPF progression information and an optimality criterion were used to weigh the CT triplets differently during the training of the model. For project II, IPF quantification map was incorporated in the construction of the model using attention mechanisms.

Explainability: In project I, we used Grad-CAM to visualize the important regions for models to make the decisions, which is a post-hoc explanation method. Based on the design of project I, Grad-CAM can provide only a two-dimensional image for each triplet. Take ResNet-50 as an example, Grad-CAM plots for one triplet are two 7×7 matrices, one for the IPF class and the other for the non-IPF class. The evaluation of Grad-CAM plots depends on the visual examinations of the selected cases. Notably, two-dimensional Grad-CAM plots are hard to trace back to the highlighted regions in the corresponding CT triplets.

For project II, attention models were employed to visualize the significant areas for the diagnosis task. Both a qualitative measure (visual examination of attention maps) and a quantitative measure (kurtosis) were used to measure the explainability. Attention mechanisms are included in the overall loss function and impact the training of the model in an end-to-end manner. These produced attention maps are three-dimensional and the attention gated output is four-dimensional. Compared with that of project I, these three-dimensional

Table 5.1: Major design differences between project I and project II. DK: domain knowledge.

	Project I	**Project II**
Convolutions	2D-CNN	3D-CNN
DK	IPF progression	IPF quantification map
How DK is incorporated	End-to-end, weigh CT triplets	End-to-end, Attention mechanisms
Explainability measures	Qualitative, Grad-CAM	Both qualitative (attention maps) and quantitative (kurtosis)
	Post-hoc evaluations	Impacts the training
	Grad-CAM: Two-dimensional	Attention maps: Three-dimensional; Attention-gated output: four-dimensional

attention maps are more informative, contrary to the two-dimension Grad-CAM plots.

5.2 Cautionary notes

We present several cautionary notes in this dissertation: (1) in clinical practice, IPF diagnosis involves the process of excluding subjects with known causes of ILD, such as occupational environmental exposures, as well as the evaluation of HRCT patterns of UIP [RRM18]. Both project I and project II do not include the clinical history evaluation and UIP pattern examination, which is our innovative attempt to build an efficient diagnosis paradigm, but more clinical validations are needed. (2) both models developed in project I and project II have not yet been validated using prospective clinical trials. We understand that the process of developing a successful and clinically applicable automated diagnosis tool may take years to accomplish. We hope that the methodology discussed in this dissertation can shed light on the process of future model development and improvement.

APPENDIX A

Supplementary files for Introduction

An example of image processing is provided in Figure A.1. More details of each step was discussed in Section 1.3.3. During the image processing steps, the dimension of this CT scan reduced from $512 \times 512 \times 62$ to $128 \times 128 \times 64$.

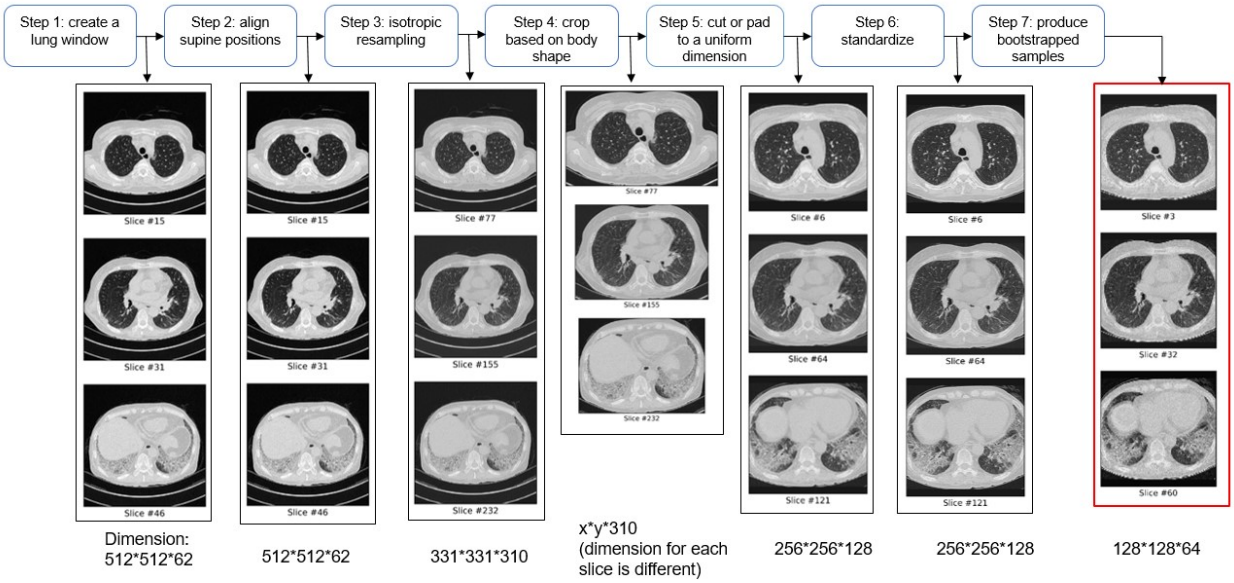


Figure A.1: An example of image processing. One of the final bootstrapped samples is highlighted with red rectangles. The dimension of each intermediate image is displayed under the figure.

APPENDIX B

Supplementary files for Project I

B.1 CT acquisition and image reconstruction conditions of the five studies.

Table B.1: CT acquisition and image reconstruction conditions of the five studies.

Study	Slice thickness	Patient positions	Manufacturers	Reconstruction kernels	Pixel spacing (mm)	Image resolution	Percentage of volumetric scans (%)
1	1, 1.25mm: 98.4% Other: 1.6%	HFP: 73.1% FFP: 20.0% HFS: 5.7% FFS: 1.2%	SEIMENS: 39.6% GE: 39.2% Philips: 12.7% TOSHIBA: 8.2%	B45f, BONE, D, STANDARD, etc.	0.72 ± 0.08	(512,512): 97.7%	97.5
2	1mm: 100%	FFS: 99.3% HFS: 0.7%	SIEMENS: 100%	B45f, Br49d, B31f, etc.	0.64 ± 0.06	(512,512): 100%	100
3	1, 1.25mm: 96.6% Other: 3.4%	FFS: 71.0% FFP: 26.1% HFS: 0.9%	SEIMENS: 93.3% GE: 6.2% Other: 0.5%	B60f, B70f, B60s, BONE, etc.	0.63 ± 0.05	(512,512): 100%	87.7
4	1mm: 97.5% Other: 2.5%	FFS: 91.4% HFS: 8.6%	Philips: 65.4% SIEMENS: 27.2% TOSHIBA: 3.7% GE: 3.7%	D, B60f, YC, YA, FC55, L, etc.	0.60 ± 0.05	(512,512): 100%	97.5
5	2.5mm: 40.6% 1, 1.25mm: 38.2% 2mm: 17.1% Other: 4.1%	FFS: 91.2% HFS: 7.1% FFP: 1.8%	GE: 71.8% TOSHIBA: 13.5% SIEMENS: 9.4% Philips: 5.3%	LUNG, STANDARD, FC56, L, etc.	0.62 ± 0.07	(512,512): 96.4%	97.6

Note: This information was retrieved from DICOM (digital imaging and communications in medicine) image header files. HFP: head first prone, FFP: feet first prone, HFS: head first supine, FFS: feet first supine. Reconstruction kernels are sorted by frequency of elements in descending order.

