UNIVERSITY OF TWENTE

MASTER THESIS

Artificial Intelligence Driven Assessment of Asbestos Exposed Patients

Author
K.B.W GROOT LIPMAN, BSc

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

TECHNICAL MEDICINE

October 20, 2020

Abstract

Even though asbestos has been banned for a long time, patients are still presenting with asbestos-related diseases due to the long incubation time. These asbestos-related diseases, such as asbestosis and pleural plaques, can be numerous and hard to identify and quantify. Moreover, diagnosing the patients requires more time and attention from clinicians due to a financial compensation provided by the Dutch Institution for Asbestos Victims. To assist them in quantifying these diseases, we developed three AI models to assess asbestos-exposed patients. First, an AI model detected the morphological lung anomalies, where the model returns an anomaly heatmap. The results suggest that these methods can be employed to detect large morphological anomalies in the lungs and could provide further insights for the clinical and methodological research on asbestos exposure. Second, we developed an AI model for the classification of asbestosis in patients. Combining the model's classifications based on the CT scan and the corresponding DLCO values was superior to the other models and reached excellent diagnostic accuracy. The results suggest that the implementation of this model in the clinical setting could benefit the patient and clinicians in terms of reproducibility, consistency, and speed of the assessment of asbestosis. Because of the promising results backed by the pulmonologists, clinical validation of this AI model is currently ongoing. Third, we developed an AI model for the automatic segmentation of the pleural plaques in CT scans to estimate the volume. The predicted volume showed a high correlation to expert readers' segmentation, but overlapping measures were lacking. The AI model can be used to decrease the workload for the expert readers and to continue to expand the dataset to get a larger sample size. Moreover, we tested the relation between the lung function and the pleural plaque volume, which suggests that patients with a higher volume of plaques have a worse lung function. Until now, the relation between pleural plaque volume and lung function has not been proven at this scale.

Graduation Committee

Chairman and Technological Supervisor University

Dr. ir. F. van der Heijden
Robotics and Mechatronics, University of Twente, Enschede

Medical Supervisors Institution

Dr. J.A. Burgers, MD PhD
Thoracic Oncology, The Netherlands Cancer Institute, Amsterdam

Dr. T.N. Boellaard, MD PhD
Radiology, The Netherlands Cancer Institute, Amsterdam

Z. Bodalal (Elkarghali), MD MSc
Radiology, The Netherlands Cancer Institute, Amsterdam

C.J. de Gooijer, MD MSc
Thoracic Oncology, The Netherlands Cancer Institute, Amsterdam

Technological Supervisor Institution

S. Trebeschi, MSc
Radiology, The Netherlands Cancer Institute, Amsterdam

Process Supervisor University

Dr. M. Groenier
Technical Medicine, University of Twente, Enschede

External Member University

B. Wermelink, MSc
Multi-Modality Medical Imaging Group, TechMed Centre, University of Twente, Enschede

Acknowledgements

My gratitude goes to Prof. Dr. Regina Beets-Tan and Prof. Dr. Paul Baas for opening up their departments for my internship. It was an honor to work with all the talented and hard-working people in your departments.

I want to thank Dr. ir. Ferdi van der Heijden for the technical supervision during the masters. Your courses got me excited about medical imaging analysis, and your ongoing supervision during my internships has been a great asset.

The guidance of Dr. Sjaak Burgers MD and Dr. Thierry Boellaard MD has enabled me to develop myself in the clinical environment and reason with the clinical relevance in mind. Your interest in the projects led to constructive discussions that kept the focus sharp every time, while allowing me to develop the projects freely and in my own way, for which I am very grateful.

I am grateful for the guidance and advice of Dr. Marleen Groenier on my personal development. Your on-point questions always challenged me to think deeper about my behaviour and how I can change for the better.

Daily supervision was provided by Stefano Trebeschi MSc, Zuhir Bodalal MD MSc, and Dianne de Gooijer MD MSc. During this year, your supervision has been excellent; every time I encounter a problem, you were there to help me resolve it. You have been putting in lots of hours over the past year in discussions with me, which enabled me to think of a new solution to a given problem, for which I cannot thank you enough.

My sincere thanks go to all the radiologists who have been working on the project. Especially considering that covid-19 slowed the process, the amount of work you performed is astonishing.

This year would not have been as delighted if it was not for my fellow peers in the O-building. It was a pleasure to have all of you around, and the amount of joy you brought each day to work put a smile on my face.

To all Antoni van Leeuwenhoek hospital / Netherlands Cancer Institute employees, thank you for being welcoming and willing to teach me the tricks of the trade. It did not matter whether I was joining the radiology department, thorax oncology, or the operation rooms, everybody was cheerful and a delight to work with.

Finally, I want to thank my parents and Kaylee van Duren for the immense support over the years. Even when I made decisions you would not have made, you always stood by me, which enabled me to bring a positive attitude from home towards work.

Kevin Groot Lipman

Amsterdam

October 20, 2020

Contents

List of Figures

List of Tables

List of Abbreviations

1D	**1 D**imensional
2D	**2 D**imensional
3D	**3 D**imensional
AI	**A**rtificial **I**ntelligence
AMA	**A**merican **M**edical **A**ssociation
AUC	**A**rea **U**nder the **C**urve
CE	**C**ross **E**ntropy
CI	**C**onfidence **I**nterval (95%)
CNN	**C**onvolutional **N**eural **N**etwork
CPU	**C**entral **P**rocessing **U**nit
CT	**C**omputed **T**omography
DLCO	**D**iffusing **L**ung Capacity for **C**arbon Mono **O**xide
DSC	**D**ice **C**oefficient **S**core
FVC	**F**orced **V**ital **C**apacity
GPU	**G**raphics **P**rocessing **U**nit
HRCT	**H**igh **R**esolution **C**omputed **T**omography
HU	**H**oundsfield **U**nits
IAS	**I**nstituut **A**sbest **S**lachtoffers
ICC	**I**ntraclass **C**orrelation **C**oefficient
ILO	**I**nternational **L**abour **O**rganization
KCO	Transfer Coefficient for **C**arbon Mono **O**xide
KL	**K**ullback **L**eibler
MRI	**M**agnetic **R**esonance **I**maging
MSE	**M**ean **S**quared **E**rror
NKI	**N**ederlands **K**anker **I**nstituut
PPV	**P**leural **P**laque **V**olume
ReLU	**Re**ctified **L**inear **U**nit
RGB	**R**ed **G**reen **B**lue
ROC	**R**eceiver **O**perating **C**haracteristic
VAE	**V**ariational **A**uto **E**ncoder
VC	**V**ital **C**apacity
VO2	Maximum **O**xygen Uptake

1

General Background

1.1 Introduction

Asbestos-related lung diseases arise as a consequence of exposure to asbestos fibers, to which approximately 125 million people are exposed worldwide[1]. These are heat resistant fibrous silicate minerals extensively used in the twentieth century for manufacturing, mining, and construction[1]. Asbestos is morphologically divided into two groups: the straight, rigid amphiboles, and the curvy flexible serpentines, each accounting for 10% and 90% of the total use, respectively. Upon inhalation, asbestos fibers enter the respiratory tracts and, some of them, deposit in the lower airways and alveoli. There, they can induce benign (asbestosis, pleural plaques) and malignant (mesothelioma, lung cancer) diseases[2–5]. This thesis will focus on asbestosis and pleural plaques (Figure 1.1). Asbestosis is a result of chronic scarring (fibrosis) of the lung due to the inhalation of asbestos fibers[6]. Pleural plaques are local areas of hyalinized collagen fibers and may vary in calcified or noncalcified form[2–8]. Currently, the exact mechanisms of the development of these diseases are not entirely understood[3,9].

