Estimating Lung Capacity in Pulmonary Fibrosis Patients via Computerized Tomography (CT) Scan Data and Machine Learning

Mark E. Earle and Mahmood Al-khassaweneh
Department of Engineering, Computing and Mathematical Sciences
Lewis University
Romeoville, USA
{markeearle, malkhassaweneh}@lewisu.edu

2021 International Conference on Information Technology (ICIT) | 978-1-6654-2870-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIT52682.2021.9491120

***Abstract*— Pulmonary fibrosis affects 1 in 200 Americans. Current diagnosis methods include spirometry testing that can be difficult for patients with debilitated lung diseases. Often, CT Scans are done on patients as part of the diagnostic test panel. This paper aims to use the non-invasive CT scans of a patient to determine the lung capacity (and forced vital capacity) without the patient needing to do a spirometry test.**

This paper looked at 172 patients with associated CT scans and spirometry lung capacity results. Using the Hounsfield Units to identify the lung tissue, training images were produced to train machine learning models. The results of the machine learning models indicate that the Hounsfield Units were not an adequate method for determining lung tissue and lung capacity in CT scan images. The reliability and accuracy of spirometry tests were also discussed.

Keywords—machine learning, Computerized Tomography, DICOM, Python, pulmonary fibrosis, image processing

I. Introduction

Pulmonary fibrosis effects an estimated 1 in 200 adults over the age of 60 in the United States. Pulmonary ('Lung') and Fibrosis ('Scar Tissue') defines the medical condition where scar tissue develops inside the lungs, limiting the functionality of the lungs to deliver oxygen to the blood stream. [1] The competition data mining website, Kaggle.com, in their OSIC Pulmonary Fibrosis Progression competition webpage, states, "Current methods make fibrotic lung diseases difficult to treat, even with access to a chest CT scan…patients suffer extreme anxiety…from the disease's opaque path of progression." This patient anxiety stems, partially, from the subjective reading of chest CT scans. Patient anxiety also stems from the spirometry test. The author of this paper has firsthand experience with asthma and dreading the spirometry test. For asthmatics and patients with pulmonary fibrosis, the spirometry test is an exceedingly difficult and invasive test.

Many other papers have been written surrounding computer aided diagnosis (CAD). This paper aims to build on the existing CAD. Furthermore, this paper plans to address the subjective chest CT (Computer Tomography) scan reading by systematically evaluate the projected lung capacity percentage obtained by the spirometry test by evaluating the patient CT scan image sets.

This research project examined the CT scans of 172 patients, all with varying degrees of pulmonary fibrosis. The images were processed to identify the lung tissue. Once the lung tissue was identified and normalized across the patient dataset, machine learning, and deep learning models were employed to predict the lung capacity percentage that doctors use in their diagnosis and treatment plans. This research project provides an opportunity to improve patient care by reducing the need for spirometry testing.

II. Background and Literature Review

A. Introduction to Pulmonary Fibrosis

Idiopathic pulmonary fibrosis is a worldwide problem. North America and Europe see 3 to 9 cases per 100,000. The United States is reported higher than even the North American results with one study showing 494 cases per 100,000. That equates to ~1,600,000 cases of IPF. [2]

Matthew Zielinski [3] and Abehsera, et al [4] explain that the presentation of pulmonary fibrosis within a CT scan has the appearance of a honeycomb shape. Figure 1 shows the formation of honeycombing within a Fpulmonary fibrosis patient.

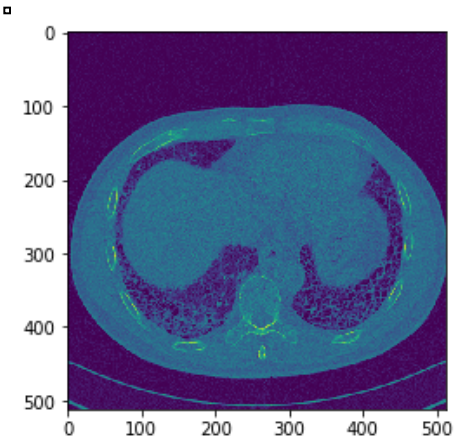

Figure 1: Image of patient with pulmonary fibrosis present in lower posterior lobes [3].