B.2 Model construction for the pilot study

Previous study contains the axial chest CT scans of 122 clinically-diagnosed IPF patients. The number of available CT slices per scan is 300 ± 92. The total number of available chest slices is 36,603. After an in-house automatic lung segmentation and denoising procedure, we record the number of segmented lung area pixels on a slice level. For all segmented lung area pixels, we predict the likelihood of progression using a machine learning technique. 1 Since the response variable (the percentage of progressive lung pixels) is not normally distributed, we consider generalized linear models (GLM). For each lung CT slice index t, we have 4 key variables: the number of predicted progressive lung area pixels (r_t), the number of lung area pixels (o_t), the true progressive rate (p_t) and the log-odds of the true progressive rate (θ_t) and the model is given by

$$r_t \sim Binomial(o_t, p_t), \tag{B.1}$$

$$\theta_t = \log(\frac{p_t}{1-p_t}), \tag{B.2}$$

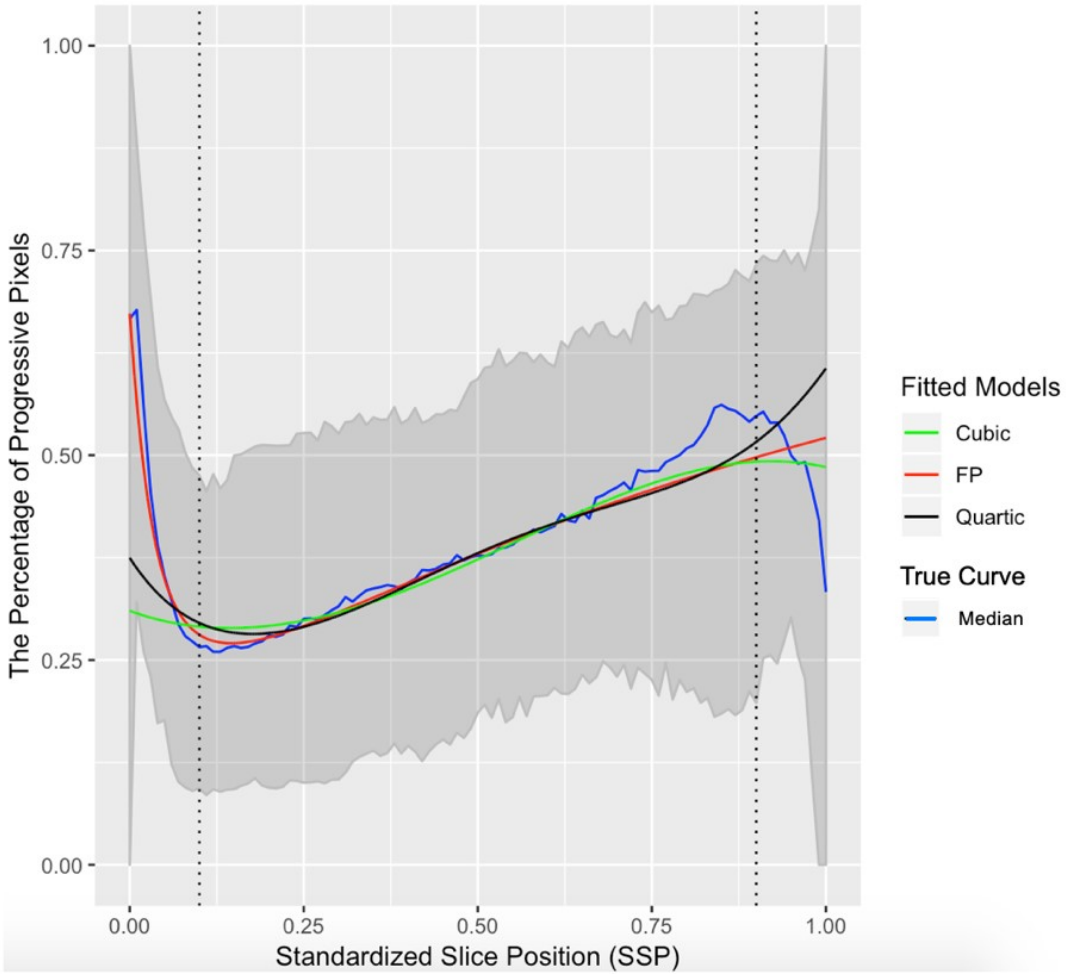
where t=1,..., 36,603 in our study. We assume that the log-odds of the true progressive rate can be adequately modeled by a logistic model with a linear predictor given by either a cubic or quartic polynomial or a fractional polynomial in the explanatory variable SSP_t representing the standardized slice positions (SSPs). Fitting the model with polynomial terms is straightforward, and for fitting the fractional polynomial models, we used the mfp package in R version 3.4.0. Supplementary Table B.2 shows the expression and estimated parameters of best fitted cubic, quartic polynomials, and fractional polynomial (FP) models. FP model is selected as the best-fitting model since it achieved the smallest AIC score.

Furthermore, Davidson-MacKinnon J tests [DM81] were used to assess the validity of model fitting in the presence of one alternative model. The principle of J test is straightforward: if one model includes the correct set of variables, then adding new regressors from another model should not be statistically significant. In our case, for example, sup-

pose the cubic model can correctly characterize the underlying distribution with its regressors (i.e. SSP_t, SSP_t^2, SSP_t^3), then adding new terms from the FP model (for example, $\log(SSP_t + 0.1)$) should not contribute to the model performance. The derivation of J test is based on asymptotic theory and the results are reported to be misleading when the sample size is small [DM04].

We used J tests to compare each of the two models: FP and cubic; FP and quartic; cubic and quartic. Results from J tests suggest that each of the two models are statistically different ($P < 0.001$ for all comparisons), except that for the FP model, adding the regressors from the quartic model do not change the model performance ($P = 0.26$). We used the function jtest from the R package lmtest to conduct all J tests [HZF15].

Figure B.1 provides a visual description of the fits for the three models and the true median curve displays the percentages of progressive pixels versus the SSPs. From the figure, we conclude that the selected fractional polynomial provides the best fit.



Notes: The gray area represents the range of 2.5^{th} percentile and 97.5^{th} percentile of the true curve and the two dotted vertical lines at SSP = 0.10 and 0.90 represent the noticeable boundary effects.

Figure B.1: The true median curve in blue shows the percentage of progressive pixels versus standardized slice position (SSP). The other colored curves are the best fits to the overall population trends from the other three models.

Table B.2: Model fitting performance: three pre-selected models and their corresponding estimated parameters and Akaike information criterion (AIC). FP: fractional polynomial. FP achieves the least AIC score and is highlighted in bold fonts.

Model	Expression	Estimated parameters	AIC
FP	$\theta_t = \beta_0 + \beta_1(SSP_t + 0.1)^{-2} + \beta_2 \log(SSP_t + 0.1) + \epsilon_t$	$\hat{\beta}_0 = -0.04$ $\hat{\beta}_1 = 0.03$ $\hat{\beta}_2 = 1.05$	**5372743**
Cubic	$\theta_t = \beta_0 + \beta_1 SSP_t + \beta_2 SSP_t^2 + \beta_3 SSP_t^3 + \epsilon_t$	$\hat{\beta}_0 = -0.80$ $\hat{\beta}_1 = -1.47$ $\hat{\beta}_2 = 5.91$ $\hat{\beta}_3 = -3.70$	5435168
Quartic	$\theta_t = \beta_0 + \beta_1 SSP_t + \beta_2 SSP_t^2 + \beta_3 SSP_t^3 + \beta_4 SSP_t^4 + \epsilon_t$	$\hat{\beta}_0 = -0.51$ $\hat{\beta}_1 = -5.58$ $\hat{\beta}_2 = 22.98$ $\hat{\beta}_3 = -30.38$ $\hat{\beta}_4 = 13.93$	5413504

