Convolutional neural network based image classification for Idiopathic pulmonary fibrosis and nonspecific interstitial pneumonia using 129Xe gas exchange imaging

Junlan Lu*
Medical Physics Graduate Program
Duke University
Durham, NC 27705
junlan.lu@duke.edu

Abstract

Hyperpolarized (HP) gas exchange imaging is a novel technique that is able to assess regional pulmonary function of the lung[1]. It is able to characterize between diseases such as chronic obstructive pulmonary disease and left heart failure[5]. We investigate the feasibility of building a convolutional neural network (CNN) machine learning model to classify between non-specific interstitial pneumonia (NSIP) or interstitial pulmonary fibrosis (IPF) in HP gas images.

1 Introduction

Hyperpolarized Gas Transfer MRI has emerged as a promising technique to assess pulmonary function which measures 129Xe in the airspaces, interstitial barrier tissues, and red blood cells[1]. We plan on investigating the ability of 129Xe images in these three compartments to classify images to be either of non-specific interstitial pneumonia (NSIP) or interstitial pulmonary fibrosis (IPF) as well as to characterize these differences. The central hypothesis is that due to the two diseases being separate entities, there will be a distinction in either the regional ventilation or regional gas exchange as it almost comprehensively characterizes lung physiology. This is of great clinical importance and interest because it is currently difficult to diagnose between the two diseases without the need for lung biopsy[2]. Differentiating between these two diseases is critical as the prognoses are different[2] and there exist specialized treatments, which are extremely costly to the patient, for each disease[3]. Although there are comparisons between the diseases using lung biopsy and conventional CT[2], no studies have been performed differentiating IPF from NSIP with HP gas MRI images.

Effectively classifying medical images plays an essential role in aiding clinical care and treatment. Deep CNNs are widely used in changing image classification tasks and have achieved significant performance since its advent as the dense structure of CNNs with learned convolutional layers and deep fully connected weights offers better data representation and generalizability than most traditional machine learning approaches[6]. Such advances in deep learning-based image analysis have spurred research in radiology, such as segmentation, detection, and diagnosis/prognosis[8][9][7]. The purpose of this study was to develop a CNN framework to classify between NSIP and IPF using HP gas exchange imaging and assess the model's features.

*junlanlu.com

2 Related work

Current literature shows that is possible to use deep learning to identify diseases such as COVID-19 from other pneumonia with high accuracy[12]. Christe et al. showed that it is possible for pulmonary fibrosis to aid in deep learning to identify features of IPF in CT scans[13]. Similarly, Walsh et al. demonstrated a deep learning algorithm for classification of fibrotic lung disease on high resolution CT[14]. However, in all of these mentioned studies, high resolution CT images were used instead of the less widely adopted HP gas MRI. We hope to be able to replicate such accuracy with HP Gas MRI rather than high resolution CT images.

3 Methods

3.1 Data preparation

At the time of writing, our dataset consists of HP MRI scans of 55 IPF patient scans and 36 NSIP patient scans. Some of the scans may be repeat scans of the same patient, but we treat repeat scans as a separate data point. This makes a total of 91 observations. Each MRI scan consist of MRI images in three phases: 129Xe dissolved in the blood, 129Xe dissolved in the lung parenchyma, and 129Xe in the alveolar airspaces. All images are reconstructed to a 128x128x128 volume. Labeled MRI images are originally provided in MAT files and then organized and consolidated into an HDF5 file. Images are then normalized to be between values of 0 to 1 or 0 to 255 depending on the different models that we implement below. Images of the three phases are coalesced together to form the 3 "color" channels.

3.2 CNN architecture and classification model for diagnosis

In this project, we implement a 2D CNN with multiple instance learning (MIL). MIL is a type of problem where a single observation is composed of numerous instances – in this case, a single observation is the entire 3D patient image and the instances are the individual slices. Thus for this project, 3D MRI volumes are re-parameterized as collection of 2D cross sectional images. Although many MIL pooling mechanisms exists such as mean and max pooling, we use a mechanism that is fully parametrized by neural networks as described in [10]. This MIL pooling mechanism implements an attention layer which has the benefit of interpretation. For the CNN structure, we use the pre-trained VGG16 base without the top classification layer for feature extraction. This method of transfer learning takes advantage of reusing features and weights learned from a similar dataset as a starting point for training a model on a dataset of interest. In our case of limited observations, it is important to train on larger datasets to extract features that are both relevant and generalizable to unseen data. In this work, we implement only one stage of transfer learning which is VGG16 trained on the ImageNet dataset. On top of the VGG16 feature extraction layers are two fully connected dense layers that serve as the classification head which feeds into the attention mechanism. The output of the attention layer is then passed through a final dense layer with sigmoid activation for classification between the two diseases. The architecture is depicted in figure 2.