Zielinski explains further that the honeycombing seen in pulmonary fibrosis patient typically presents in the lower posterior lobes of the lungs.

B. Forced Vital Capacity (FVC)

The forced vital capacity (FVC) is the total amount of air a patient's lungs can expel after inhaling as deeply as possible. [5] The forced vital capacity is a lung function text that is

measure by a spirometry unity. It is used to diagnose obstructive lung diseases such as asthma, chronic obstructive pulmonary disease (COPD), and pulmonary fibrosis.

C. DICOM File format

Roni [6] explains what DICOM is by saying, "DICOM is a software integration standard that is used in Medical Imaging. All modern medical imaging systems (AKA Imaging Modalities) like X-Rays, Ultrasounds, CT (Computed Tomography), and MRI (Magnetic Resonance Imaging) support DICOM and use it extensively." There are two core features of the DICOM standard, the file format and the networking protocol. This paper will focus solely on the file format and will not delve into the networking aspect of DICOM. With that said, all references to DICOM files will be assumed to reside on the local or removable disk drive of the operating computer. Section A.3 of the DICOM standard [7] defines the Information Object Definition (IOD) of a DICOM CT image. There are six Information Entities (IE) within a DICOM CT image file. These IEs are: Patient, Study, Series, Frame of Reference, Equipment, and Image. We can further combine the IEs into two categories, Metadata attributes (contains IEs: Patient, Study, Series, Frame of Reference, and Equipment) and Image attributes (contains IE: Image).

D. Hounsfield Units

The unit of measurement in CT scans is the (HU), which is a measure of radiodensity. CT scanners are carefully calibrated to accurately measure this. [8] Dr. Kazerooni and Dr. Gross explain that the normal Hounsfield unit range for normal lung tissue is -700 to -500. [9, p. 379] This means that, after the pixel array within the DICOM file is adjusted with the Rescale Intercept (0028, 1052) and the Rescale Slope (0028, 1052), any pixel value falling within the -700 to -500 can be tagged as lung tissue.

E. Image Processing

Since CT images are strictly monochrome, image processing research was concentrated on monochrome applications. Intensity Transformations, Contrast Stretching, and Thresholding are all image processing techniques used for evaluation and image preprocessing.

F. Object Detection

Object detection of medical images started as early as the 1970s. These object detection routines were merely if-then-else statements of a rudimentary nature. By the mid-1990s, experts started to understand the need to implement deep learning networks to the problem of object detection. [10] Krizhevsky, et al [11], were one of the first groups to successfully develop a deep neural network and test it against a properly annotated image dataset. The deep learning model class Krizhevsky, et al, used was the convolutional neural network (CNN). Sumit Saha [12], does pose the question of, "[with] an image [as] nothing but a matrix of pixel values…why not flatten the image and feed it to a multi-level perceptron for classification purposes". Sumit continues by explaining the issue with multi-level perceptron used in image classification is that most images are too complex for a standard perceptron neural network to be effective. Thus, complex deep neural networks are needed for image object detection.

G. Existing CAD Software Applications with Medical Imaging

There are many existing computer-aided diagnoses (CAD) applications within the medical field. One such application is by Christe, et al [13]. The authors developed the INTACT system for idiopathic pulmonary fibrosis (IPF) on high resolution CT images (HRCT). This system consisted of 3 stages: lung anatomy segmentation, lung tissue characterization, and diagnosis. Lung anatomy segmentation utilized the following steps: extraction of large airways, segmentation of lung regions, separation of the left and right lungs, and morphological 3-dimensional smoothing. Tissue characterization was achieved by a convolutional neural network. For the final step, diagnosis, there was an additional sub-step that segmented the lungs into 12 regions each. These region's tissue was then run through another convolutional neural network. The authors also utilized two human radiologists as a test system against the INTACT system. The results of the INTACT system against the radiologists was a consistent tie.

