Convolutional Neural Networks for Efficient Localization of Interstitial Lung Disease Patterns in HRCT Images

Sunita Agarwala[1], Abhishek Kumar[2], Debashis Nandi[1], Ashis Kumar Dhara[3], Anup Sadhu[4], Sumitra Basu Thakur[4] and Ashok Kumar Bhadra[4]

[1]Computer Science & Engineering, National Institute of Technology Durgapur,INDIA.
[2]School of Computer and Information Sciences, University of Hyderabad, INDIA.
[3]Centre for Image Analysis, Uppsala University, SE-751 05 Uppsala, SWEDEN.
[4]Medical College Kolkata,INDIA.

Abstract. Lung field segmentation is the first step towards the development of any computer aided diagnosis (CAD) system for interstitial lung diseases (ILD) observed in chest high resolution computed tomography (HRCT) images. If the segmentation is not done efficiently it will compromise the accuracy of CAD system. In this paper, a deep learning-based method is proposed to localize several interstitial lung disease patterns (ILD) in HRCT images without performing lung field segmentation. In this paper, localization of several ILD patterns is performed in image slice. The pretrained models of ZF and VGG networks were fine-tuned in order to localize ILD patterns using Faster R-CNN framework. The three most difficult ILD patterns consolidation, emphysema, and fibrosis have been used for this study and the accuracy of the method has been evaluated in terms of mean average precision (mAP) and free receiver operating characteristic (FROC) curve. The model achieved mAP value of 75% and 83% on ZF and VGG networks, respectively. The result obtained shows the effectiveness of the method in the localization of different ILD patterns.

1 Introduction

Interstitial lung diseases, also known as diffused parenchymal lung disease is a generalized term used to refer to a group of diseases encompassing nearly 200 different pulmonary diseases based on the patterns formed by them in the lung field. Though the interstitial lung diseases are a heterogeneous group they show similar clinical manifestations which make the differentiation among them difficult and also accounts for the greater inter and intra observer variability. This makes the differential diagnosis difficult even for the experienced medical experts. The main symptom common to this group is inflammation of interstitium, which is responsible for providing support to the lungs microscopic air sacs. If diagnosed early, they can be treated and recovery time will also be less. HRCT images are considered best imaging tools to study ILD patterns, due to their high resolution. Most common ILD patterns found in HRCT images are consolidation, emphysema and fibrosis. Figure. 1 shows these patterns together with a normal lung image. ILD patterns are generally considered as a textural aberration in lung tissues. Therefore, the majority of the earlier works focused on the extraction of feature by using different techniques and then further classification of these extracted features with the help

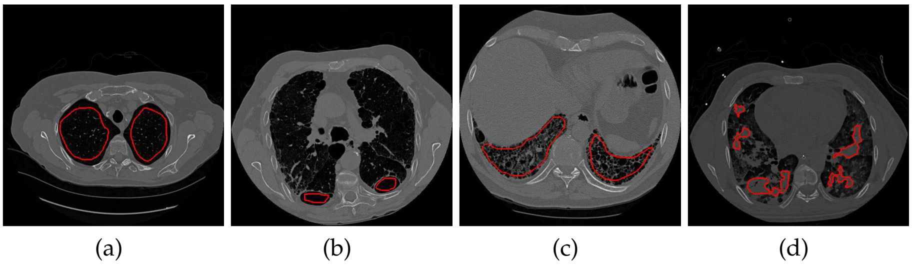


Fig. 1: Different type of ILD patterns: (a) normal lung (b) emphysema (c) fibrosis and (d) consolidation

of different classifiers.

Recently, after the impressive performance of deep learning methods in several image recognition and classification contests, the researchers have turned their attention towards deep learning based methods. In [1], some features of CNNs had been incorporated such as weight sharing among the hidden layers, which were further totally connected to the output neurons. The network was trained in a supervised manner using gradient descent. The main task of feature extraction and classification, however, was done by a modified RBM. In [2], a pre-trained model of AlexNet was used to fine tune the lung data and later in the classification of lung slices. To fit the architectural design of AlexNet they have rescaled the input images and artificially generated the three channels with the help of different Hounsfield unit (HU) windows. In [3], the authors tried to classify the six ILD patterns by proposing a convolutional neural network. The network design is not very complex with 5 convolutional layers with the kernel size of 2×2 and LeakyRelu is used as an activation function. The pooling layer size is equal to the size of the final feature maps. In the last, there are three fully connected layers. The network achieved a performance of nearly 85.5% in the classification of lung patterns. In [4], the authors proposed a CNN with only one convolutional layer and three fully connected layers. The shallow architecture of the network prevented it from taking the full benefit of the power of deep CNN layers. In all the above mentioned methods, the main challenge lies in automatic localization of ILD patterns.