1.2 Characteristics of Asbestos-Related Diseases

Studies have shown that the effects of asbestos inhalation are correlated to a number of factors, such as cumulative dose of exposure, time since exposure, and characteristics of the type of fiber that is inhaled[10]. Differences in dosage over time, for example, will spike different reactions from the immune system, where a high dose over a short period of time triggers mostly an acute neutrophilic reaction (i.e. immune cells commonly seen in acute inflammations), whereas a lower dose over a more extended period promotes a chronic alveolar macrophage response (i.e. immune cells commonly seen in chronic inflammations)[11]. The fiber size is especially important since it determines both the depth in the respiratory tract the fibers can reach and the ability for macrophages to clear them through phagocytose[11]. Only fibers <0.4 micrometers in diameter and <10 micrometers in length can reach the alveoli, where phagocytosis is limited by the size of the macrophages (14 micrometers to 21 micrometers usually)[11]. Depending on the type and size of the asbestos fibers, it can take weeks to years to clear these fibers through phagocytosis. During this process, the macrophages release growth factors resulting in collagen deposition, alveolar epithelial cell

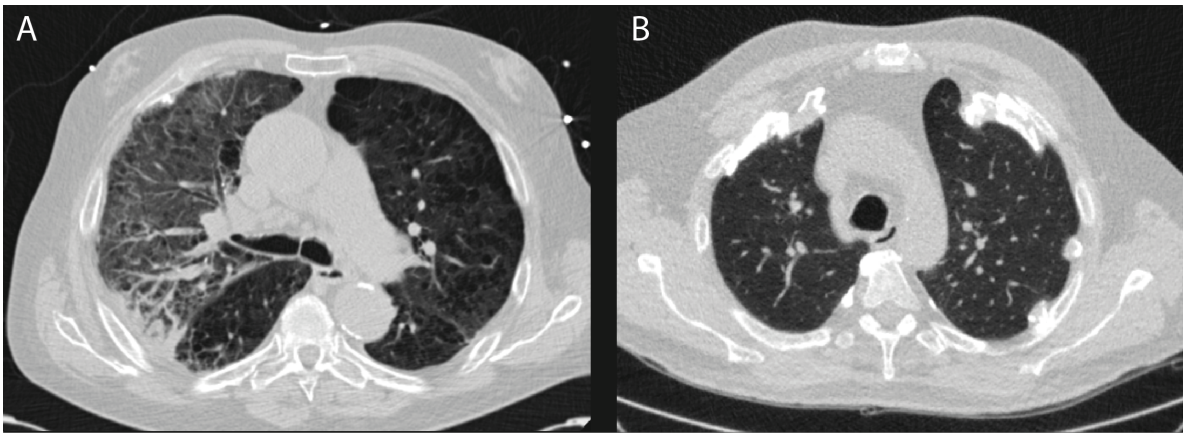

FIGURE 1.1: Visualization of asbestos-related diseases. (A) Patient with fibrosis and asbestosis. (B) Patient with (calcified) pleural plaques.

damage, and fibroblast proliferation[11]. Our body reacts by activating an inflammatory response to the damage of the epithelial cells, marking one of the early phase characteristics in the pathogenesis of asbestosis[11]. This leads to diffuse foci of fibrosis in the bronchial walls and alveolar ducts, which can include asbestos bodies of iron-rich proteins encapsulated in the fiber-ingested macrophages. Asbestos bodies can be either histologically assessed or by bronchoalveolar lavage fluid inspection and are suggestive of significant asbestos exposure. These asbestos bodies are diagnostic of asbestosis when the concentration of asbestosis bodies reaches 1 per milliliter.

However, a biopsy to confirm the presence of asbestos bodies is invasive and not necessary to diagnose asbestosis if the following criteria are met: (1) sufficient asbestos exposure with substantial time since first exposure, usually more than 20 years, (2) abnormalities detected in chest images, such as fibrosis and typically pleural plaques and (3) reduced lung function[11]. Pleural plaques (depending on the extension of the disease in the pleura) can lead to a restrictive lung function[7], preventing the patient from breathing optimally. Pleural plaque extension can be measured by volume, and the volume is correlated to the cumulative asbestos exposure[6].

1.3 Challenges

Asbestos-related diseases can be numerous and hard to identify and quantify. The diagnosis of asbestosis would make a patient eligible for compensation in the Netherlands. However,

it is especially challenging to diagnose the patients with asbestosis in a noninvasive manner. As the current diagnostic workup involves a team of multiple experts and a set of different examinations, it leads not only to the longer diagnostic time for both radiologists and pulmonologists, but also high inter- and intraobserver variability[12]. This is especially the case for patients with limited visible disease, with published reports evidencing small pleural abnormalities are quite often missed in patients that were exposed to asbestos[12]. Lawmakers are currently pushing for alternative and/or supplementary diagnostic methods to process the applications of patients fairly and timely, at reasonable costs. Currently, only patients that are diagnosed with asbestosis are eligible for compensation, unlike patients with only pleural plaques.

1.4 Current Assessment

In this context, imaging offers a quick, cheap, and reproducible way for assessing asbestos-related pulmonary diseases. The International Labour Organisation (ILO) released the International Classification of Radiographs of Pneumoconioses guidelines, currently considered the gold standard to quantify these diseases in chest X-rays[13]. Pneumoconiosis is the term used for all interstitial lung diseases caused by the dust inhalation, of which asbestosis is one. The ILO score is constructed through the reader's assessment of (1) radiographic quality, (2) pleural abnormalities concerning location, calcification size and extent, and (3) parenchymal abnormalities. Parenchymal abnormalities can be further classified by small or large opacity categorization based on shape and size, the lung zones where they are located, and the opacities' profusion. Following the ILO score, the radiologist or pulmonologist can describe asbestosis related diseases systematically and in a reproducible manner. Radiologists and pulmonologists use a derivative of the ILO score to assess the asbestos-related anomalies systematically. This score is still time-consuming and fails to represent the 3D structure due to the chest X-rays' 2D nature. 3D Computed Tomography (CT) would be preferred over 2D chest X-rays, providing better sensitivity and specificity for detecting asbestos-related diseases[13]. The High-Resolution CT (HRCT) is currently the preferred image modality to visualize asbestos-related diseases. This is a CT acquisition technique where

a reconstruction algorithm post-processes the chest's obtained thin-slices with high-spatial-frequencies, yielding better resolution[14]. Within these CT-scans, we can perform a volumetric measure on pleural plaques. Finding a correlation between the volume of these plaques and reduced lung function parameters could explain symptoms of patients and change the application process to reward compensation for pleural plaque volume. This improvement can be achieved by determining the critical volume correlated sufficiently to a significant lung function reduction. By automatic volumetry of these plaques, we could provide a precise quantitate method to determine whether the patient meets a criterion, unlike the 5% fibrotic surface.

1.5 Proposed Solution

Automatic solutions based on artificial intelligence (AI) have the potential to quantify lung anomalies and discern the ones connected to relevant asbestos exposure, namely volumetry of pleural plaques, and detection of asbestosis. This would simultaneously provide a standardized and fully automatic solution, allowing to reduce intra- and interobserver variability[15]. AI aims to mimic cognitive, labor-intensive tasks via complex computational models trained on top of existing datasets. Specifically, in the field of biomedical imaging, convolutional neural networks (CNN) approaches have been proven incredibly successful for their ability to process imaging data with different levels of abstractions, and automatically learn imaging features from data. These properties enable us to navigate and explore massive datasets and discover complex structures and patterns that can be used for prediction, segmentation, and classification. Current state-of-the-art technologies on this subject are based on convolutional neural networks. Through subsequent filtering operations which will down-sample the input size, the model squashes images down to a lower, highly informative dimensional space, where quantitative features are used for classification or regression tasks. For image segmentation, the procedure is complemented by a decoder, which by applying the inverse operation, mapping the low dimensional space to the full resolution image, yields the label map in this case.

1.6 Research Aim and Outline of Thesis

This thesis aims to improve the decision process to compensate patients that suffer from benign asbestos-related lung diseases, which account for 100 patients yearly in the Netherlands. Furthermore, we aim to quantify pleural plaque volumes and test for correlation with lung function parameters. This will be achieved in three-fold. First, we will develop a preliminary identification method for lung anomalies based on variational autoencoders. Second, an automatic classification algorithm will be trained to classify CT scans as positive or negative for compensation. Third, we implement a fully convolutional neural network to automatically segment the pleural plaques in the CT scan and perform volumetric measures on them, after which we test for correlation to the respective lung function parameters.

2

General Technological Background

2.1 Machine & Deep Learning

In machine learning, the aim is to construct an algorithm that automatically extracts patterns out of data and makes a prediction, classification or segmentation from these discovered patterns. Mathematically, given the random variables x and y, we aim to learn the function $f(x) = y$ through a series of predefined computational steps. There are many techniques used in machine learning, where neural networks are one of them. Neural networks take inputs of variable size (e.g. pixel values of one image). Each input value connects to every node in the next layer. In this connection, the value of that input is multiplied by a weight factor, after which a bias is added to it. This value is subsequently passed through an activation function that applies nonlinearity (i.e. make all negative values zero), which is necessary to model complex relations between the input (e.g. the image) and the output (e.g. classification of diseases). Deep Learning is the overarching term for neural networks with multiple layers stacked on top of each other, creating a deep network.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) is a Deep Learning method and is currently the state-of-the-art method for automatic imaging processing. The weights are shared in convolutional filters, enabling the network to retrieve features out of the image where pixels are locally correlated. More specifically, it takes an image as input and downsamples it through convolutional operations to a highly informative feature representation, which can be described as a smaller image with characteristics of the original image. For example, for an apple image, its feature representation could contain values that describe the color, shape, and size. For an image of a thorax CT scan, it could represent characteristics like the upper body's size, the smoothness of the pleura, or the shape of the rib cage. Depending on the task at hand, this feature representation can be complemented by a classification layer to predict what disease is located in the image, or to upsampling layers to produce a segmentation or reconstruction of the image. In this technological background, we discuss the different types of layers that could be implemented in a CNN, loss functions to quantify the performance of the CNN for a specific goal, the optimization techniques to reach that goal, and the 'learning part' of the CNN, which is called backpropagation.