H.J. Kim, et al [14], developed a computer-aided diagnosis system (simply called the CAD system) for providing quantitative lung fibrosis scores base on thin-section CT images. This team also utilized two expert radiologists for comparative performance measurement. This CAD system was broken down into 5 steps: images were denoised, images were grid sampled, the characteristics of grid intensities were converted to texture features, texter features classified pixels as fibrotic or non-fibrotic, and fibrotic pixels were reported as a percentage. Their results indicated that their CAD system could be useful for reproducible objective measurements of fibrosis in clinical trials.

[15] uses Homography estimations to trace objects between frames of a video capture. This process could also be adapted to CT scan analysis with the objects of interest being tracked throughout the patient's body.

[16] examines wavelet and curvelet transforms specifically in detecting human organs in CT scan images. Curvelet transforms have the ability to detect structures with curved attributes, however, pulmonary fibrosis presents with hexagonal shapes.

III. Methodology

Two methodologies were explored, Machine Learning and Image Processing. The machine learning methodology took a training/testing image per patient and built an unsupervised image detection model for evaluation. The image processing methodology filtered images using standard image filtering and object detection techniques to identify the honeycomb patterns.

A. Read Binary Files, Parse DICOM Attributes

The pydicom library was used to import and parse the DICOM attributes. The code utilizes a loop through the target directory to read in the DICOM files through the pydicom class library.

B. Ordering Image sets

The image slices need to be ordered based on the instance number represented by the x.InstanceNumber (0020,0013) of the DICOM file. This requirement is to allow for proper

iteration through the images sets to accelerate the identification of lung images.

Once the images are sorted by instance number, the middle image of the set is selected for preliminary evaluation. The reasoning behind selecting the middle image is that it has the best chance of containing lung tissue in a blind selection given the constraint that (0018, 0015) Body Part Examined equals 'Chest'.

C. Rescale Slope and Intercept Adjustment

The images were then adjusted base on the Rescale Intercept (0028, 1052) and the Rescale Slope (0028, 1053) attributes with equation (1). Figure 10 and Figure 11 are a comparison of the same image before and after the rescale transformation. Although the two images look identical, comparing Figure 2 and Figure 3 show that the array values have been adjusted. This likeness in the image figures is due to the python plotting library normalizing the images to a specific cmap.

$$U = m * SV + b \tag{1}$$

where, **U** is the output unit (scaled pixel value), **m** is the Rescale Slope (0028, 1052), **SV** is the stored value (value on disk), and **b** is the rescale intercept (0028, 1053).