B.3 D-optimal design under generalized linear models (GLM) setting

Since 2D-CNN architectures are used for this task, we explore a three-point design with equal weights. Each such design Z_{ij} is equally weighted at the three sampled lung CT ordered standardized slice position z_{ijk} (in ascending order), $k = 1, 2, 3$, i is the subject index and j is the triplet index. For simplicity, we omit the subscript i and j in this section. The triplet contains the three slices sampled from the top, middle, and bottom of the lungs respectively; this means that their positions satisfy the constraints: $0.1 \leqslant z_{ij1} < 0.37$, $0.37 \leqslant z_{ij2} < 0.64$,and $0.64 \leqslant z_{ij3} < 0.9$. Let y be the 3×1 vector of observed percentages of progressive lung area pixels at these three lung CT positions and, let μ be its mean vector of size 3×1 with each component between 0 and 1. The statistical model in matrix form

can be written in two parts:

$$E(y) = \mu \text{ and } \log(\frac{\mu}{1-\mu}) = F\beta, \tag{B.3}$$

with the interpretation that the equations are interpreted row-wise. The matrix F is the 3×3 loading matrix for the standardized slice positions and β is the 3×1 vector of unknown parameters. As an illustration, consider the case when we have the selected fractional polynomial with

$$F = \begin{pmatrix} f(z_1)^T \\ f(z_2)^T \\ f(z_3)^T \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2, \end{pmatrix}$$

and $f(z_j)^T = \Big(1, \quad (z_j + 0.1)^{-2}, \quad \log(z_j + 0.1)\Big)$ are functions of the standardized slice positions z_j, $j = 1, 2, 3$. The D-optimality criterion focuses on the determinant of the information matrix, which is defined by the negative of the expectation of the second derivatives of the total log likelihood function with respect to the model parameters. More specifically, if Z is a n-point design, the D-criterion is defined by $D(Z) = |F^T W_l F|$, where the weight matrix W_l is a diagonal matrix of size $n \times n$, with diagonal elements $w_j = \hat{\mu}_j(1 - \hat{\mu}_j)$, where $\hat{\mu}_j = \frac{exp(f(z_j)^T \hat{\beta})}{1+exp(f(z_j)^T \hat{\beta})}$. D-optimal design maximizes $D(Z)$ over all possible designs and a minimally-supported D-optimal design maximizes $D(Z)$ over all 3-point designs. Supporting Table B.2 shows the estimated parameters $\hat{\beta}$ for the three selected models.

For example, for the one specific triplet (triplet 1) provided in Figure 2.4 (b), based on the standardized slice positions, we have $\hat{z}_1 = 0.34$, $\hat{z}_2 = 0.63$, $\hat{z}_3 = 0.67$, and $D(\hat{Z}) = |\hat{F}^T \hat{W}_l \hat{F}| = 1.19 \times 10^{-4}$, where the calculation of $\hat{F}$ and $\hat{W}_l$ for this specific triplet is provided:

$$\hat{F} = \begin{pmatrix} 1 & (\hat{z}_1 + 0.1)^{-2} & \log(\hat{z}_1 + 0.1) \\ 1 & (\hat{z}_2 + 0.1)^{-2} & \log(\hat{z}_2 + 0.1) \\ 1 & (\hat{z}_3 + 0.1)^{-2} & \log(\hat{z}_3 + 0.1) \end{pmatrix} = \begin{pmatrix} 1 & 5.17 & -0.82 \\ 1 & 1.88 & -0.31 \\ 1 & 1.69 & -0.26 \end{pmatrix}$$

and $\hat{W}_l = \begin{pmatrix} 0.22 & 00 & \\ 0 & 0.24 & 0 \\ 0 & 0 & 0.25 \end{pmatrix}$. Here $|\hat{F}^T \hat{W}_l \hat{F}|$ denotes the determinant of the matrix $\hat{F}^T \hat{W}_l \hat{F}$.

B.4 Visualization of D-criterion values

To better visualize the relationship between lung CT positions and their corresponding D-criterion values, we plot the distribution of D-criterion values while fixing one CT position (z_1, z_2, z_3) at a time. As discussed before, each triplet contains three slices which are sampled one from each zone (see Figure 2 (a)). Fixing one slice at the midpoint of that zone, we then evenly sample 20 slices from the other two zones, and calculate the D-criterion values for a total of 400 (=20 $\times$ 20) triplet combinations for each subplot, as shown in Figure B.2.

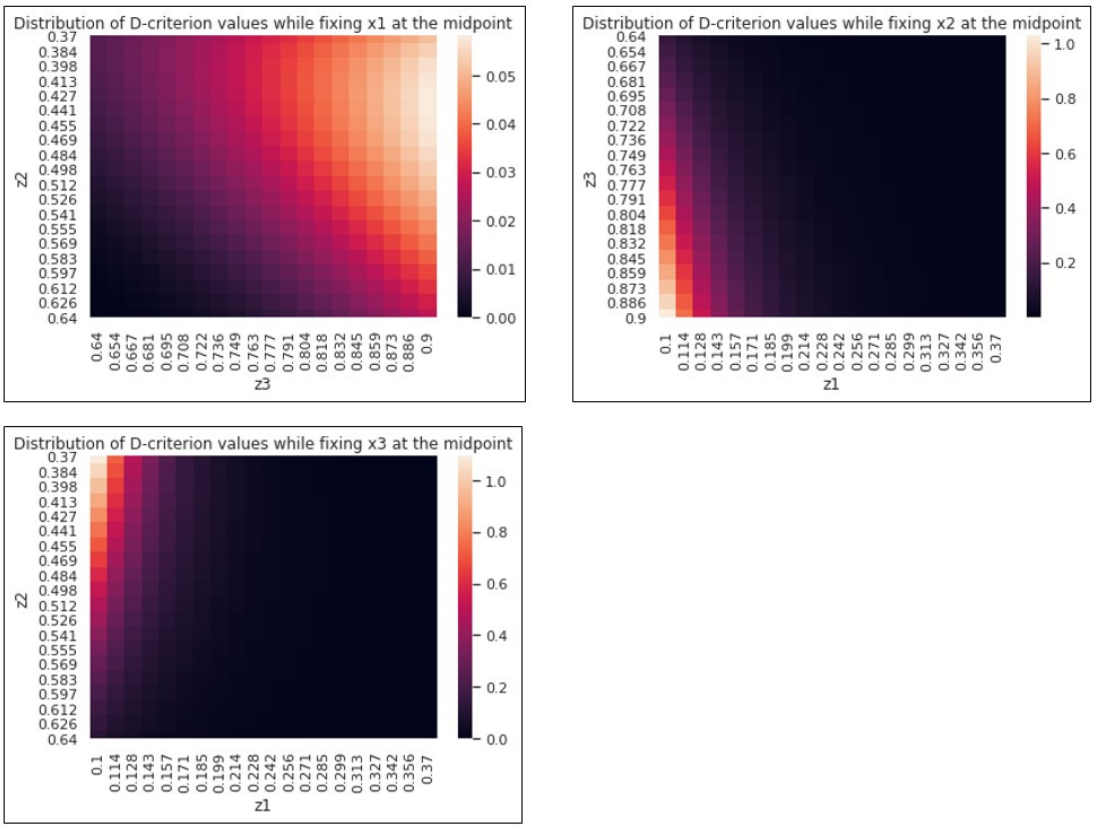


Figure B.2: Distributions of D-criterion values while fixing z_1, z_2, and z_3 one at a time, respectively.

Figure B.2 shows a general trend that when we fix one slice, the D-criterion values of neighboring slices are smaller. This is consistent with the common knowledge that when two CT slices are closer, they offer less meaningful medical information than two distant slices. Therefore, the triplet with two slices closer to each other should be assigned with less weight in the training of deep learning systems. This also justifies our assumption that using triplets collected from three separate zones can, to some extent, avoid sampling slices that are adjacent to each other. This suggests that optimal design ideas can provide a better sampling strategy.

B.5 Sensitivity analysis results

Sensitivity analysis results of scenario 1, 2, 3 are presented in the Table B.3, B.4, and B.5, respectively.

B.6 Model generalizability testing

Generalizability is defined as the model performance on unseen test scenarios. For example, deploying a developed model on CT images acquired under a new institution, protocols, patient cohort, etc. The evaluation of generalizability is a key factor for the successful deployment of deep learning models in healthcare settings.