In this paper, the three ILD patterns have been localized and classified using different networks, without doing lung segmentation. The rest of the paper is organized as follows: In section II, database part has been discussed along with different network architectures and training methods. In the next section, the implementation details and results obtained have been discussed in the context of various relevant metrices. Finally, the paper is concluded with a conclusion and future work part.

2 MATERIALS AND METHODS

2.1 Database Description

A publicly available database from MedGIFT [5] has been used for this study. In the preparation of database, initially a raw list of 1266 patients was taken. Up to now, more than 700 cases were revised and 128 cases were stored in the database. For the cases where the number of patterns is more than one in the same image slice, the patterns are labelled to identify which ROI belongs to which pattern. Only those patterns were considered by the annotators which sharply resembles the class considered and ambiguous tissue area has not been taken into consideration.

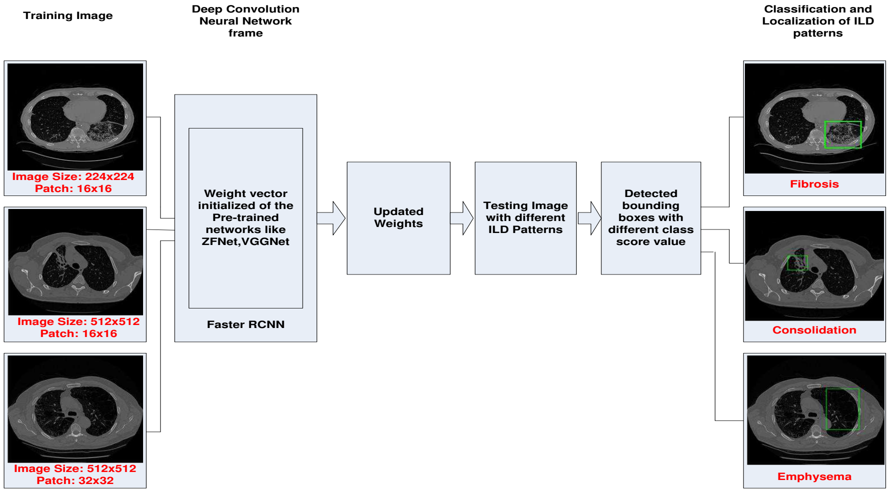

Training Image
Deep Convolution Neural Network frame
Classification and Localization of ILD patterns
Image Size: 224x224 Patch: 16x16
Image Size: 512x512 Patch: 16x16
Image Size: 512x512 Patch: 32x32
Weight vector initialized of the Pre-trained networks like ZFNet,VGGNet
Faster RCNN
Updated Weights
Testing Image with different ILD Patterns
Detected bounding boxes with different class score value
Fibrosis
Consolidation
Emphysema

Fig. 2: System block diagram

2.2 Networks used in proposed model

Due to low availability of quality annotated data, instead of training the model from scratch pre-trained weights of ZF and VGG net have been used and fine-tuned to fit our data. The architectural design of all these networks is shown in brief:

ZF Net: The design of the ZF network [6] was inspired by its predecessor AlexNet [7] to improve the performance of image classification. The network consists of an input layer, five convolution layers, two fully-connected layers, and a softmax layer. The input layer of ZF network supports the image size of 224×224. Therefore, we have considered image size of 224×224 as well as 512×512 in this paper. In the first convolution layer, ZF net used 96 filters of size 7×7 with decreased stride value as compared to 11×11 filter size of AlexNet. The smaller size 7×7 filter is more efficient to fetch fine details of features compared to big size 11×11 filters. The number of filters used in subsequent convolution layers is 256, 384, 384 and 256. The size of filters in subsequent layers are 5×5, 3×3, 3×3, 3×3. ReLUs have been used for activation functions, a cross-entropy loss for the error function, and trained using batch stochastic gradient descent algorithm.