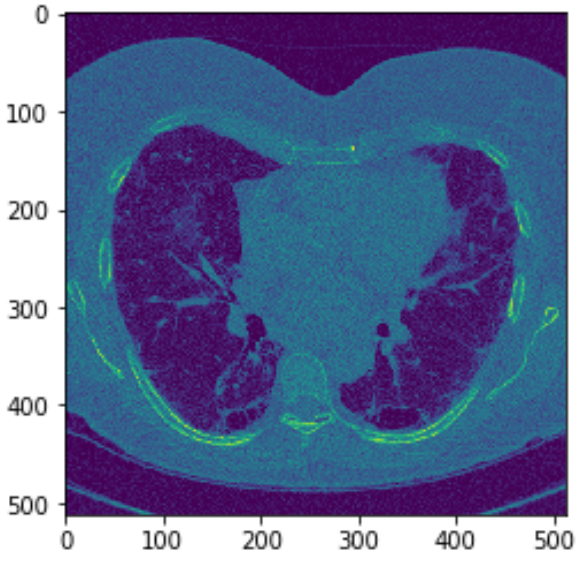

Figure 2: Original DICOM Image

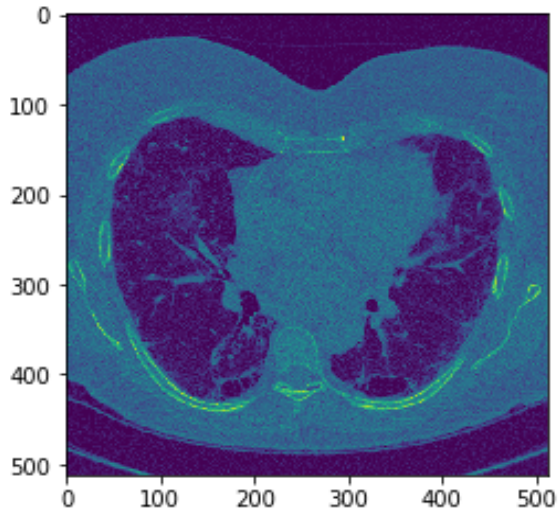

Figure 3: Rescale Slope and Interupt Adjusted DICOM Image

D. Hounsfield Unit adjustment

As stated previously, the Hounsfield Units for lung tissue are between -700 and -500. Figure 4 shows the image in Figure 3 with the Hounsfield Unit filter applied. This HU adjustment significantly accentuates the lung region of the CT scan image. This Hounsfield Unit filter produces a binary image file.

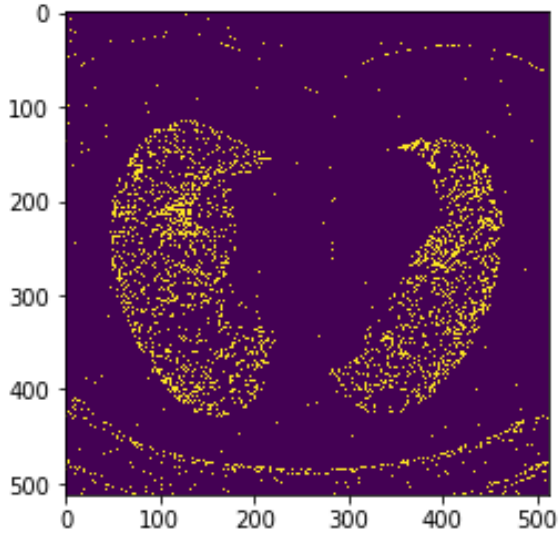

Figure 4: Image filtered with HU

E. Calculate Training Images

In order to train the convolutional neural network (CNN), a database of images was needed. This database consisted of one large image per patient containing three CT scans. This 'training' image was developed by concatenating the top (Patient Image Array * 0.25), middle (Patient Image Array * 0.5), and bottom (Patient Image Array * 0.75) binary lung images into one (1536 X 512) pixel image.

Figure 5 is an example of a training image for the CNN training and testing.

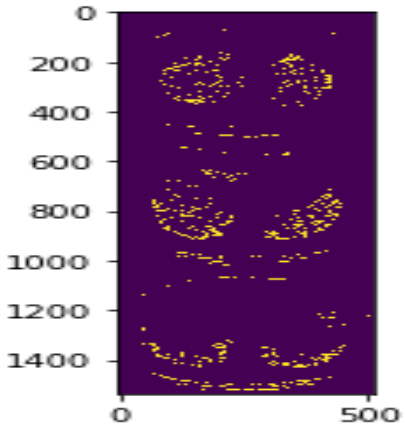

Figure 5: Training Image example

F. *Machine Learning*

Orange 3 was used to develop the image machine learning model. 172 images, in 67 categories, were used as the training and testing data. Both Logistic Regression and Neural Network models were used in the training. The target classification was set to the patient's lung output percentage rounded down to the nearest integer. This produced a category range from 47% to 146%, in increments of 1%.

G. *Image Filtering*

Image filtering was also tested against the CT image datasets. Figure 6 shows the result of the OpenCV library with the OTSU threshold filter applied to a standard CT image.