In order to evaluate our model generalizability to an unseen scenario, we add the following experiments by selecting one study at a time as the holdout set. These tests are constructed to evaluate if the constructed model is able to achieve good inference for CT images acquired from a new institution, i.e. the holdout set. Experimental design and model accuracy in the holdout study are provided in Supplementary Table B.6. For example, under experiment number 1, we use all IPF patients and non-IPF patients from study 4 and study 5 to construct a new IPF diagnosis model. Afterwards, we evaluate the model performance under another holdout study (i.e. study 3 in this case), which has not been included in the model construction stage. Due to the sample size limitation of IPF patients, we did not conduct the experiment using study 1 as the holdout study.

According to the results in Supplementary Table B.6, except for the experiment number 3, other experiments can successfully classify more than 90% of patients in the holdout study (accuracies greater than 90%). Experiment number 3, which uses study 5 as the holdout study, only achieves an average accuracy of 0.73. This may due to the fact that study 5 has 61.8% of CT scans that are greater than 1.25mm, while CT scans collected from all four other studies only have less than 5%. Under this scenario, the deep learning model constructed using thin (1-1.25mm) CT scans does not generalize well to thicker (greater than 1.25) CT scans in the holdout study. Other studies have also reported this lack of generalizability in

Table B.3: Study-wise model performance and overall model performance with *an adaptive selection of triplets per scan.*

	Sensitivity (IPF patients)		Specificity (Non-IPF ILD patients)			Overall model performance		
Model (Loss function)	Study 1	Study 2	Study 3	Study 4	Study 5	Sensitivity	Specificity	Accuracy
Baseline CNN (CE)	0.94 (0.03)	0.86 (0.08)	0.93 (0.06)	0.95 (0.03)	0.88 (0.05)	**0.91 (0.03)**	**0.92 (0.04)**	**0.92 (0.02)**
Baseline CNN (DK)	0.86 (0.16)	0.81 (0.21)	0.84 (0.16)	0.84 (0.14)	0.78 (0.17)	0.84 (0.17)	0.82 (0.15)	0.83 (0.06)
MobileNet (CE)	0.98 (0.01)	0.97 (0.04)	0.99 (0.01)	0.99 (0.03)	0.96 (0.03)	0.98 (0.01)	0.98 (0.01)	0.98 (0.01)
MobileNet (DK)	0.99 (0.02)	0.92 (0.06)	0.98 (0.02)	1 (0)	0.95 (0.04)	0.97 (0.02)	0.98 (0.02)	0.97 (0.01)
VGG16 (CE)	0.98 (0.02)	0.97 (0.03)	0.99 (0.02)	1 (0)	0.98 (0.02)	0.98 (0.01)	0.99 (0.01)	0.98 (0.01)
VGG16 (DK)	0.98 (0.02)	0.96 (0.03)	0.99 (0.01)	1 (0)	0.98 (0.02)	0.97 (0.01)	0.99 (0.01)	0.98 (0.01)
ResNet-50 (CE)	0.98 (0.01)	0.89 (0.08)	0.93 (0.13)	0.87 (0.25)	0.91 (0.11)	0.95 (0.03)	0.91 (0.14)	0.93 (0.08)
ResNet-50 (DK)	0.97 (0.02)	0.86 (0.06)	0.97 (0.02)	0.99 (0.02)	0.96 (0.02)	0.94 (0.02)	**0.97 (0.02)**	**0.96 (0.01)**
DenseNet-121 (CE)	0.96 (0.04)	0.84 (0.22)	0.99 (0.01)	1 (0)	0.97 (0.03)	**0.92 (0.08)**	0.99 (0.01)	**0.96 (0.03)**
DenseNet-121 (DK)	0.88 (0.20)	0.77 (0.31)	0.98 (0.02)	0.99 (0.03)	0.98 (0.02)	0.84 (0.24)	0.98 (0.02)	0.93 (0.09)

Note: Mean and standard deviations shown in brackets are calculated across the results from each testing fold. CE: cross entropy loss without domain knowledge-enhanced loss function; DK: domain knowledge-enhanced loss function. Statistically significant results ($P < 0.017$) are highlighted in bold font. The significance cutoff 0.017 is decided by Bonferroni correction for multiple testing, which is dividing the pre-specified significance level 0.05 by the number of tests (3, including the overall sensitivity, specificity, and accuracy) for each model.

Table B.4: Study-wise model performance and overall model performance by *adding a re-sampling step during the preprocessing procedure.*

	Sensitivity (IPF patients)		Specificity (Non-IPF ILD patients)			Overall model performance		
Model (Loss function)	Study 1	Study 2	Study 3	Study 4	Study 5	Sensitivity	Specificity	Accuracy
Baseline CNN (CE)	0.89 (0.05)	0.81 (0.08)	0.96 (0.03)	0.95 (0.04)	0.92 (0.04)	0.86 (0.05)	0.95 (0.02)	0.91 (0.01)
Baseline CNN (DK)	0.88 (0.03)	0.82 (0.11)	0.96 (0.03)	0.96 (0.03)	0.93 (0.03)	0.86 (0.04)	0.95 (0.01)	0.92 (0.01)
MobileNet (CE)	0.98 (0.01)	0.96 (0.03)	1 (0)	1 (0)	0.97 (0.03)	0.98 (0.01)	0.99 (0.01)	0.98 (0)
MobileNet (DK)	0.99 (0.01)	0.98 (0.03)	0.97 (0.04)	1 (0)	0.97 (0.03)	0.99 (0)	0.98 (0.03)	0.98 (0.01)
VGG16 (CE)	0.98 (0.02)	0.95 (0.04)	1 (0)	1 (0)	0.98 (0.02)	0.97 (0.01)	0.99 (0.01)	0.98 (0)
VGG16 (DK)	0.98 (0.01)	0.96 (0.03)	1 (0)	1 (0)	0.98 (0.02)	0.98 (0.01)	0.99 (0.01)	0.99 (0)
ResNet-50 (CE)	0.98 (0.01)	0.90 (0.07)	0.99 (0.01)	0.97 (0.03)	0.97 (0.02)	0.95 (0.02)	0.98 (0.01)	0.97 (0.01)
ResNet-50 (DK)	0.97 (0.01)	0.88 (0.11)	0.98 (0.01)	0.94 (0.09)	0.95 (0.01)	0.95 (0.03)	0.97 (0.01)	0.96 (0.02)
DenseNet-121 (CE)	0.98 (0.01)	0.91 (0.06)	0.98 (0.03)	0.99 (0.03)	0.97 (0.02)	0.96 (0.01)	0.98 (0.02)	0.97 (0.01)
DenseNet-121 (DK)	0.97 (0.03)	0.88 (0.11)	0.99 (0.01)	1 (0)	0.97 (0.02)	0.94 (0.03)	0.98 (0.01)	0.97 (0.01)

Note: Mean and standard deviations shown in brackets are calculated across the results from each testing fold. CE: cross entropy loss without domain knowledge-enhanced loss function; DK: domain knowledge-enhanced loss function. No statistically significant results ($P < 0.017$) were identified in this table.

Table B.5: Study-wise model performance and overall model performance using triplets collected from *lower zones only*.