2.3 Layers

2.3.1 Input Layer

The input layer defines the input shape of the network and is determined by the size of the images or hardware limitations. Specific for CNNs, all images in the dataset should have the same size during training, and have to be resized if they have different sizes. Due to the massive computational power required, limited by the available hardware, it is quite common to downsample medical images like CT or MRI scans. Besides the size of the image, the channels are defined as well. For grayscale images like CT, there is one channel needed, but for RGB images, each color needs a separate channel. Like RGB images, which are actually three stacked images to form 1 image, medical images of multiple corresponding scans can also be stacked in channels. This is especially common in MRI, where multiple sequences of the same imaged body structure can be present. Segmentations or attention maps can also be implemented in these channels, as long as they have the same size as the original image and the imaged structures do have the same location within the images.

2.3.2 Convolutional Layer

The convolutional layer consists of filters that slide over the image, where this technique is based on the assumption that pixels nearby each other relate more than distant pixels (or voxels in case of 3D images). These filters, also known as weights, usually consist of a matrix of 3x3 values, where 3x3 pixels of the input are element-wise multiplied by the values in the filter. These multiplied values are subsequently added and this sum will represent the degree of similarity between the filter and that part of the image. A bias can be implemented as well, but is often omitted due to the superiority of batch normalization, which will be further explained later. After this operation is performed for all 3x3 blocks in the image, a new filtered image is retrieved. The number of filters per convolutional layer is defined during the construction of the CNN. An increasing number of filters directly introduces more features the CNN can capture, but at the cost of computational expense and risk of overfitting due to unnecessary complexity of the CNN. Overfitting is the process of performing well on the images the CNN is trained on, but lacking performance for images it has never seen. The weights of all filters are learned, i.e. they are adjusted based on

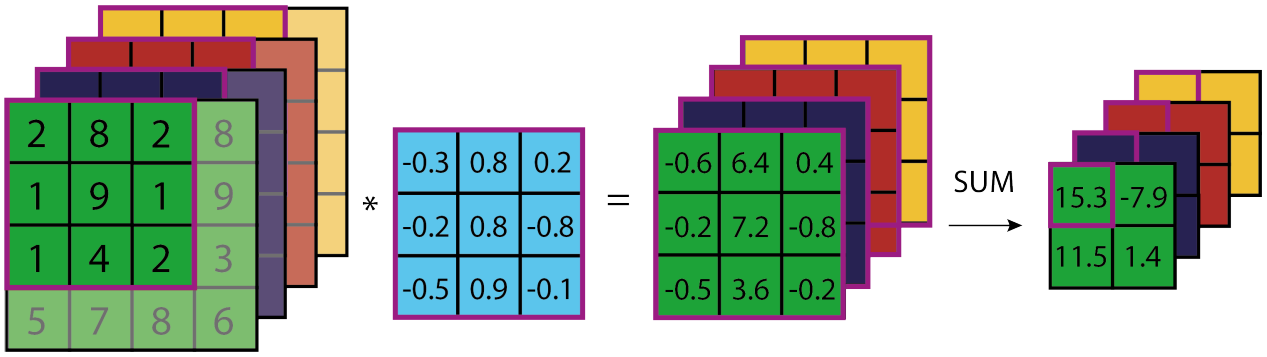

FIGURE 2.1: The Convolutional Layer visualized. The layer performs convolutions on the input, which contains four feature maps denoted in the different colors. The lightblue matrix contains the convolutional weights, which are learned. The field of view of the weights on the images is visualized by the purple outline. The 3x3 outputs are derived by element wise multiplication between the field of view in the image and the weights. Subsequently, the sum is taken of each 3x3 output, resulting in one value in the new feature map at the right. The remaining values in the right feature map are derived by moving the field of view on the original image to top-right, bottom-left, and bottom-right, respectively. Each feature map is convoluted with another set of weights.

the difference in what the model predicted in the image and what the label was. Since 3x3 blocks of convolutional filters fit x minus 2 times in the image with x the number of pixels in one direction, zero-padding is often applied. This operation surrounds the image with zeros before convolutions are applied to ensure that the dimensions are equal after filtering. Normally, convolutions have a stride of one, meaning that they move one pixel at a time after they perform a convolution. Increasing the stride to two would result in each filter moving two pixels before performing the next convolution. This will lead to fewer convolutions and, therefore, a reduction of the output of the convolutional layer by factor two. An overview of a convolutional layer is given in Figure 2.1

2.3.3 Activation Layer

After the convolutional layer filters the image, an activation function applies nonlinearity to the filtered image to enable CNN to learn complex relations. Complex relation are defined by interactions that cannot be solved by linear equations. Rectified Linear Unit (ReLU)[16] is one of the most used activation functions for applying nonlinearity after the convolutional layers, which makes all the negative values in the filtered image zero and keeps the positive value unaffected (Figure 2.2). Spin-offs of the concept of ReLU yield even more performance, which were LeakyReLU[17] that scales negative values by a fixed parameter, and PReLU[18]

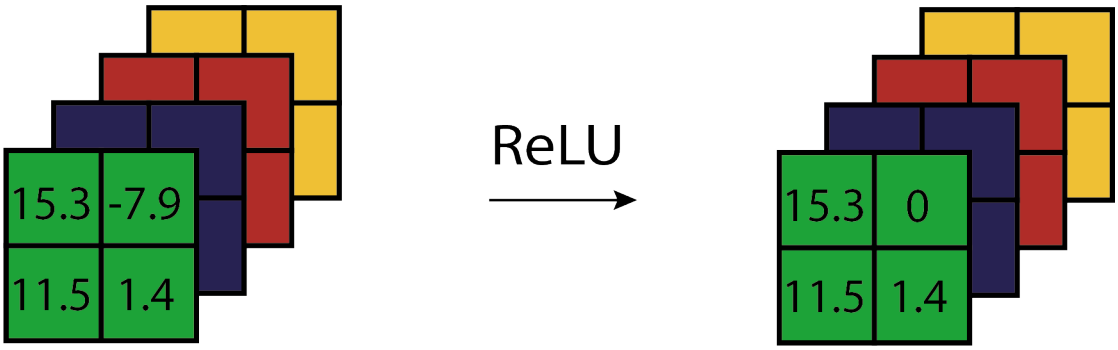

FIGURE 2.2: The Rectified Linear Unit converts all negative activation values to zero.

that scales the negative values by a learnable parameter alpha. For the final classification of images, the sigmoid function is often used. It scales all incoming values in the range [0, 1], which we interpret as the probability of a disease present in the image. If we have multiple options in our classification that are mutually exclusive, the sigmoid function will not be sufficient since the probabilities should be dependent on each other. An extension of the sigmoid function, the softmax function, is necessary to convert all incoming activations to a probability with the sum of 1. The sigmoid is implemented in the case of pixel-wise classification like segmentation since the segmented pixels are not dependent on each other.

2.3.4 Batch Normalization Layer

Batch normalization[19] is the process of standardizing the output of a layer. It subtracts the mean of all activations of each individual activation and divides it by the standard deviation of all activations, with two learnable parameters that orchestrate the whole process. In this way, it stabilizes the distribution of each layer, and activations will not reach extremely high or low values. The training process of a CNN is more likely to converge faster with batch normalization, which means it finds an optimal solution in less iterations. However, batch normalization requires sizable batch sizes during training to work, since the variance within the batch size should reflect the variances in the total dataset.