VGG Net: VGG Net [8] was created by Karen Simonyan and Andrew Zisserman of the University of Oxford with a basic idea to keep the architectural design, *simple* and *deeper*. The model was runner-up at ILSVRC 2014 with 7.3% error rate. They have created a bunch of networks with layer weights varying between 11 to 19. The best result was obtained by 16 weight layers. They have used a filter size of 3×3 throughout all the layers, with stride and pad size of 1. The size of a max-pooling layer is also fixed *i.e.* 2 × 2 with stride 2. The reasoning behind this decreased filter size is that a stack of 2, 3 × 3 convolution layers offers an effective receptive field of 5 × 5, a stack of 3 such layers offers a receptive field of size 7 × 7. The benefit of such design is that here 2 or 3 nonlinear activation layers can be inserted depending on the filter size instead of just 1 layer. The increased nonlinear activation layers can make the decision function more discriminative. One more benefit of such design is that the number of parameters also got reduced with reduced filter size.

2.3 Training of Network

The network was simulated for ZF and VGG Net on FASTER R-CNN framework provided by [9]. This model of FASTER R-CNN is simply an advanced and more robust version of FAST RCNN [10] by the same author. In this work, the Region Proposal Network(RPN) and FAST R-CNN both are merged into a single network by sharing their convolutional features. After finding the region boundary for an object with the help of region proposal network module, class score for each region is calculated and then the region having highest score for that particular object is kept. A lot of smaller changes also has been made to fit the FASTER-RCNN framework with this database. All these changes are listed below:
Base Network: Pre-trained weights from ZF, VGG have been used for calculating region-of-interest (ROI).
Training/Testing: The default end-to-end training and testing scheme is used. Learning rate starts with .001 and is kept constant throughout the process. The process is evaluated on $10,000$ iterations.

The proposed model consists of two parts training and testing as shown in Figure 2. The input images got divided between training and testing images into separate text files sequentially. In the database, the ROIs have been labelled. These ROIs were further divided into patch sizes of 16 × 16 and 32 × 32. The image size of 224 × 224 and 512 × 512 have been used in this study. When the image size is 224 × 224 we have used patch size of only 16 × 16 because a lesser or greater patch size is either too small to capture the relevant features inside a single patch or too big that it will capture irrelevant feature also. In case of image size 512 × 512, the patch size is 32 × 32. Later on, the annotated ROIs, training images and pre-trained weights of the network used were fed into Faster-RCNN framework. After training, the model got updated weights trained on relevant features. In the testing part, the text file containing information about testing images has been given as an input. With the help of Faster-RCNN region proposal network, the number of regions will be generated for an object. Further, a class score for each region is calculated and the region having highest score is kept.

2.4 Description of Training and Test Data set

The whole dataset for each pattern has been divided into two. One half is kept for training and its number has been increased artificially by using image augmentation techniques such as flip, translation, rotation etc. Testing data is kept as it is without any augmentation. Table 1. gives the total number of slices for training and testing purposes of each pattern after augmentation.

After getting the polygon for the ROIs, non-overlapping patches of size 16 × 16 and 32 × 32 have been extracted. The patches with more than 75%

Table 1: Augmented data obtained after applying data augmentation techniques on original data.

Pattern	Training data without augmentation	Training data after augmentation	Testing data without augmentation
Consolidation	58	348	58
Emphysema	36	216	35
fibrosis	193	1158	100
Total	287	1722	193

of area lying outside the ROI region were discarded. Figure. 3 shows patch extraction from annotated CT lung slice. Table 2. gives the total number of extracted patches for individual pattern used as training data.

Table 2: Total number of extracted patches from ROIs for each pattern

Pattern	Total number of ROIs	# of 16×16 patches for image size 224×224	# of 32×32 patches for image size 512×512
Consolidation	636	1204	1466
Emphysema	372	1057	1572
fibrosis	1782	8047	10870
Total	2790	10308	13908

3 EXPERIMENTAL SETUP AND RESULTS

3.1 Implementation

The whole setup was trained in linux environment using NVIDIA GTX 1070 6 GB GPU on a system with 16 GB RAM and having corei5 7th generation @ 3.50GHz processor. For all the networks FASTER R-CNN is used as a framework. Methods where convolutional networks are not used are coded in python.