	Sensitivity (IPF patients)		Specificity (Non-IPF ILD patients)			Overall model performance		
Model (Loss function)	Study 1	Study 2	Study 3	Study 4	Study 5	Sensitivity	Specificity	Accuracy
Baseline CNN (CE)	0.9 (0.05)	0.86 (0.06)	0.96 (0.05)	0.97 (0.06)	0.96 (0.02)	0.89 (0.04)	0.96 (0.04)	0.93 (0.02)
Baseline CNN (DK)	0.89 (0.06)	0.86 (0.07)	0.96 (0.02)	0.87 (0.04)	0.93 (0.06)	0.88 (0.06)	0.94 (0.03)	0.91 (0.01)
MobileNet (CE)	0.99 (0.01)	0.94 (0.04)	0.99 (0.01)	1 (0)	0.98 (0.02)	0.97 (0.02)	0.99 (0.01)	0.98 (0.01)
MobileNet (DK)	0.98 (0.01)	0.98 (0.02)	0.99 (0.01)	1 (0)	0.97 (0.03)	0.98 (0.01)	0.98 (0.01)	0.98 (0.01)
VGG16 (CE)	0.98 (0.01)	0.93 (0.05)	0.99 (0.01)	1 (0)	0.98 (0.02)	0.96 (0.02)	0.99 (0.01)	0.98 (0.01)
VGG16 (DK)	0.97 (0.01)	0.95 (0.03)	0.99 (0.01)	1 (0)	0.99 (0.02)	0.97 (0.02)	0.99 (0.01)	0.98 (0.01)
ResNet-50 (CE)	0.96 (0.03)	0.89 (0.01)	0.96 (0.02)	0.96 (0.03)	0.96 (0.03)	0.94 (0.03)	0.96 (0.02)	0.95 (0.02)
ResNet-50 (DK)	0.97 (0.03)	0.81 (0.16)	0.98 (0.01)	0.94 (0.04)	0.95 (0.04)	0.92 (0.06)	0.96 (0.02)	0.95 (0.02)
DenseNet-121 (CE)	0.97 (0.02)	0.83 (0.13)	0.99 (0.01)	1 (0)	0.97 (0.03)	0.93 (0.05)	**0.99 (0.01)**	**0.96 (0.02)**
DenseNet-121 (DK)	0.93 (0.11)	0.77 (0.19)	0.98 (0.01)	0.95 (0.04)	0.97 (0.02)	0.88 (0.14)	0.97 (0.02)	0.93 (0.05)

Note: Mean and standard deviations shown in brackets are calculated across the results from each testing fold. CE: cross entropy loss without domain knowledge-enhanced loss function; DK: domain knowledge-enhanced loss function. Statistically significant results ($P < 0.017$) are highlighted in bold font.

Table B.6: Experimental setup and results for model generalizability testing by using one study at a time as the holdout test study.

Experiment number	IPF cohorts	Non-IPF cohorts	Holdout study	Model (Loss function)	Model accuracy in the holdout study (standard deviation)
1	1 & 2	4 & 5	3	MobileNet (CE)	0.93 (0.03)
1	1 & 2	4 & 5	3	MobileNet (DK)	0.93 (0.03)
2	1 & 2	3 & 5	4	MobileNet (CE)	0.97 (0.03)
2	1 & 2	3 & 5	4	MobileNet (DK)	0.95 (0.07)
3	1 & 2	3 & 4	5	MobileNet (CE)	0.73 (0.17)
3	1 & 2	3 & 4	5	MobileNet (DK)	0.73 (0.16)
4	1	3, 4 & 5	2	MobileNet (CE)	0.99 (0.01)
4	1	3, 4 & 5	2	MobileNet (DK)	0.99 (0.01)

Note: Mean and standard deviations shown in brackets are calculated across the results from each fold, based on five-fold cross validation. CE: cross entropy loss without domain knowledge-enhanced loss function; DK: domain knowledge-enhanced loss function.

CT texture features caused by variations in CT slice thicknesses [CKL15].

APPENDIX C

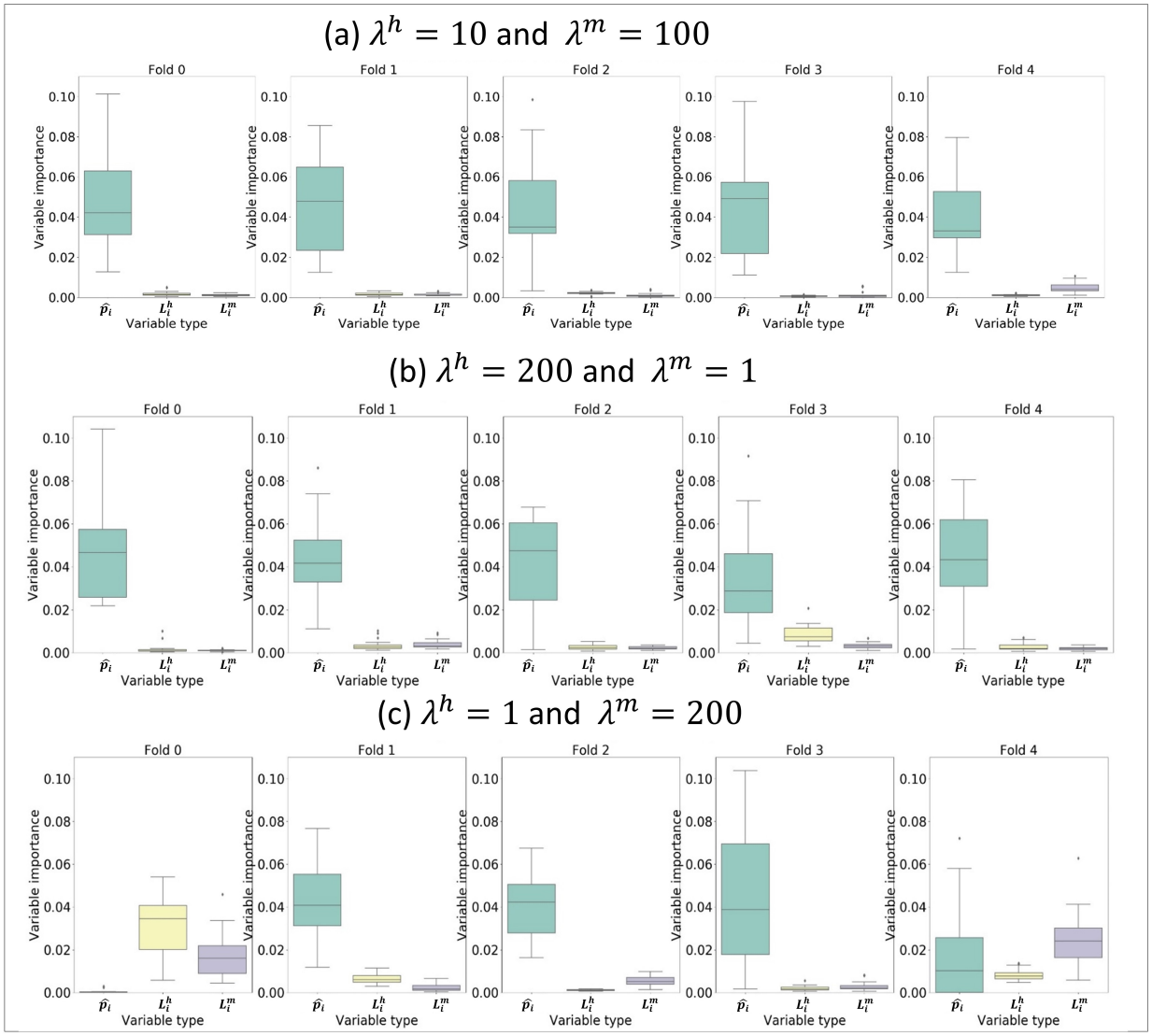
Supplementary files for project II

C.1 Random forest (RF) analysis

We calculated and plotted the variable importance for the constructed RF using the normalized total reduction of Gini impurity brought by each feature.

We provide a series of boxplots representing the variable importance analysis for the constructed RF, as shown in Figure C.1. For each RF, the variables come from the concatenated MSGA output collected from all M number of CT samples, where $M = 20$ in our work. Furthermore, the learned features from MSGA (a vector of size $1 \times 3M$) can be separated into three types: $\hat{p}_i$ (size $1 \times M$), L_i^h (size $1 \times M$), and L_i^m (size $1 \times M$), where $\hat{p}_i$, L_i^h, L_i^m represent the predicted probability of being IPF at the last layer of MSGA, attention-based loss function at a high- and medium- resolution for subject i, respectively. Each boxplot shows the variability across the variable importance for these M variables.