2.3.5 Max Pooling / Average Pooling Layer

Pooling operations are implemented to reduce the dimensionality of the input. Convolutional and activation layers filter the image and return the value of activation. Max pooling

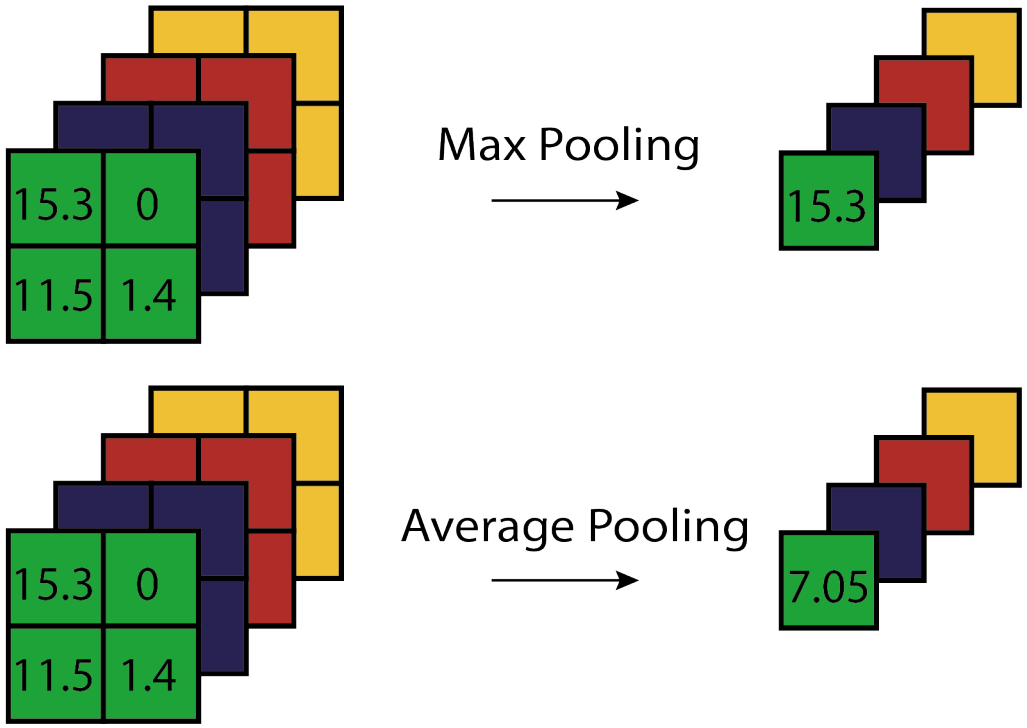

FIGURE 2.3: Pooling operations with 2x2 field of view visualized. Top: Max Pooling operation, which only retains the maximum value per feature map. Bottom: Average Pooling operation, which averages the input per feature map. Feature maps are color-coded.

layer subsequently takes the highest activation in blocks of commonly 2x2 and discards the rest of the activations. With these pooling operations, dimension reduction is achieved at the cost of 75% of the activation that is lost after each max pooling layer. When dimension reduction is no longer feasible, usually around the dimension of 3x3 - 7x7 for 2D images, average pooling is implemented for classification tasks. It returns the average value of the remaining activations, after which one vector is created of the size 1xN, where N is the number of features we define. Each filter in the last layer leads to one feature of the original image before connection to the fully connected layers. Both pooling operations are visualized in Figure 2.3. Nowadays, pooling operations are quite controversial, where the state-of-the-art networks seem to favor convolutional layers with stride 2 over the max pooling layer for dimension reduction.

2.3.6 Fully Connected Layer

The fully connected layers consist of nodes, where each node is connected to all nodes in the previous and next layer. Fully connected layers do not process any spatial information, i.e.

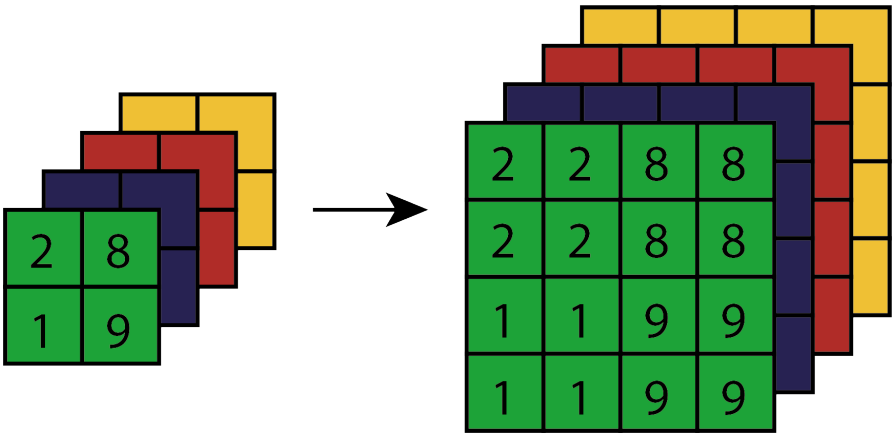

FIGURE 2.4: The Upsampling layer generates larger feature maps by copying the values of the input.

they have no prior knowledge about which pixels in an image are next to each other, unlike convolutional layers. In CNNs, they are used after the average pooling layer to process the values of all features and construct the output. A classification layer is most often a fully connected layer with the number of nodes equal to the number of categories to classify with a sigmoid or softmax activation function.

2.3.7 Upsample Layer

Upsample layers are used to enlarge the highly informative feature representation, which is necessary to generate a reconstruction or segmentation of the same size as the original image. It copies the value of the input it receives by the upsampling factor that is given (Figure 2.4). For a 2x2 upsample factor, an input of the value 3 would become a square of 2x2 consisting of all 3's. The workings of this layer are simple, and more sophisticated upsampling techniques are nowadays favoured over this layer.

2.3.8 SubPixelUpscaling Layer

The SubPixelUpscaling[20] layer is an extension of the normal upsample layer. Instead of copying the values multiple times, it uses the values of multiple feature maps and combines them in a new upsampled feature map (Figure 2.4). However, this will lead to a decrease in feature maps compared to the normal upsampling layer by the square of the scale factor.

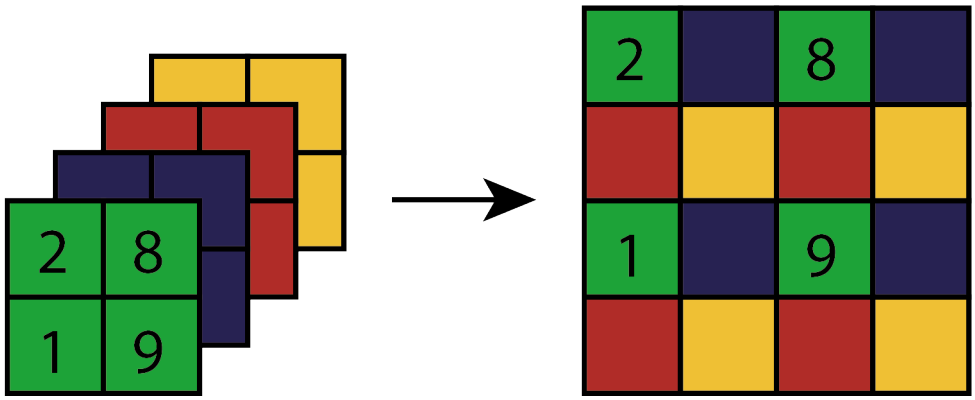

FIGURE 2.5: SubPixelUpscaling integrates multiple feature maps to generate an upsampled feature map.

2.3.9 Deconvolution Layer

The deconvolution layer can be implemented for more refined upsampling. Similar to the convolutional layer, the deconvolutional layer has convolutional filters that are learned to upsample the feature representation. The size of the filters depends on the upsampling factor, which is usually 2x2. The filters will go over one value in the feature representation at the same time and multiple this value by the learned filter weights.

2.4 Loss Function

CNNs are trained by changing the weights and the biases each time after an output is returned (e.g. a prediction about the disease in the image). The prediction of the CNN is compared to the label (the 'ground truth'). This allows us to quantify the error of the model with respect to a predefined error function. Using the derivative of the loss function, we can minimize the error in the process of optimization of the network. The label for a positive finding of disease is binarized to the value 1, the absence of the disease is defined as 0. The prediction of the model will be a value between 0 and 1, indicating the probability the model gives to the presence of the disease. The further the prediction is from the label, the more change in the weights will follow. The learning process, which is called backpropagation, is performed by calculating the gradient. The gradient is the loss, or difference in prediction and label, for all training examples in one batch. It is calculated through the chain rule, where the weights that have the most impact on the wrong prediction can be localized

and adjusted the most. One constraint for this concept is that there cannot be any loops in which the gradient is flowing. It flows from the output back to the input layer. To get the best update of the weights, all training examples should be predicted. However, this is computationally expensive and a slow method to approximate the best possible setup of weights. Therefore, training is performed in mini-batches of usually 8 to 64 examples at a time for CNNs, where the loss is defined as the difference of all predictions and their respective labels. The difference is called the loss and is calculated by the loss function. This function should be set according to the goal one wants to achieve with the training of a CNN model. Noteworthy, it should be differentiable in order to update the weights, which makes accuracy, for example, a non-optimizable function.