To train the model, we first load the weights pre-trained on the ImageNet dataset into the VGG16 network without the top classification head. This equips our model to handle lines, borders, and textures often found in the real world that will generalize well to new problems. Although a second stage of transfer learning should be done to learn more about the dataset in the HP gas imaging domain, we did not have the time and computational resources to do so. After loading weights, the base VGG16 network was imported into the deep MIL model and the attention layer + new fully connected classification head was trained on our HP Gas MRI dataset. For training, the data was split into 5 folds to be used in a k-fold cross validation scheme. We used the Adam optimizer with a learning rate of 0.0001 for 50 epochs and a batch size of 1 observation per step.

4 Results

From the parameters above, an average training accuracy of 100% and validation accuracy of 86.8% is achieved respectively. The results of the physical layer weights initialized to be $[0.5, 0.5, 0.5]$ and constrained to be between 0 - 1 are $[0.51, 0.50, 0.49]$ in the end with the relative importance being 129Xe in the RBC, barrier, and gas phase respectively.

5 Discussion

There are extreme challenges with evaluating the features that the model is looking at. For example, the model could be memorizing the shape of the lung rather than the intensities inside it. Another challenge is the limited size of our dataset. Although we have 55 observations of IPF, many of these observations are repeat scans which will inherently produce very similar features like lung volume shape. Thus, our assumptions of treating repeated scans as separate independent observations might be a hindrance. The issue of repeated scans could artificially inflate the validation accuracy as scans of the same patient but different time points could be both in the test and validation batch. If the model learns shapes rather than intensities, it would be possible to memorize the shape of the lung and classify correctly. Parallel to the problem of memorizing lung shapes as a result of having too few training data, we need to be able to assess the important features in the image. This can be done by implementing a class activation map, of which, gradient-weighted class activation maps are known to be useful [11]. Ultimately, in order to turn this work to a future publication, it will be important to identify features in the image that are characteristic of each disease – features simple enough that a radiologist would be able to understand.

In addition, the current model assumes that there is no relationship between each bag of features which is not true in our case because each bag comes from the ordered slice of the same image. There will naturally be features that will be very similar from slice to slice. Another issue is determining the best set of image channels to look at. We initialize the weights of the physical layer such that each image channel has equal importance, but that may not be a physiologically correct assumption.

In the future, we plan to implement a number of additions and improvements to help with the challenge of dealing with a small radiological dataset. Firstly, we will incorporate an additional layer of transfer learning, unfreezing the weights of the convolutional base of VGG16 to train on the rest of the HP gas image dataset. Secondly, we need to incorporate image augmentation such that we deform images such that the network will be more generalizable to more lung shapes. An aside to image augmentation is image simulation, where we simulate images of healthy patients and diseased patients to aid in transfer learning. A simple task to examine how much lung shape matters in classification is to pass through a binary mask of the lung shape and evaluate the accuracy of the model. To deal with the issue of our data, we will explore different ways to preprocess the data, perhaps look at the lung in different orientations. We must also deal with the issue of an imbalanced dataset in evaluating our model performance. One possibility to explore is the use of 3D CNNs, however, 3D CNNs are still in its nascent and have higher memory requirement.

In this work we have demonstrated a preliminary framework for building a classifier for NSIP and IPF. However, significant work needs to be completed in the realms of data preparation and organization and model evaluation.

6 Figures

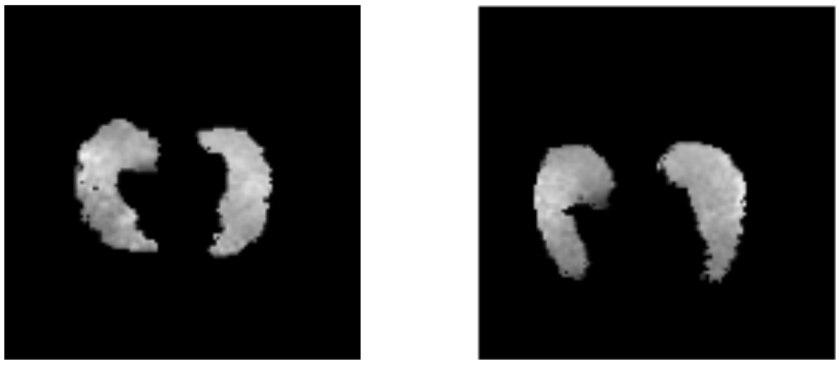

Figure 1: Example of Xe129 MRI scan cross sectional image in the dissolved, barrier phase taken from a subject with IPF (left) and a subject with NSIP (right).