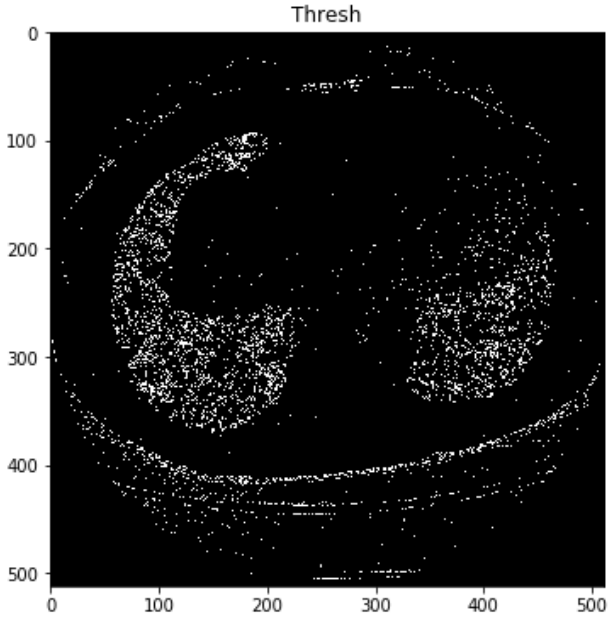

Figure 6: openCV OTSU thresholding result

The OTSU threshold image was then filtered using the OpenCV dilate and erode functions. Figure 7 shows the results of the OpenCV dilate and erode functions. The image in Figure 7 was first dilated with 5 iterations of a (2x2) ones' kernel. That result was then eroded by a (4x4) ones' kernel with an iteration set to 3. Finally, the image was dilated again using a (4x4) ones' kernel and 5 iterations. This sequence produced a binary mask that was combined with the original file. The OpenCV 'findcontours' with closed shape function was ran on the masked image. The resulting list of contours was evaluated based on the number of sides present in the contour. Any contour that had greater than 5 sides was kept for further evaluation. The diameter of the remaining contours was calculated and only contours with a diameter of 2 mm to 50 mm were counted.

The honeycomb counts obtained from the filtered image set were summed together and divided by the number of images in the set, as explained in equation **Error! Reference source not found.**).

$$HoneyCombRating = \frac{\sum_{0}^{Num\ of\ Images} HoneyCom\ per\ image}{Num\ of\ Images} \quad (2)$$

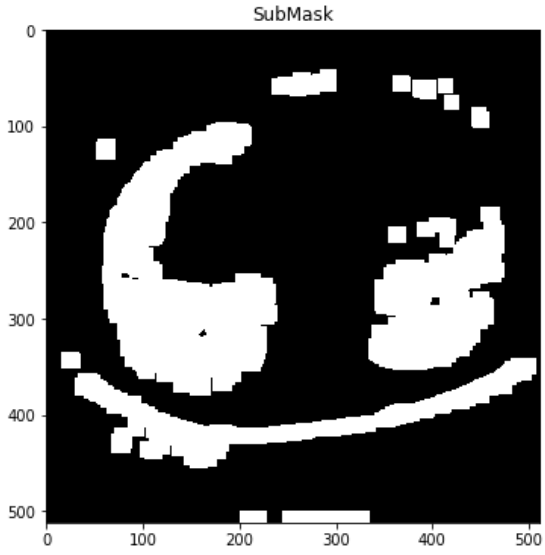

Figure 7: OpneCV dilate and erode image result

Figure 8 shows the results of the contours filtering described above. The red crosses indicate the location of a verified object of interest (i.e., honeycomb structure).

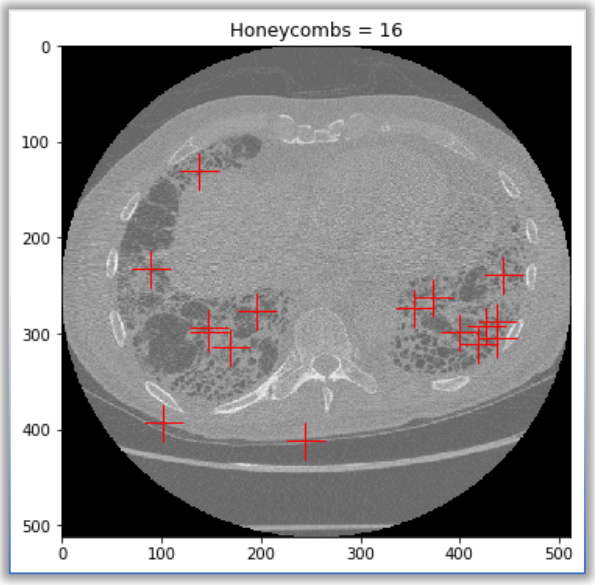

Figure 8: OpenCV contour detection result

IV. Results

The results for the different algorithm carried out in this paper are presented in the following subsections.