3.2 Performance metrics

Mean Average Precision (mAP): It has been shown to have especially good discrimination [11]. The mAP for a set of detections is the mean over classes, of the interpolated AP for each class. Recall is defined as the ratio of true-positive detection to ground-truth instances, and precision as the ratio of true-positive detection to all detection. For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. That is, if the set of relevant documents for an

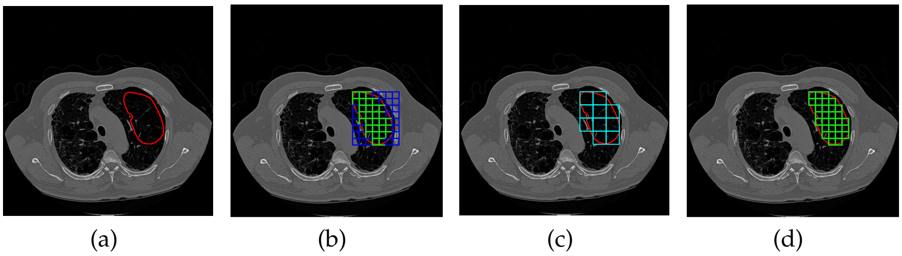


Fig. 3: Patch extraction from ROI. (a) The polygon in the red is the ROI.(b) The non-overlapping square boxes are the patches. The blue color boxes have been discarded because more than 75% of their area is outside the ROI region. (c) The cyan color boxes are the final patches of size 32 × 32. (d) The green color boxes are the final patches of size 16 × 16

information need $q_j \in Q$ is $\{d_1,...dm_j\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then where,

$$mAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\text{Precision}(\text{R}_{\text{jk}}) \tag{1}$$

$$\text{P} = \frac{N_{tp}}{N_{tp}+N_{fp}} \tag{2}$$

$$\text{R} = \frac{N_{tp}}{N_{tp}+N_{fn}} \tag{3}$$

where

N_{tp} - Number of true-positive objects
N_{tn} - Number of true-negative objects
N_{fp} - Number of false-positive objects
N_{fn} - Number of false-negative objects

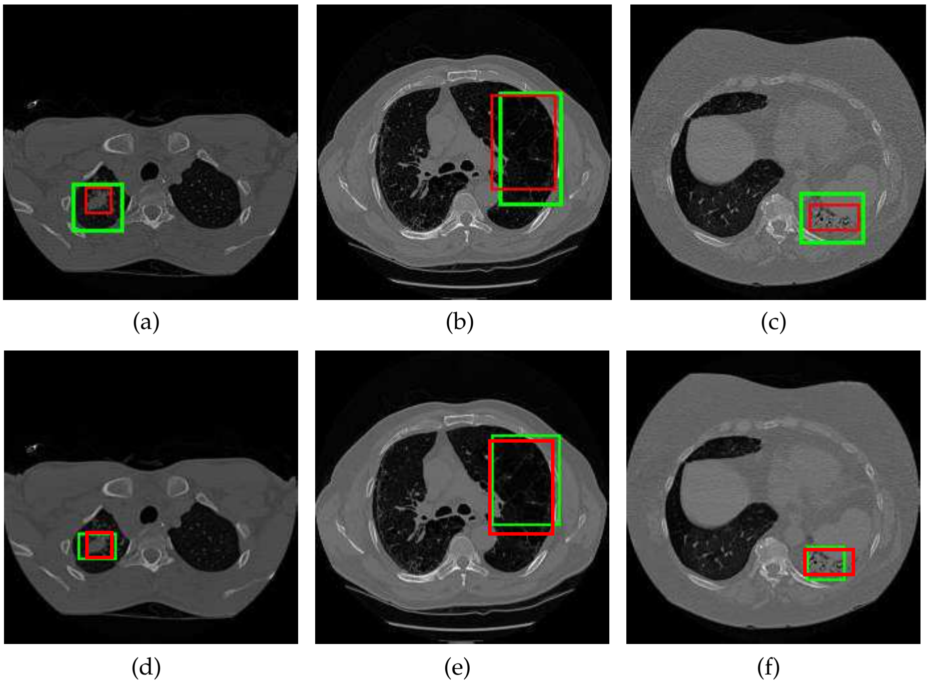
Here, first we have calculated precision of the three patterns individually, by dividing true positives with the total number of detected bounding boxes. After calculating precision, it has to be averaged for whole class and then mean is calculated by dividing average precision value with the total number of classes.