Validation set performance (AUC for each fold) under MSGA and MSGA+RF is reported in Supplementary Table C.1.



Notes: M is the number of samples produced for each CT scan, where we select $M = 20$ in this work. We separate the learned features from MSGA (a vector of size $1 \times 3M$) into three types: $\hat{p}_i$ (size $1 \times M$), L_i^h (size $1 \times M$), and L_i^m (size $1 \times M$), where $\hat{p}_i$, L_i^h, L_i^m represent the predicted probability of being IPF at the last layer of MSGA, attention-based loss function at a high- and medium- resolution, respectively. Each boxplot shows the variability across the variable importance for these M variables.

Figure C.1: Variable importance plots under the RF model using three hyperparameter settings as illustrative examples (a, $\lambda^h = 10$ and $\lambda^m = 100$; b, $\lambda^h = 200$ and $\lambda^m = 1$; c, $\lambda^h = 1$ and $\lambda^m = 200$). Variable importance is plotted for each fold.

The histogram of the estimated attention-based loss functions among all training samples using one fold model as an example are plotted in Figure C.1. The figure shows a clear

Table C.1: Validation set performance (AUC for each fold) of both MSGA and MSGA+RF under three hyperparameter collections, including a, $\lambda^h = 10$ and $\lambda^m = 100$; b, $\lambda^h = 200$ and $\lambda^m = 1$; c, $\lambda^h = 1$ and $\lambda^m = 200$.

			Fold 0		Fold 1		Fold 2		Fold 3		Fold 4	
	λ^h	λ^m	AUC of MSGA	AUC of MSGA +RF	AUC of MSGA	AUC of MSGA +RF	AUC of MSGA	AUC of MSGA +RF	AUC of MSGA	AUC of MSGA +RF	AUC of MSGA	AUC of MSGA +RF
(a)	10	100	0.983	0.995	0.986	0.987	0.999	0.986	0.986	0.989	0.980	0.984
(b)	200	1	0.999	0.984	0.967	0.985	0.987	0.974	0.944	0.967	0.993	0.971
(c)	1	200	0.513	0.943	0.949	0.977	0.985	0.982	0.831	0.920	0.500	0.804

distinction between IPF and non-IPF subjects regarding the estimated attention-based loss functions for both high- (a) and medium (b) resolutions. This provides an explanation why, for this fold, adding an RF classifier can greatly improve the model performance for the validation set (AUC from 0.513 to 0.943).

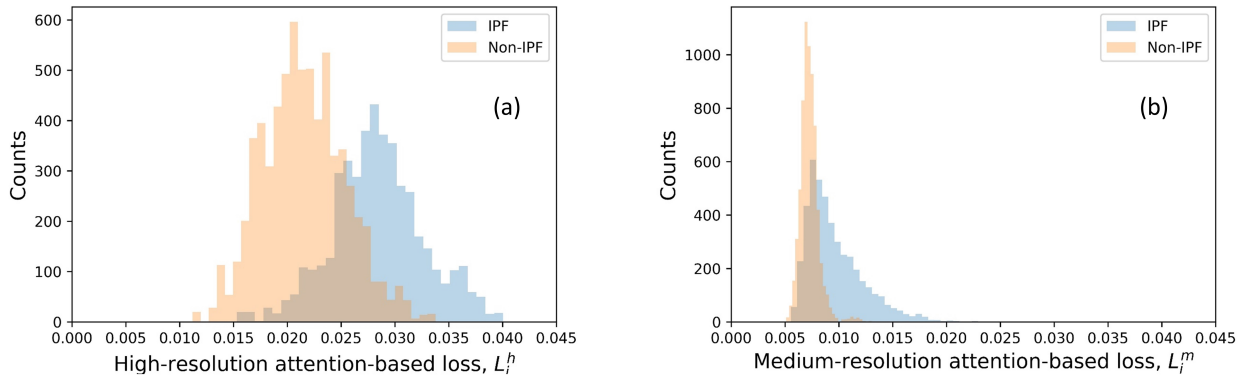


Figure C.2: Histogram of the estimated attention-based loss function at high- (a, L_i^h) and medium- resolution (b, L_i^m) when $\lambda^h = 1$ and $\lambda^m = 200$, at fold 0, among all training samples.

Visual examination of the variable importance illustrated that when MSGA performance was satisfactory (we provided two examples here: $\lambda^h = 10$ and $\lambda^m = 100$; $\lambda^h = 200$ and $\lambda^m = 1$), RF mostly leveraged information from the predicted probability of being IPF produced at the last layer of MSGA ($\hat{p}_i$), with minimal information borrowed from the estimated attention-based loss (including L_i^h and L_i^m).When MSGA performance was not

satisfactory (for example, $\lambda^h = 1$ and $\lambda^m = 200$), RF took the estimated attention loss into consideration with large variable importance, especially for fold 0 and fold 4.

[ACE16] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network." *IEEE transactions on medical imaging*, **35**(5):1207–1216, 2016.

[AKB19] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, and others. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." *Nature medicine*, **25**(6):954–961, 2019. Publisher: Nature Publishing Group.

[AKK20] S Agarwala, M Kale, D Kumar, R Swaroop, A Kumar, A Kumar Dhara, S Basu Thakur, A Sadhu, and D Nandi. "Deep learning for screening of interstitial lung disease patterns in high-resolution CT images." *Clinical radiology*, **75**(6):481–e1, 2020.

[ANT05] Arata Azuma, Toshihiro Nukiwa, Eiyasu Tsuboi, Moritaka Suga, Shosaku Abe, Koichiro Nakata, Yoshio Taguchi, Sonoko Nagai, Harumi Itoh, Motoharu Ohi, et al. "Double-blind, placebo-controlled trial of pirfenidone in patients with idiopathic pulmonary fibrosis." *American journal of respiratory and critical care medicine*, **171**(9):1040–1047, 2005.

[AR01] Gareth Ambler and Patrick Royston. "Fractional polynomial model selection procedures: investigation of Type I error rate." *Journal of Statistical Computation and Simulation*, **69**(1):89–108, 2001.

[BLB03] Alan C Best, Anne M Lynch, Carmen M Bozic, David Miller, Gary K Grunwald, and David A Lynch. "Quantitative CT indexes in idiopathic pulmonary fibrosis: relationship with physiologic impairment." *Radiology*, **228**(2):407–414, 2003.

[BMB18] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, Gabor Grothendieck, Peter Green, and others. "Package 'lme4'." *Version*, **1**:17, 2018.

[Bre01] Leo Breiman. "Random forests." *Machine learning*, **45**(1):5–32, 2001.

[BW05] Martijn PF Berger and Weng-Kee Wong. *Applied optimal designs*. John Wiley & Sons, 2005.

[BW09] Martijn PF Berger and Weng-Kee Wong. *An introduction to optimal designs for social and biomedical research*, volume 83. John Wiley & Sons, 2009.

[BZO19] Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M

Snyder, and Joel T Dudley. "Deep learning predicts hip fracture using confounding patient and healthcare variables." *NPJ digital medicine*, **2**(1):1–10, 2019. Publisher: Nature Publishing Group.

[CKL15] Daniel Y Chong, Hyun J Kim, Pechin Lo, Stefano Young, Michael F McNitt-Gray, Fereidoun Abtin, Jonathan G Goldin, and Matthew S Brown. "Robustness-driven feature selection in classification of fibrotic interstitial lung disease patterns in computed tomography using 3D texture features." *IEEE transactions on medical imaging*, **35**(1):144–157, 2015.

[Cle01] William S Cleveland. "Data science: an action plan for expanding the technical areas of the field of statistics." *International statistical review*, **69**(1):21–26, 2001.