2.4.1 Cross Entropy

Cross entropy is commonly used in classification problems. Cross entropy is an unbounded loss function — i.e. the range is [0, ∞] instead of [0, 1] — which makes it vulnerable for instances where the images have the wrong label (e.g. the image does contain the disease but it is not labeled as such). This loss for binary classification is given by:

$$CE = -\sum_i (y'_i \log(y_i) + (1 - y'_i) \log(1 - y_i))$$

2.4.2 Dice Coefficient Score

Dice Coefficient Score (DSC) is often used with the segmentation of images. A common problem in segmentation is a class imbalance, which means that there are orders of magnitude more voxels unlabeled than labeled (i.e. pleural plaque segmentation). Cross entropy does not perform well at huge class imbalance problems. Moreover, a huge class imbalance can lead to an event where the CNN will predict all voxels in a CT scan as background, and it would achieve an accuracy of 99%+. Dice bypasses this problem by negating the true negative. It focuses only on true positives, false positives, and false negatives. The formula for the dice is given by:

$$DSC = \frac{2\sum_i^N y'_i y_i}{\sum_i^N y'_i + \sum_i^N y_i}$$

2.4.3 Mean Squared Error

The mean squared error is commonly used for reconstruction. It calculates the squared error between the original image and the reconstruction and returns the mean as loss, given by:

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - y_i')^2$$

2.5 Training Procedure

Training of the CNN can be initialized once the CNN is constructed through the described layers, and the loss function is defined. However, training on the entire dataset would be infeasible since we would not have any objective method to determine the model's performance. Therefore, we split the dataset into a training set, a validation set, and a test set. The CNN extracts features from the data in the training set and makes a prediction. Subsequently, the corresponding CT label is compared to the CNN's prediction, after which it adjusts the weights to get closer to the label next time. The maximum training time is defined by epochs, where an epoch is defined as the moment when the CNN has processed every CT scan in the training set once. This iterative process goes on, but can lead to overfitting, which means the CNN learns features of individual scans that do not generalize well to unseen CT scans. To counter this problem, the CNN has to process the CT scans in the validation set after each epoch, resulting in a validation loss. This loss indicates the current performance of the CNN during training. The model can not learn from the CT scans in the validation set, which means no weights are adjusted based on the score it reached. After training is completed, we retrieve the CNN's weights when it reached the maximum performance on the validation set, to extract the CNN that generalizes best. However, this introduces a bias where we pick the best version during training. Therefore, the CNN's performance is determined by predicting the CT scans in the independent test set, with CT scans the CNN has never processed before. When the performance of a CNN is described in this thesis, it will always be the result of predicting outcomes of CT scans in the test set.

3

Preliminary Identification of Lung Anomalies

3.1 Introduction

As asbestos fibers deposit deep into the tissues of the lungs, it causes a series of known lung diseases. The most common pathological manifestations of asbestos exposure include fibrosis, pleural plaques (occasionally calcified), and atelectasis[2–5]. While some are common and can be directly addressed by AI-diagnostic models, others are rarer and may require a more generic approach to capture and quantify anomalies in the lung parenchyma and pleura. To explore all possible variations present in the imaging dataset in relation to the exposure to asbestos, we turned our attention to AI-based anomaly detection methods.

One of the most commonly used AI models architectures for anomaly detection in imaging is deep convolutional variational autoencoders[21]. The goal of these networks is to learn salient key features in the original image and encode them in the latent space in a meaningful, semantic, quantitative manner, that can be used to reconstruct the original image. Unsupervised reconstruction loss between input and generated output is used for training.

Variational autoencoders (VAE) aim to model a training dataset by mapping its imaging features to a Gaussian distribution. This is done by adding constraints on the internal representation of the model[21]. This concept enables us to perform anomaly detection: samples that deviate from the normal training population will not be recognized by the model and labeled as abnormal. In practice, as the network reconstructs the abnormal image, unrecognized features will be ignored, and will stand out in the reconstruction when compared to the input. Another advantage of the VAE over traditional autoencoders is that we can sample any arbitrary point in the distribution of the latent space, and still get a credible output after decoding. By training the VAE on only images that have no pathology, i.e. healthy lungs, the VAE should learn how to encode only the features of normal tissues. As the network will not be able to reconstruct pathological manifestations of lung diseases, these will result in high reconstruction errors. These errors can be visually assessed, as well as quantified on a per slice or CT scan level.

This study aims to employ variational autoencoder networks for anomaly detection on CT scans of the lungs. We hypothesize that, when CT scans that containing visible pathological structure are processed, the algorithm will not be able to reconstruct them, therefore generating a reconstruction error at the location of the pathology.

3.2 Technological Background

Variational autoencoders are subtypes of autoencoder networks. Autoencoders map the input data to points to a lower-dimensional latent space representation, which contains information needed to reconstruct the original input sample. More specifically, autoencoders are composed of two separate components: an encoder $e(\cdot)$ that maps an input sample (e.g. image) x to an internal representation $e(x) \rightarrow h$ (latent space), and a decoder $d(\cdot)$ that uses that internal representation to reconstruct the original sample $d(h) \rightarrow x'$. The latent space h is commonly designed as a shorter 1D encoding of the original input samples, which contains enough information regarding the input for the decoder to be able to reconstruct it.

The size of the latent space determines the capability of the VAE to reconstruct the input image, among other factors. Small representations might lead to inadequate reconstructions of healthy structures, whereas too large representations might lead the VAE to reconstruct abnormal tissue using bits of healthy features, i.e. reconstruction of consolidations in the lung by using patches of heart-like tissue. Therefore, tuning of the network architecture is often required. The distribution of the latent space can take up any arbitrary form in standard autoencoders. In variational autoencoders, however, we further constrain the latent space to follow a predefined prior, often Gaussian. Approximating the latent space to follow a Gaussian distribution has two main advantages: (1) it provides a continuous latent space where we can sample from any point in the distribution of the latent space and generate data from it, and (2) it allows quantifying outlierness by simply measuring the deviation from the mean.

To enforce the internal representation to follow a predefined (Gaussian) distribution, the encoder is tasked to map the input sample to a distribution parameterized by the mean and standard deviation. The latent space does not consist of a 1D vector, but rather multiple vectors. This is implemented as a two-head vectorial output: $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ from which we sample the data distribution to compare against our prior. In practice, since sampling is not a differentiable operation, and backpropagation requires differentiability, a reparameterization trick has to be employed. This adds a sampling layer, where $\boldsymbol{\epsilon}$ is sampled from an independent Gaussian distribution $N(0, I)$. The output layer $\mathbf{z}$ is subsequently defined as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}$. By generating the independent sampling vector ϵ, we ensure the differentiability of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

and, therefore, the ability to backpropagate.

The loss function consists of two separate terms: the reconstruction loss, which ensures that the features learned in the latent space are a meaningful representation of the original sample, and the divergence loss, which ensures that the distribution of samples in the latent space follows the prior distribution. The reconstruction loss is given by the mean squared error between the original image and the reconstruction. The divergence loss is estimated through the Kullback Leibler (KL) divergence between the distribution parametrized by the mean and variance layers, and a prior Gaussian distribution *N(0, I)* (Equation 1). The combined loss results in a trade-off between informativeness and normality latent space.

$$D_{KL}[N(\mu_h, \Sigma_h)||N(0, I)] = \frac{1}{2}\sum^{B}\left(exp(\Sigma_h) + \mu_h^2 - 1 - \Sigma_h\right) \tag{3.1}$$

Where D_{KL} is the KL divergence, N the standard Gaussian distribution, μ the mean, h the latent space, Σ the variance. The formula has been written to calculate the exponent of the variance, instead of the natural logarithm, since the exponent is more numerically stable and easier to compute. The conversion of log variance to standard deviation is given by $\sigma = \sqrt{e^{\Sigma}}$.

One problem that can arise from VAE training is posterior collapse. A posterior collapse happens when the model minimizes the loss by minimizing the KL divergence to zero and returning an average image as reconstruction, independently of the input. To prevent posterior collapse (i.e. the collapse of the latent space to a constant value), weight annealing is used. Weight annealing prescribes the use of a weighting factor on the KL term, which starts small and increases during the training. Ideally, we want the variance of each mean large enough to acquire a continuous latent space, but the coefficient of variation should be low enough that different means cannot overlap too much after sampling.