3.3 Quantitative results

Table 3 depicts the quantitative results of different networks with different patch sizes measured in terms of *mAP*. The results were calculated for both image sizes *i.e* 224 × 224 as well as 512 × 512. Moreover, for image size 224 × 224 we have considered the patch size of 16 × 16 whereas for image size of 512 × 512 patch size of 32 × 32 has been taken. Here, we can see that the best mAP value of 83% is obtained with image size of 224 × 224 and patch size 16 × 16 by VGG Net. Moreover, with smaller image size the VGG performs better than ZF net. This could be attributed to the fact that VGG was originally designed for image size 224 × 224 and gives a better result if applied to that size. When image size is 512 × 512 the ZF and VGG give similar results which are in the range of 75%. One more interesting point is that the image size of 224 × 224 with patch size 16 × 16 gives the better result than patch size of 32 × 32 whereas if the image size is 512 × 512 patch size of 32 × 32 gives a better result than patch size of 16 × 16. This could be due to the fact that when the image size is small the smaller patch size captures all the relevant features efficiently whereas the larger patch size includes some irrelevant features also resulting in overall lower accuracy. With bigger image size the larger patch size is more suitable to extract the relevant features. Free Receiver Operating Characteristic (FROC) curve analysis is widely used for the performance evaluation of classification [12, 13]. FROC curve plots the values of sensitivity along the y-axis and false positive per image along a x-axis. The curve can extend indefinitely in the right direction but the ordinate remains to unity or less. The curve shows a non-decreasing nature and tends to converge after some time. The sensitivity and false positive per image have been calculated for different threshold values. The threshold values are the class scores of each bounding box and it lies between 0 to 1. When the threshold is at 0 the sensitivity is at maximum but the number of false positive per image also get increased. This is because of increase in the number of bounding boxes taken as ROIs even for low class scores which often results in false positive. When the threshold value is increased gradually, the sensitivity value, as well as false positive per image, also get reduced due to the lower number of bounding boxes. In Figure 5 FROC plots for selected ILD patterns has been shown for ZF and VGG networks. Figure 5 (a) is the plot for consolidation. Figure 5 (b) is the plot for fibrosis. Here, we can see that all the networks are showing the impressive FROC curve. The sensitivity achieved here is generally 90% by both the networks. Overall, we can see that the best plot is observed on fibrosis with high sensitivity value. This is because of a large number of training data available in case of fibrosis as compared to other two patterns.

Table 3: Quantitative results of different networks with different patch sizes measured in terms of *mAP*.

Image Size	**Patch Size**	**Network**	**Pattern**	**mAP** 10K
512 x 512	32 x 32	ZF	consolidation	73%
			emphysema	
			fibrosis	
		VGG	consolidation	75%
			emphysema	
			fibrosis	
224 x 224	16 x 16	ZF	consolidation	75%
			emphysema	
			fibrosis	
		VGG	consolidation	83%
			emphysema	
			fibrosis	



(a) (b) (c)

(d) (e) (f)

Fig. 4: Segmented output obtained for selected ILD patterns on ZF and VGG net.(a)-(c) is the output image for consolidation, emphysema and fibrosis, respectively on ZF net. (d)-(f) is the output for consolidation, emphysema and fibrosis, respectively on VGG net. The red color box is the original ground truth and green color box represents the detected bounding box.