A. Logistic Regression

The machine learning results were less than desired. Several different models and settings were used, Logistic Regression and Neural Networks. Image Embedding was also utilized within the Orange 3 software. The patient image dataset was split into 121 training images and 51 test images. Table I shows the results of the 1% Incremental Category testing.

TABLE I. TABLE TYPE STYLES

1% Increments		
Model	***Correct Classification***	***Incorrect Classification***
Logistic Regression	1	50
Nueral Netword	2	49

With the poor performance of the 1% Incremental Categories, the training and testing images were then reconfigured into 10% Incremental Categories. Table II shows the results of these tests.

TABLE II. TABLE TYPE STYLES

10% Increments		
Model	***Correct Classification***	***Incorrect Classification***
Logistic Regression	10	41
Nueral Netword	7	44

B. Image Filtering

The Honeycomb Rating calculated in equation **Error! Reference source not found.** was compared to the patient's lung output percentage. Figure 9 shows what an ideal and statistically significant honeycomb rating vs. lung output percentage would look like. Patients with low honeycomb ratings should have high lung output percentage and patients with high honeycomb ratings should have low lung output percentages. This is strictly for comparison and are not actual results.

Figure 10 shows the actual honeycomb rating vs. lung output percentage. These results did not match with the hypothesis presented in the ideal plot (Figure 9). Both patients with high and low honeycomb ratings had low lung output percentages.

V. Conclusions

The machine learning and image processes models performed less than expected. The authors of this paper believe that the main cause of this low performance is the misrepresentation of the lung tissue by the Hounsfield Unit technique. Although scientific research has stated that the Hounsfield Unit of -700 to -500 shows lung tissue, it was demonstrated in the image processing activities of this project that this simple is not the case in all instances. When the Hounsfield Unit filter was applied, some CT images were clearly defining the lung tissue in the binary file, but others resolved to a field of random pixel data.

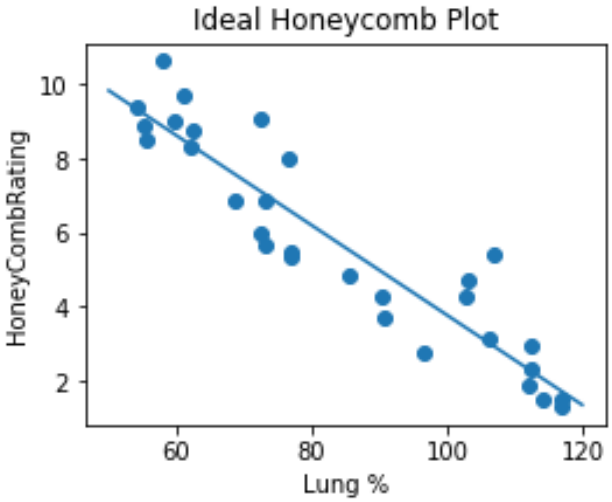

Figure 9: Ideal Honeycomb vs. Lung output percentage plot

Figure 10: Actual honeycomb rating vs. lung output percentage

Alternative image processing activities should be considered. Edge detection with centroid clustering could possibly identify the lung tissue more accurately and, therefore, provide the machine learning and deep neural network models with more accurate image data.

Another issue stems from the variable size of the patient CT Scan images. Both in image resolution and image count. Standardizing the images of various resolutions and counts indubitably cause data loss. The opportunity exists to convert the lung tissue found by proper image processing techniques to build a true-to-life 3D representation of the patient lungs.

The results of the image filtering with honeycomb rating leads the authors to believe that spirometry testing lacks the accuracy to properly determine the future progression of pulmonary fibrosis. In-depth studies need to be executed regarding the reliability and accuracy of standard spirometry tests.