3.3 Material and Methods

3.3.1 Datasets

To train a variational autoencoder network (VAE) to model healthy lung tissues, we collected a publicly-available CT dataset of lymphadenopathy patients[22]. This will be referred

to as the discovery set throughout the rest of the chapter. CT slices containing labeled enlarged lymph nodes were removed, since the dataset should only contain healthy CT slices. The study dataset consisted of patients with suspicion of asbestosis, collected at the Netherlands Cancer Institute (NKI; Amsterdam). To set a benchmark to which the reconstruction error in the study dataset can be compared, we collected a control dataset of patients and their CT scans of healthy lungs collected at the NKI.

The discovery set contained N=867 patients, corresponding to a total of N=205 519 CT scan slices. The study set contained a total of 523 patients. Patients were excluded due to the absence of CT scans (N=74), slice thickness >5 mm (N=10), and insufficient quality of the lung segmentation (N=16). This resulted in a total of 423 patients (and corresponding CT scans) from the study dataset. All 76 patients in the control set were included.

3.3.2 Data Curation

To mitigate differences of imaging protocols, all CT density histograms were clipped between -1024 and 3072 Hounsfield Units (HU) and scaled on the interval [0, 1]. Slices were also resampled to 256 x 256 due to hardware constraints. To focus the attention of the VAE on the lungs, we performed segmentation of the lungs, and we blackened the background region. The segmentation was performed using a publicly-available deep learning segmentation network by Rodney et al[23]. Lung segmentations were dilated through morphological operators with a kernel of 20 x 20 x 5 voxels to include adjacent tissue (i.e. thoracic wall) where pleural plaques are commonly found.

3.3.3 Network Design

The proposed network design follows the standard architecture of the variational autoencoder[21], where encoder $e(\cdot)$, latent space h, and decoder $d(\cdot)$ are placed in subsequent order. The encoder is composed of 6 convolutional blocks. Blocks are composed of repeated layers of convolutions, batch normalization, and the LeakyReLU activation function[17]. Downsampling is implemented through striding. The first block starts with 16 filters. Each subsequent block adds 16 filters. The decoder is composed of the mirrored architecture of the encoder, where the convolutional layer with stride 2 is replaced with a convolutional layer with a single stride and a subpixel upscaling layer[20] at the end of the convolutional block.

Sigmoid is used on the last layer of the reconstruction to constrain the image on the interval [0, 1]. While there has been some advancement in the architecture, most notably the usage of fully-convolutional layers in the latent space for medical image reconstruction[24], we kept fully connected nodes in the latent representation. This might seem disadvantageous to spatial representations, but through internal experiments, we observed that the fully connected architecture prevents the VAE from reconstructing anomalies with patches and features learned from healthy tissue. Values in the latent space are reshaped to a 4x4x96 format and passed forward to the decoder part. The decoder upsamples this vector through convolutional layers and subpixel upscaling to reconstruct the full-size image. An overview of the network design is shown in Figure 3.1.

3.4 Experiment

3.4.1 Network Implementation

The VAE network was designed and trained using Tensorflow (v1.15.0) and Keras (v2.3.1) libraries on an NVIDIA GeForce RTX 2080Ti. N=195 519 slices were assigned to the training set and N=10 000 to the validation set for monitoring the training process. The batch size was set to 48, and Adam was used as optimizer, with an initial learning rate of 1.5e-3. The VAE was trained for 200 epochs, where the weight of the KL term in the loss was increased by 0.05 after each epoch, reaching a maximum value of 1.0 in total. Best model checkpoint at the end of every epoch was performed. Data augmentation with rotation (up to 20°) and horizontal flipping of the image was implemented during training.

3.4.2 Analysis

After training, the model was used to reconstruct slices from CT scans of the study dataset. The squared error between the reconstructed image and the original image was derived. The assessment was performed both visually and quantitatively. By comparing the squared errors of the asbestosis study dataset and the control dataset, we gained insight into the distribution of the anomalies in the study dataset. A student t-test was performed to determine the difference between the reconstruction error of the two datasets.

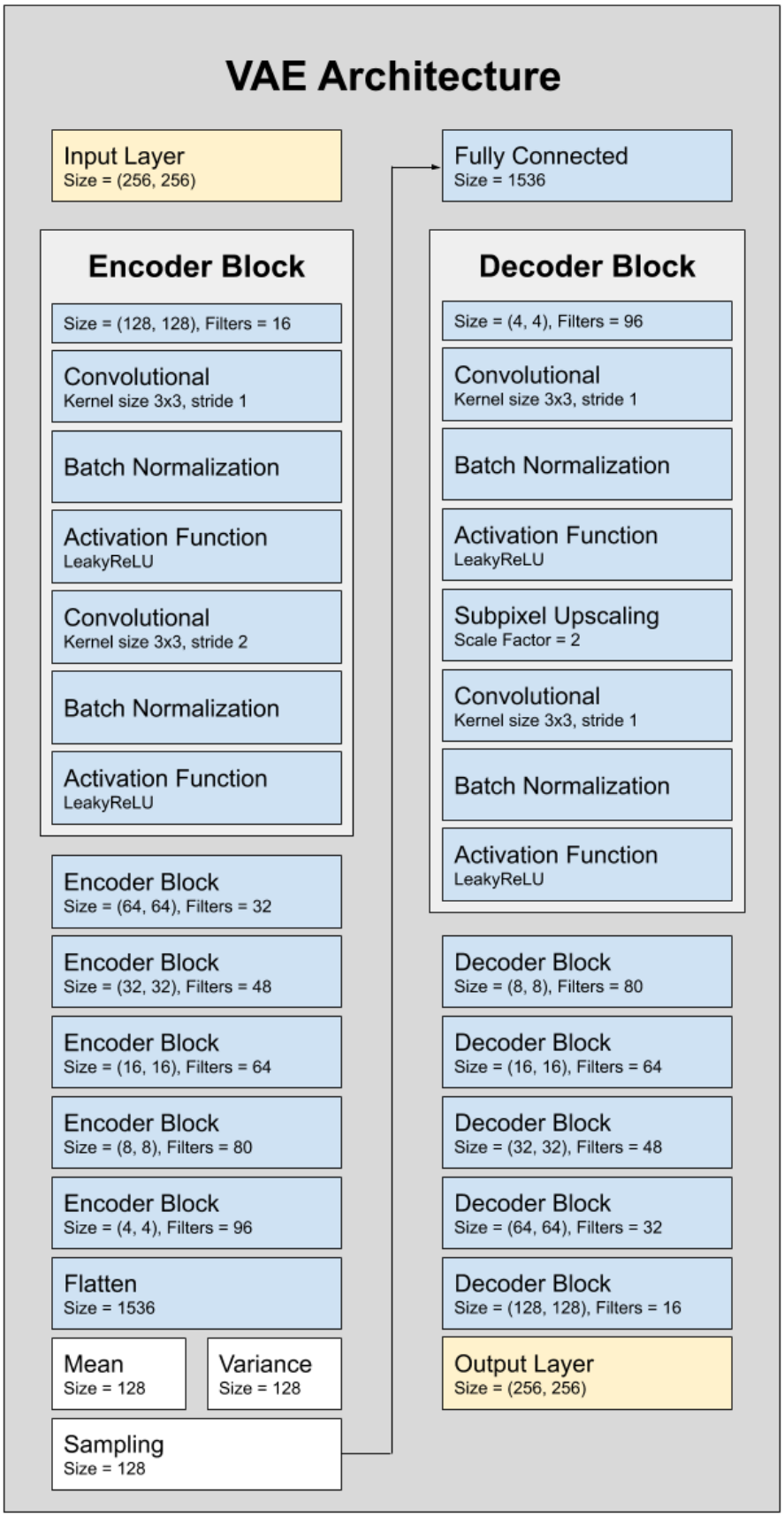

FIGURE 3.1: The architecture of the implemented variational autoencoder. Left column shows the encoder, where the image is downsampled through subsequent encoder blocks. The latent space is shown in white. The right column shows the decoder, where upsampling of the feature representation to the reconstructed image is performed through subsequent decoder blocks.

3.5 Results

The network was constructed with a latent space of 128 nodes, proven experimentally to be the best trade-off between high reconstruction error at the location of the anomalies, and low reconstruction error at the location of healthy tissue. Higher dimensionality would allow the network to reconstruct anomalies as deformed conglomerates of healthy tissues, whereas lower dimensionalities yielded the network to be unable to model healthy tissue nor anomalies. Once the training of the network on the discovery dataset was completed, we applied it to the study dataset. Specifically, we reconstructed the lung CT and estimated the reconstruction error per voxel of the network, which resulted in a heatmap of the anomalies in the lungs. An example is shown in Figure 3.2. Here two cases are presented: one from the study dataset (A) and one from the healthy control (B). For each case, the reconstruction produced by the VAE, and corresponding anomaly heatmap (squared error map) are displayed. Positive findings are marked with green boxes — i.e. the anomalies that should not be reconstructed and should, therefore, light up in the error map. Negative findings are highlighted with red boxes — i.e. normal tissue poorly reconstructed, with a high error as a result. As expected, no hotspots were found in the anomaly heatmap of the healthy control case, suggesting the absence of morphological signs of lung diseases.