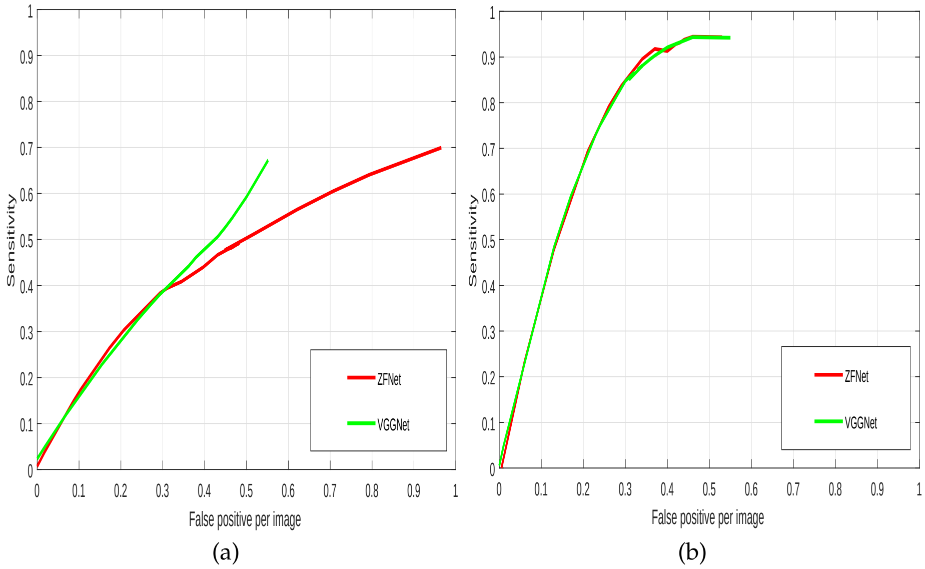


Fig. 5: FROC plot of selected ILD patterns for ZF and VGG networks. (a) is the FROC plot for consolidation with image size 512 × 512 and patch size 32 × 32 and (b) is the FROC plot for fibrosis with image size 512 × 512 and patch size 32 × 32 on ZF net and VGG net.

3.4 Qualitative results

Figure 4 shows the final output of the proposed model for different ILD patterns. The number of patches has been reduced by applying some threshold value and finally, all the patches are merged to form one big rectangular box to compare with ground truth value. Only those patches are merged together which are overlapping with each other a minimum amount of pixel value in common and those patches have been discarded whose more than 75% of the area remains outside the ROI region. Similarly, a rectangular box is drawn from ROI of ground truth which is shown in red color. There are a number of false positive cases also. But, most of the false positive happens because the database itself is partially labeled. Also, if the radiologists contradict each other for any particular region, they have not drawn any ground truth polygon for that region and it remains unlabeled. Due to the ambiguity in ILD patterns most of the times this happens which results in the low accuracy. In all these places the model is bound to give false positives. In general, out of three chosen patterns, the model is showing best results for fibrosis. This is because fibrosis has the largest training data as compared to other two patterns. VGG Net is more accurate as compared to ZF Net due to their lesser filter size properly adopt feature fully and advanced or deep architectural designs.

4 Conclusions

In this paper, an idea has been proposed to localize and classify the different ILD patterns in lung HRCT images. Deep learning based methods such as CNNs, ZF Net, and VGG Net have been used to train the dataset and Faster-RCNN have been used to draw bounding boxes around the ILD patterns. Originally these networks were designed for natural color images which everybody is familiar with so any normal human being can become an annotator for those images. But, in case of medical imaging only highly trained individuals can do the task of annotation and validation. This accounts for the low availability of quality data. To solve this problem, the concept of transfer learning has been used and the effectiveness of the method shows that even with low data good results can be obtained if the network is trained properly. Instead of training the network from scratch or taking some random weights to the network, weights from some well known pre-trained networks trained on sufficiently bigger data size has been taken. The performance can be further improved by labeling the database fully. The future work includes covering more number of ILD patterns and increasing the number of patches and training data.

Acknowledgments

This work is supported by National Institute of Technology, Durgapur by providing the adequate resources and proper scientific environment. Funding is provided by Visvesvaraya PhD scheme of DeitY (Department of Electronics & Information Technology), Govt. of India. The authors are thankful to Medical College Kolkata and EKO DIAGNOSTICS, Kolkata for providing valuable input and feedback for this work.

References

1. G. van Tulder and M. de Bruijne, "Learning features for tissue classification with the classification restricted boltzmann machine," in *International MICCAI Workshop on Medical Computer Vision*. Springer, 2014, pp. 47–58.
2. M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–6, 2016.
3. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.
4. Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*. IEEE, 2014, pp. 844–848.
5. A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized medical imaging and graphics*, vol. 36, no. 3, pp. 227–238, 2012.
6. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
8. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
9. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
10. R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
11. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
12. D. C. Edwards, M. A. Kupinski, C. E. Metz, and R. M. Nishikawa, "Maximum likelihood fitting of froc curves under an initial-detection-and-candidate-analysis model," *Medical physics*, vol. 29, no. 12, pp. 2861–2870, 2002.
13. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.