[CLX18] Yidong Chai, Hongyan Liu, and Jie Xu. "Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models." *Knowledge-Based Systems*, **161**:147–156, 2018.

[CMP19] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. "An attentive survey of attention models." *arXiv preprint arXiv:1904.02874*, 2019.

[CPD19] Andreas Christe, Alan A Peters, Dionysios Drakopoulos, Johannes T Heverhagen, Thomas Geiser, Thomai Stathopoulou, Stergios Christodoulidis, Marios Anthimopoulos, Stavroula G Mougiakakou, and Lukas Ebner. "Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images." *Investigative radiology*, **54**(10):627, 2019.

[DCK19] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. "Domain generalization via model-agnostic learning of semantic features." *arXiv preprint arXiv:1910.13580*, 2019.

[DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[Dha13] Vasant Dhar. "Data science and prediction." *Communications of the ACM*, **56**(12):64–73, 2013.

[DM81] Russell Davidson and James G MacKinnon. "Several tests for model specification in the presence of alternative hypotheses." *Econometrica: Journal of the Econometric Society*, pp. 781–793, 1981.

[DM04] Russell Davidson, James G MacKinnon, et al. *Econometric theory and methods*, volume 5. Oxford University Press New York, 2004.

[DOC18] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss." *arXiv preprint arXiv:1804.10916*, 2018.

[DYL12] Fergus Davnall, Connie SP Yip, Gunnar Ljungqvist, Mariyah Selmi, Francesca Ng, Bal Sanghera, Balaji Ganeshan, Kenneth A Miles, Gary J Cook, and Vicky Goh. "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?" *Insights into imaging*, **3**(6):573–589, 2012.

[FDA19] FDA. "About Biomarkers and Qualification." `https://www.fda.gov/drugs/biomarker-qualification-program/about-biomarkers-and-qualification/`, 2019. [Online; accessed May 6, 2021].

[Fed13] Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 2013.

[GBC16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[GGN11] Vicky Goh, Balaji Ganeshan, Paul Nathan, Jaspal K Juttla, Anup Vinayan, and Kenneth A Miles. "Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker." *Radiology*, **261**(1):165–171, 2011.

[Gua18] The Guardian. "Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian." `https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe/`, 2018. [Online; accessed May 9, 2021].

[HFM16] John P Hutchinson, Andrew W Fogarty, Tricia M McKeever, and Richard B Hubbard. "In-hospital mortality after surgical lung biopsy for interstitial lung disease in the United States. 2000 to 2011." *American journal of respiratory and critical care medicine*, **193**(10):1161–1167, 2016.

[HLV17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[Hyu07] Jung Kim Hyun. *Classification in Thoracic Computed Tomography Image Data*. University of California, Los Angeles, 2007.

[HZC17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861*, 2017.

[HZF15] Torsten Hothorn, Achim Zeileis, Richard W Farebrother, Clint Cummins, Giovanni Millo, David Mitchell, and Maintainer Achim Zeileis. "Package 'lmtest'." *Testing linear regression models. https://cran. r-project. org/web/packages/lmtest/lmtest. pdf. Accessed*, **6**, 2015.

[HZR16] 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.

[JLL18] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. "Learn to pay attention." *arXiv preprint arXiv:1804.02391*, 2018.

[JNS99] Yulei Jiang, Robert M Nishikawa, Robert A Schmidt, Charles E Metz, Maryellen L Giger, and Kunio Doi. "Improving breast cancer diagnosis with computer-aided diagnosis." *Academic radiology*, **6**(1):22–33, 1999.

[KBC15] Hyun J Kim, Matthew S Brown, Daniel Chong, David W Gjertson, Peiyun Lu, Hak J Kim, Heidi Coy, and Jonathan G Goldin. "Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months." *Academic radiology*, **22**(1):70–80, 2015.

[KBE11] Hyun J Kim, Matthew S Brown, Robert Elashoff, Gang Li, David W Gjertson, David A Lynch, Diane C Strollo, Eric Kleerup, Daniel Chong, Sumit K Shah, et al. "Quantitative texture-based assessment of one-year changes in fibrotic reticular patterns on HRCT in scleroderma lung disease treated with oral cyclophosphamide." *European radiology*, **21**(12):2455–2465, 2011.

[KBL17] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks." In *International conference on information processing in medical imaging*, pp. 597–609. Springer, 2017.

[KPS11] Talmadge E King Jr, Annie Pardo, and Moisés Selman. "Idiopathic pulmonary fibrosis." *The Lancet*, **378**(9807):1949–1961, 2011.

[KTC10] HJ Kim, DP Tashkin, P Clements, G Li, MS Brown, R Elashoff, DW Gjertson, F Abtin, DA Lynch, DC Strollo, et al. "A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients." *Clinical and experimental rheumatology*, **28**(5 Suppl 62):S26, 2010.

[KWB20] Grace Hyun J Kim, Stephan S Weigt, John A Belperio, Matthew S Brown, Yu Shi, Joshua H Lai, and Jonathan G Goldin. "Prediction of idiopathic pulmonary fibrosis progression using early quantitative changes on CT imaging for a short term of clinical 18–24-month follow-ups." *European radiology*, **30**(2):726–734, 2020.

[LDT20] Yang Lei, Xue Dong, Zhen Tian, Yingzi Liu, Sibo Tian, Tonghe Wang, Xiaojun Jiang, Pretesh Patel, Ashesh B Jani, Hui Mao, et al. "CT prostate segmentation based on synthetic MRI-aided deep attention fully convolution network." *Medical physics*, **47**(2):530–540, 2020.

[LFK19] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, and others. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis." *The lancet digital health*, **1**(6):e271–e297, 2019. Publisher: Elsevier.

[LGL20] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. "Explainable artificial intelligence: Concepts, applications, research challenges and visions." In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 1–16. Springer, 2020.

[LGT20] Lisa Lancaster, Jonathan Goldin, Matthias Trampisch, Grace Hyun Kim, Jonathan Ilowite, Lawrence Homik, David L Hotchkin, Mitchell Kaye, Christopher J Ryerson, Nesrin Mogulkoc, et al. "Effects of Nintedanib on Quantitative Lung Fibrosis Score in Idiopathic Pulmonary Fibrosis." *The open respiratory medicine journal*, **14**:22, 2020.

[LGY08] Chang Hyun Lee, Jin Mo Goo, Hyun Ju Ye, Sung-Joon Ye, Chang Min Park, Eun Ju Chun, and Jung-Gi Im. "Radiation dose modulation techniques in the multidetector CT era: from basics to practice." *Radiographics*, **28**(5):1451–1459, 2008. Publisher: Radiological Society of North America.

[Lip18] Zachary C Lipton. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, **16**(3):31–57, 2018.

[LJC17] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. "Deep learning in medical imaging: general overview." *Korean journal of radiology*, **18**(4):570, 2017.

[LPM15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025*, 2015.

[LWB15] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, Diana L Miglioretti, Breast Cancer Surveillance Consortium, et al. "Diagnostic accuracy of digital screening mammography with and without computer-aided detection." *JAMA internal medicine*, **175**(11):1828–1837, 2015.

[LWP18] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. "Tell me where to look: Guided attention inference network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9215–9223, 2018.

[LWV14] AV Lebedev, Eric Westman, GJP Van Westen, MG Kramberger, Arvid Lundervold, Dag Aarsland, H Soininen, I Kłoszewska, P Mecocci, M Tsolaki, et al. "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness." *NeuroImage: Clinical*, **6**:115–125, 2014.

[LYM19] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets." *Nature biomedical engineering*, **3**(3):173–182, 2019.

[MBK06] Cynthia H McCollough, Michael R Bruesewitz, and James M Kofler Jr. "CT dose reduction and dose management tools: overview of available options." *Radiographics*, **26**(2):503–512, 2006. Publisher: Radiological Society of North America.