To evaluate the anomaly heatmaps in a quantitative manner, we compared the distribution of the signal in the heatmaps between the study dataset and the healthy control group (Figure 3.3). Comparison of the average signal in each group yields a significant difference between the means of the two groups (3.39 vs 2.89 [x 10-3], p=0.001). Further analysis in the signal distribution (percentiles) revealed a larger difference between the heatmaps of the two groups, with incremental levels of significance associated with higher percentiles until the 99th percentile ($p<0.001$ at the median, 75th, 90th, 95th, and $p=0.03$ at 99th). These results suggest the regions of hotspots of anomalies in the study group that were not found in the control group.

3.6 Discussion

Aim of this study was to develop novel AI techniques for the identification of morphological lung anomalies in patients with a recorded history of asbestos exposure. Specifically,

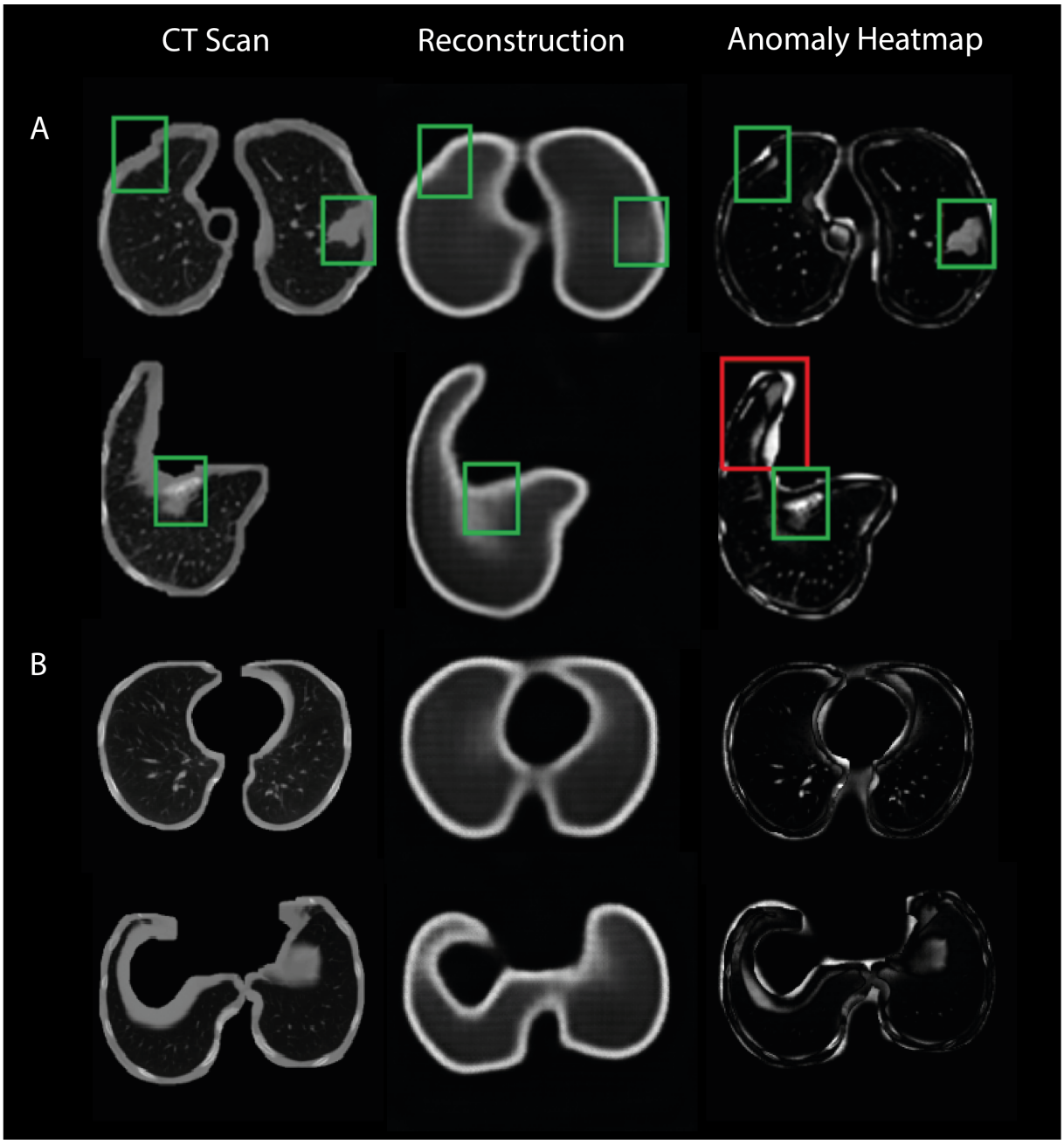

FIGURE 3.2: The input (CT scan), output (Reconstruction), and the squared error between them (Anomaly Heatmap) visualized. (A) Two slices of the same patient out of the study dataset with anomalies. The first row shows a pleural plaque (left box) and a mass (right box), where the second image shows a calcified pleural plaque located at the diaphragm. The green boxes highlight anomalies that should return a high reconstruction error, where the red box shows normal tissue poorly reconstructed. (B) Two slices of the same patient out of the control dataset. The first row shows a slice around the heart, where the second row shows a slice around the diaphragm, which yields a higher reconstruction error of normal tissue.