Giving the medical community the ability to determine FVC (without the need for spirometry measurement) and identifying degradation in a patients with pulmonary fibrosis is a worthwhile endeavor and should be continued.

REFERENCES

[1] Pulmonary Fibrosis Foundation, "Pulmonary Fibrosis Overview," Pulmonary FIbrosis Foundation, [Online]. Available: https://www.pulmonaryfibrosis.org/life-with-pf/about-pf. [Accessed 7 September 2020].

[2] Pypi.org, "MedPy 0.4.0," Pypi.org, 14 Febrary 2019. [Online]. Available: https://pypi.org/project/MedPy/. [Accessed 05 September 2020].

[3] M. Zielinski MD, Interviewee, Medical Aspects of Pulmonary Fibrosis. [Interview]. 4 September 2020.

[4] M. Abehsera, D. Valeryre, P. Grenier, H. jailet, J. P. Battesti and M. W. Brauner, "Sarcidosis with pulmonary fibrosis: CT patterns and correlation with pulmonary function," American Journal of Roentgenology, vol. 174, no. 6, pp. 1751-1757, 2000.

[5] Univeristy of Michigan, Michigan Medicine, "Forced Expiratory Volume and Forced Vital Capacity," 9 June 2019. [Online]. Available: https://www.uofmhealth.org/health-library/aa73564#:~:text=Forced%20vital%20capacity%20(FVC)%20is,important%20measurement%20of%20lung%20function..

[6] R. "Introduction to DICOM - Chapter 1 - Introduction," 11 October 2011. [Online]. Available: http://dicomiseasy.blogspot.com/2011/10/introduction-to-dicom-chapter-1.html.

[7] dicom.nema.org, "DICOM PS3.3 2020d - Information Object Definitions," 2020. [Online]. Available: http://dicom.nema.org/medical/dicom/current/output/chtml/part03/sect_C.7.6.2.html#sect_C.7.6.2.1.1.

[8] G. Zuidhof, "Full Preprocessing Tutorial," 2017. [Online]. Available: https://www.kaggle.com/gzuidhof/full-preprocessing-tutorial/notebook.

[9] E. A. Kazerooni and B. H. Gross, Coardiopulmonary Imaging, Philadelphia: Lippincott Williams & Wilkins, 2004.

[10] G. Litjens, T. Kooi, B. Ehteshami, A. A. A. Setio, F. Ciompi, M. Ghafoorian, A. van der Laak, B. van Ginneken and C. I. Sanchez, "A survey on deep learning in medical image analysis," Medical image Analysis, pp. 60-88, 2017.

[11] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet calssification with deep convolutional neural networks," Advances in neural information processing, pp. 1097-1105, 2012.

[12] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks - the ELI5 way," 15 December 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[13] A. Christe, A. A. Peters, D. Drakopoulos, J. T. Heverhagen, T. Geiser, T. Stathopoulou, S. Christodoulidis, M. Anthimopoulos, S. G. Mougiakokou and L. Ebner, "Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images," Investigative Radiology, p. 627, 2019.

[14] H. Kim, D. Tashkin, P. Clements, G. Li, M. Brown, R. Elashoff, D. Gjertson, F. Abtin, D. Lynch, D. Strollo and J. Goldin, "A Computer-aided Diagnosis System for Quantitative Scoring of Extent of Lung Fibrosis in Scleroderma Patients," Clin Exp Rheumatol, vol. 28, no. 62, pp. 26-35, 2010.

[15] H. Razaee, A. Aghagolzadeh, M. H. Seyedarabi and S. Al Zu'bi, "Tracking and Occlusion Handling in Multi-sensor Networks by Particle Filter," *IEEE GCC,* February 2011

[16] S. AlZu'bi, S. Sharif, N. Islam and M. F. Abbod, "Multi-resolution analysis using curvelet and wavelet transforms for medical imaging," IEEE International Symposium on Medical Measurements and Applications, pp. 188-191, May 2011.