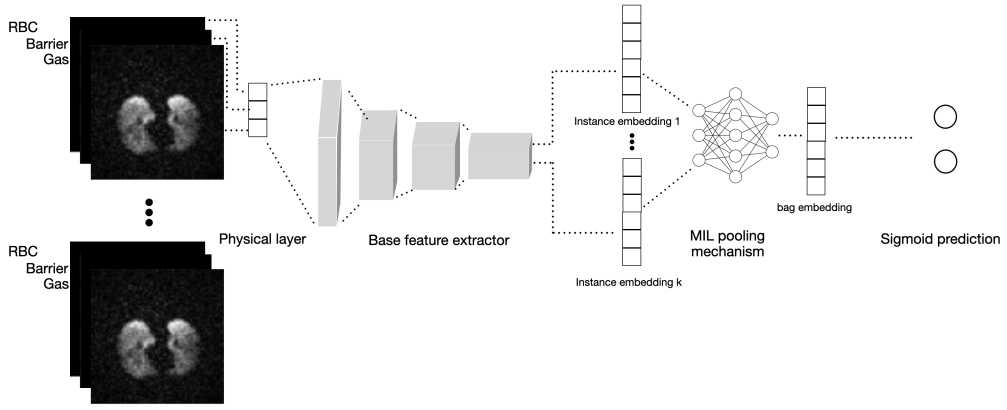

Figure 2: Deep MIL model architecture. Each slice serves as an instance. There is a physical layer that collapses the 3 color channels to 1 dimension. The K128x128x1 sized instance is passed through the base feature extractor consisting of weights pre-trained using the VGG16 architecture trained on the ImageNet database. The feature extractor encodes each isntance into a N-dimensional embedding and the K N-dimensional instance embeddings are pooled in the attention layer as proposed by Ilse et al.[10]. The attention layer combines each instance embedding into a multi-slice prediction which is used ot make a class prediction of other NSIP and IPF.

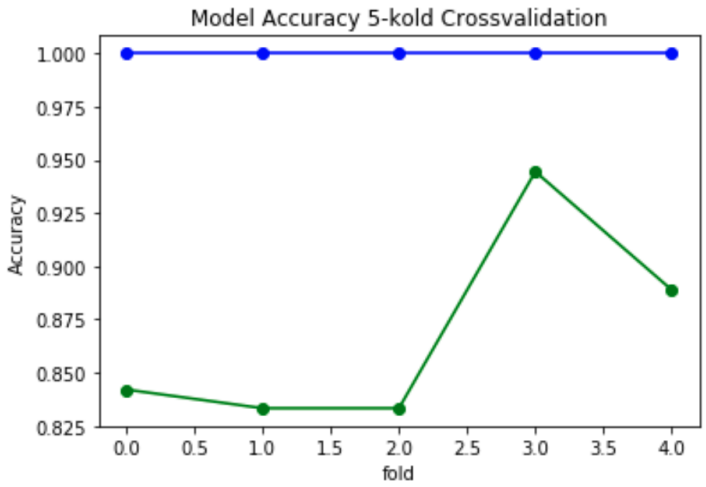

Figure 3: Accuracy as a function of fold with 5-fold cross validation. Train accuracy is shown in blue and validation accuracy is shown in green. Average training accuracy is 100% and average validation accuracy is 86.8%.

Acknowledgments

The author is funded by the NSF graduate research fellowship. Additional acknowledgment goes to Billy Carson from the BME department for guidance in this project.

References

[1] Wang, Z. et al. Quantitative analysis of hyperpolarized 129Xe gas transfer MRI. Med. Phys. 44, 2415–2428 (2017).

[2] Glaspole, I. & Goh, N. S. L. Differentiating between IPF and NSIP. Chron. Respir. Dis. 7, 187–195 (2010).

[3] Margaritopoulos, G. A., Vasarmidi, E. & Antoniou, K. M. Pirfenidone in the treatment of idiopathic pulmonary fibrosis: An evidence-based review of its place in therapy. Core Evid. 11, 11–22 (2016)

[4] Robertson, S. H. et al. Optimizing 3D noncartesian gridding reconstruction for hyperpolarized 129 Xe MRI-focus on preclinical applications. Concepts Magn. Reson. Part A 44, 190–202 (2015).

[5] Wang, Z. et al. Diverse cardiopulmonary diseases are associated with distinct xenon magnetic resonance imaging signatures. doi:10.1183/13993003.00831-2019

[6] Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review neural computation; 2017.

[7] T. Retson, A.H. Besser, S. Sall, D. Golden, A. Hsiao Machine learning and deep neural networks in thoracic and cardiovascular imaging J. Thorac. Imaging, 34 (3) (2019), pp. 192-201

[8] P. Lakhani, B. Sundaram Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks Radiology, 284 (2) (2017), pp. 574-582

[9] L.M. Prevedello, B.S. Erdal, J.L. Ryu, K.J. Little, M. Demirer, R.D. White Automated critical test findings identification and online notification system using artificial intelligence Radiology, 285 (2017), pp. 923-931

[10] Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. 35th Int. Conf. Mach. Learn. ICML 2018 5, 3376–3391 (2018).

[11] Localization, G., Dec, C. V & Cogswell, M. Grad-CAM : Visual Explanations from Deep Networks. 1–23

[12] Bai, H., Wang, R. & Liao, W.-H. AI Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Etiology on Chest CT. Radiology (2020). doi:10.14358/PERS.80.2.000

[13] Christe, A. et al. Computer-Aided Diagnosis of Pulmonary Fibrosis Using Deep Learning and CT Images. Invest. Radiol. 54, 627–632 (2019).

[14] Walsh, S. L. F., Calandriello, L., Silva, M. & Sverzellati, N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir. Med. 6, 837–845 (2018).