[MDS18] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. "Learning visual question answering by bootstrapping hard attention." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–20, 2018.

[MGA15] James L Mulshine, David S Gierada, Samuel G Armato III, Rick S Avila, David F Yankelevitz, Ella A Kazerooni, Michael F McNitt-Gray, Andrew J Buckler, and Daniel C Sullivan. "Role of the quantitative imaging biomarker alliance in optimizing ct for the evaluation of lung cancer screen–detected nodules." *Journal of the American College of Radiology*, **12**(4):390–395, 2015.

[MMK18] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation." In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pp. 1038–1042. IEEE, 2018.

[MSM18] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks." *Digital Signal Processing*, **73**:1–15, 2018.

[NAB11] Paul W Noble, Carlo Albera, Williamson Z Bradford, Ulrich Costabel, Marilyn K Glassberg, David Kardatzke, Talmadge E King Jr, Lisa Lancaster, Steven A Sahn, Javier Szwarcberg, et al. "Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials." *The Lancet*, **377**(9779):1760–1769, 2011.

[NCR12] Luba Nalysnyk, Javier Cid-Ruzafa, Philip Rotella, and Dirk Esser. "Incidence and prevalence of idiopathic pulmonary fibrosis: review of the literature." *European Respiratory Review*, **21**(126):355–361, 2012.

[PAM19] Ian Pan, Saurabh Agarwal, and Derek Merck. "Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks." *Journal of digital imaging*, **32**(5):888–896, 2019. Publisher: Springer.

[PF13] Foster Provost and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." *Big data*, **1**(1):51–59, 2013.

[PMW19] Constantin Pape, Alex Matskevych, Adrian Wolny, Julian Hennies, Giulia Mizzon, Marion Louveaux, Jacob Musser, Alexis Maizel, Detlev Arendt, and Anna Kreshuk. "Leveraging domain knowledge to improve microscopy image segmentation with lifted multicuts." *Frontiers in Computer Science*, **1**:6, 2019.

[Puk06] Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

[QSS09] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[RA94] Patrick Royston and Douglas G Altman. "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **43**(3):429–453, 1994.

[RAS99] Patrick Royston, Gareth Ambler, and Willi Sauerbrei. "The use of fractional polynomials to model continuous risk variables in epidemiology." *International journal of epidemiology*, **28**(5):964–974, 1999.

[RCE11] Ganesh Raghu, Harold R Collard, Jim J Egan, Fernando J Martinez, Juergen Behr, Kevin K Brown, Thomas V Colby, Jean-François Cordier, Kevin R Flaherty, Joseph A Lasky, et al. "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management." *American journal of respiratory and critical care medicine*, **183**(6):788–824, 2011.

[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.

[RRM18] Ganesh Raghu, Martine Remy-Jardin, Jeffrey L Myers, Luca Richeldi, Christopher J Ryerson, David J Lederer, Juergen Behr, Vincent Cottin, Sonye K Danoff, Ferran Morell, et al. "Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline." *American journal of respiratory and critical care medicine*, **198**(5):e44–e68, 2018.

[RS08] Patrick Royston and Willi Sauerbrei. *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, volume 777. John Wiley & Sons, 2008.

[RSN21] RSNA. "Quantitative Imaging Biomarkers Alliance." `https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance/`, 2021. [Online; accessed May 6, 2021].

[SCD17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[SD19] Ashish Sinha and Jose Dolz. "Multi-scale guided attention for medical image segmentation." *arXiv preprint arXiv:1906.02849*, 2019.

[Sha95] Juliet Popper Shaffer. "Multiple hypothesis testing." *Annual review of psychology*, **46**(1):561–584, 1995.

[SIK86] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. "Akaike information criterion statistics." *Dordrecht, The Netherlands: D. Reidel*, **81**(10.5555):26853, 1986.

[SOS19] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention gated networks: Learning to leverage salient regions in medical images." *Medical image analysis*, **53**:197–207, 2019.

[SWG19] Yu Shi, Weng Kee Wong, Jonathan G Goldin, Matthew S Brown, and Grace Hyun J Kim. "Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization-Random forest approach." *Artificial intelligence in medicine*, **100**:101709, 2019.

[SWS17] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis." *Annual review of biomedical engineering*, **19**:221–248, 2017.

[SWU18] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. "Maximum classifier discrepancy for unsupervised domain adaptation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*, 2014.

[VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[WCS16] Simon LF Walsh, Lucio Calandriello, Nicola Sverzellati, Athol U Wells, and David M Hansell. "Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT." *Thorax*, **71**(1):45–51, 2016.

[WCS18] Simon LF Walsh, Lucio Calandriello, Mario Silva, and Nicola Sverzellati. "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study." *The Lancet Respiratory Medicine*, **6**(11):837–845, 2018.

[WGG18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

[WL20] Jonas Widell and Mats Lidén. "Interobserver variability in high-resolution CT of the lungs." *European journal of radiology open*, **7**:100228, 2020.

[WLR19] Simon LF Walsh, David J Lederer, Christopher J Ryerson, Martin Kolb, Toby M Maher, Richard Nusser, Venerino Poletti, Luca Richeldi, Carlo Vancheri, Margaret L Wilsher, et al. "Diagnostic likelihood thresholds that define a working diagnosis of idiopathic pulmonary fibrosis." *American journal of respiratory and critical care medicine*, **200**(9):1146–1153, 2019.

[WLZ20] Xiaoqin Wang, Gongbo Liang, Yu Zhang, Hunter Blanton, Zachary Bessinger, and Nathan Jacobs. "Inconsistent Performance of Deep Learning Models on Mammogram Classification." *Journal of the American College of Radiology*, 2020. Publisher: Elsevier.

[WSN14] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. "scikit-image: image processing in Python." *PeerJ*, **2**:e453, 2014. Publisher: PeerJ Inc.

[YKH19] Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. "Melanoma recognition via visual attention." In *International Conference on Information Processing in Medical Imaging*, pp. 793–804. Springer, 2019.

[YKK19] Heechan Yang, Ji-Ye Kim, Hyongsuk Kim, and Shyam P Adhikari. "Guided soft attention network for classification of breast cancer histopathology images." *IEEE transactions on medical imaging*, **39**(5):1306–1315, 2019.

[YZC21] Wenxi Yu, Hua Zhou, Youngwon Choi, Jonathan G Goldin, Pangyu Teng, and Grace Hyun J Kim. "An automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using domain knowledge-guided attention models in HRCT images." In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, p. 115971Y. International Society for Optics and Photonics, 2021.

[YZG21] Wenxi Yu, Hua Zhou, Jonathan G Goldin, Weng Kee Wong, and Grace Hyun J Kim. "End-to-end Domain Knowledge Assisted Automatic Diagnosis of Idiopathic Pulmonary Fibrosis (IPF) Using Computed Tomography (CT)." *Medical Physics*, 2021. Publisher: Wiley Online Library.

[ZBL18] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." *PLoS medicine*, **15**(11):e1002683, 2018.

[ZDG19] Martin Zlocha, Qi Dou, and Ben Glocker. "Improving retinanet for ct lesion detection with dense masks from weak recist labels." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 402–410. Springer, 2019.

[ZGM19] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. "Self-attention generative adversarial networks." In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

[ZHL19] Shu Zhang, Fangfang Han, Zhengrong Liang, Jiaxing Tan, Weiguo Cao, Yongfeng Gao, Marc Pomeroy, Kenneth Ng, and Wei Hou. "An investigation of CNN models for differentiating malignant from benign lesions using small pathologically proven datasets." *Computerized Medical Imaging and Graphics*, **77**:101645, 2019.

[Zho18] Zhi-Hua Zhou. "A brief introduction to weakly supervised learning." *National science review*, **5**(1):44–53, 2018.

[ZKL16] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.