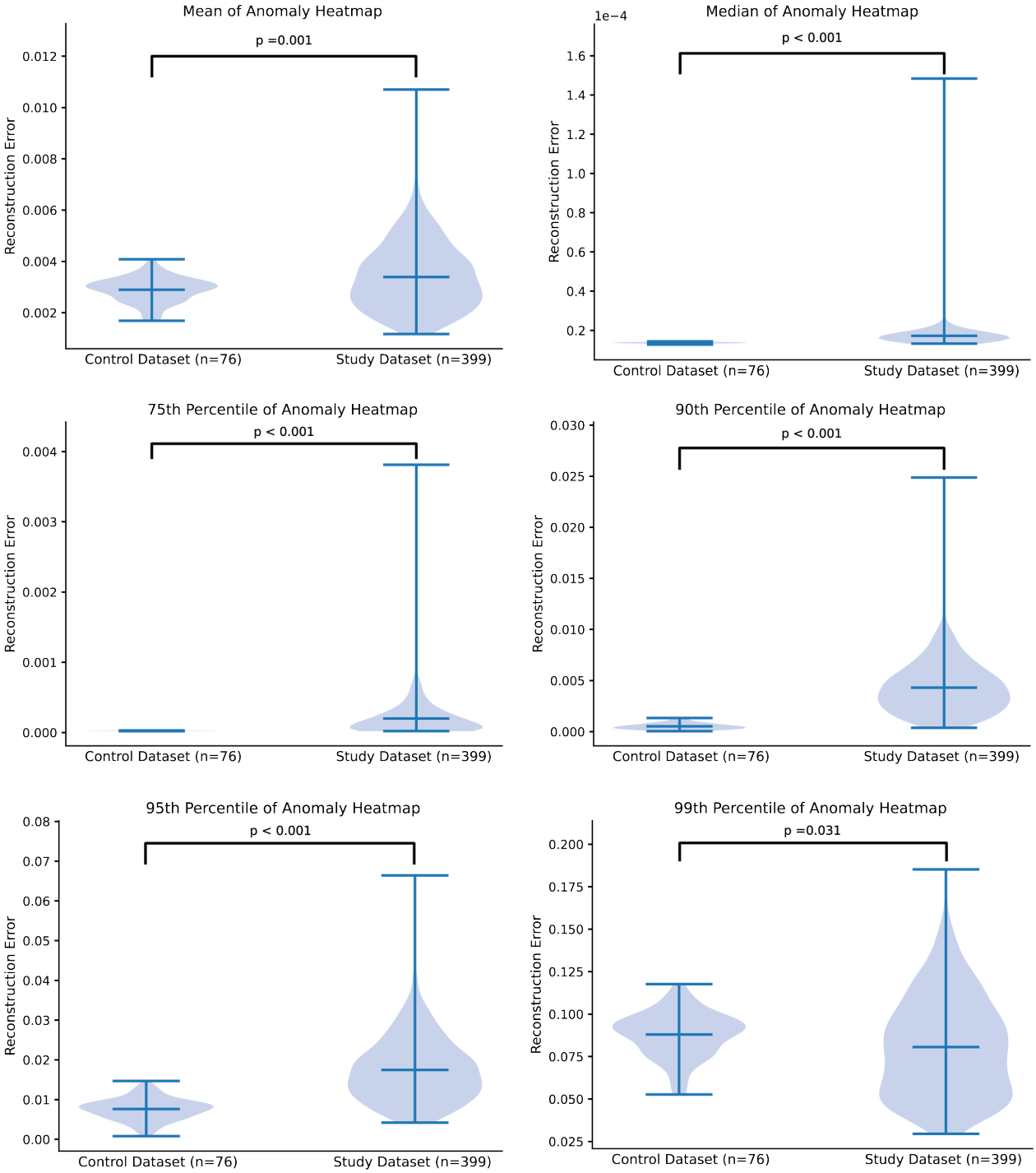

FIGURE 3.3: Violin plots of the distribution of the anomaly heatmap for multiple percentiles. The y-axis shows the reconstruction error for the given percentile. The x-axis shows an approximation of the frequency in which the reconstruction error occurs for the healthy control dataset (left) and the asbestosis study dataset (right).

we aimed to detect all lung morphological anomalies — allowing occasional false-negative results. To this aim, we employed a variational autoencoder network (VAE) to generate anomalies heatmaps on lung CT images.

The study dataset yielded a significantly higher reconstruction error than the control dataset, suggesting the presence of sufficiently large morphological anomalies in the study dataset localized via the usage of our model. The more the percentiles of both datasets for testing increased, the more the dataset diverged, peaking at the 90th percentile, after which it declined. This is expected since the CT scans in the asbestosis dataset contain more voxels of anomalies that are poorly reconstructed, leading to higher values at the 75-95th percentiles. At 99th, the approximated maximum error is returned for both datasets, which occurs partly due to false positive results, resulting in more uniform distributions. The same concept holds for percentiles lower than the median, where the air in the CT scans is reconstructed, yielding no differences in reconstruction error between the datasets.

Visual inspection of the reconstructed lungs shows the ability of the model to identify anomalies along the thoracic wall, as well as the presence of extensive fibrotic tissue. As lungs are typically smooth along the borders, pleural plaques tend to interrupt the smooth curvature of the pleura, resulting in a higher reconstruction error, and therefore a hotspot on the anomaly map. This is also shown in our exemplary figures, where the deformation of the thoracic wall is noticeable at the location of the pleural plaque. Fine level details as well as texture features in large masses are also poorly reconstructed, with blurry results often seen in the parenchyma, as also shown in the example cases we present. This is the result of the size of the latent space ($|h| = 128$) as a trade-off between reconstruction quality and semanticity of the model: larger sizes (or even fully-convolutional alternatives) would allow the network to reconstruct morphological anomalies as conglomerations of patches of healthy tissues. The reconstruction of these kinds of masses led to the preference for a restricted, fully connected latent space instead of a large, or fully-convolutional one. This reduction came at the cost of the quality of the reconstruction, which resulted in blurry images. In addition to this, standard VAE are known to intrinsically return blurry reconstructions, when compared to other network architectures[25]. However, it serves its purpose better as the blurry form that does not reconstruct the anomalies. As we observed in our dataset, normal fine-grained structures like the trachea and bronchi should ideally have

been reconstructed perfectly. Further methodological research on more advanced methods for the representation of fine-grained healthy structures is beyond the scope of the present study.

An additional point of discussion is the choice for a 2D instead of a 3D architecture. This was based on: (1) the absence of publicly available datasets of healthy subjects, (1) the larger amount of 2D vs 3D images (multiple slices per volume), and (3) the intrinsic batch-size requirements of the VAE model which would be limited to a couple of samples per batch in case of 3D images.

The histogram of the output values of the layers in the latent space is noteworthy since the mean layer shows the expected behavior, but the variance layer does not. It seems to favor the small regions of uncertainty to keep a predictable reconstruction without too much deviation. This phenomenon can be countered with the tuning of the weight put on the KL-loss or the reconstruction loss each, but this will come at the cost of lower quality of the reconstructions and is not desirable. The sampled vector is different for each CT scan relative to the mean layer but has approximately the same mean overall CT scans combined, which is expected. The sample factor epsilon is a random number out of the Gaussian distribution, which is expected to yield a mean of zero over infinite iterations.

The slices that were annotated in the LIDC-IDRI supplementary spreadsheet were removed during preprocessing. However, few anomalies were still visible in a number of CT scans, which were not in the annotation spreadsheet. These visible anomalies are, however, not expected to have any significant impact on the training, since these were only a few slices and the batch size was sufficiently large. Further research should focus on more robust alternatives of standard VAE[26], which, however, fall outside the scope of the present research.

Further development will focus on using the resulting anomaly heatmap to strengthen the accuracy of diagnostic specific networks developed in the subsequent chapters of this thesis.

3.7 Conclusion

In this study, we developed an AI model for the detection of morphological lung anomalies applied to patients with a recorded history of asbestos exposure. For each chest CT scan, our model returns an anomaly heatmap, where hotspots represent imaging features and structures that are not present in the model's training population. We studied these heatmaps qualitatively and quantitatively. Our results suggest that these methods can be employed for the detection of large morphological anomalies in the lungs, and could provide further insights for the clinical and methodological research on asbestos exposure.

In preparation for submission for publication to the American Journal of Respiratory and Critical Care Medicine

4

Automated Classification of Asbestosis

4.1 Introduction

Construction and manufacturing industries have long abandoned the use of asbestos — pushed by the government. However, due to the long incubation time, many patients are now presenting in the clinical setting with shortness of breath, persistent dry cough, and chest pain. Imaging of the chest reveals typically (calcified) pleural plaques, fibrosis, and atelectasis[2–5]. Patients that endured occupational asbestos exposure and are diagnosed with asbestosis could be eligible for compensation. In the Netherlands, the Institute for Asbestos Victims (IAS) facilitates these compensations under the directives of the government. A committee consisting of three to five pulmonologists out of a total pool of twenty pulmonologists evaluates the application of patients with the CT scan, lung function, and the history of occupational asbestos exposure. Here, the pulmonologists are blinded to each other's verdict. There is no unanimity required: two out of three pulmonologists positive is sufficient to get the diagnosis of asbestosis. The approval process for asbestosis can be troublesome since most applications only involve noninvasive measures that are less specific than the invasive ones as described in the Chapter 1.

Currently, the members of the committee give their approval for a positive asbestosis diagnosis if three criteria are met: (1) the patient has a sufficient history of occupational asbestos exposure, (2) the surface of the lung parenchyma in the CT scan of the patient is at least 5% covered with fibrosis, and (3) the patient has a reduced lung function. This is the legal diagnosis for asbestosis, rather than the clinical one. For the first criterion, a risk matrix was developed to state the intensity of asbestos of the most common occupations per decade, for the period of 1945-1995. More specifically, the years of work are multiplied by the corresponding intensity factor for the patient's occupations during that time, leading to an overall grade of the intensity of total asbestos exposure, which can be converted to fiber years. This value has to be higher than five fiber years to meet the criterion on sufficient history of occupational asbestos exposure[27]. The second criterion of lung parenchyma fibrosis is evaluated through visual radiological inspection, where an experienced reader estimates the 3D volume of fibrosis, from the 2D slices of the CT scan. The fibrosis has to cover at least 5% of the pleural surface. The third criterion is lung function loss, which is estimated on a 5-point scale based on the criteria by the American Medical Association (AMA) and "Guides

to the evolution of permanent impairment," 6th edition 2008. These guidelines describe the three most indicative parameters of lung function loss of patients with asbestosis: (1) forced vital capacity (FVC), (2) diffusing capacity for carbon monoxide (DLCO), and (3) the maximal oxygen consumption (VO2 max). FVC is the total amount of air the patient can exhale by force after a full inhalation in liters. The DLCO describes the ability of the carbon monoxide (as a substitute for oxygen) to transfer into the blood in ml/min/kPa. VO2 max is the maximal uptake of oxygen during incremental exercise in ml/min/kg. The lowest-scoring one determines the lung function loss category (Table 4.1). AMA class > 1 is required to meet the third criterion. Besides the AMA-classification and their corresponding lung function tests, the vital capacity (VC) and the carbon monoxide transfer coefficient (KCO) are often given to assist the pulmonologists in their assessment of the lung function of the patient. VC is the total amount of air the patient can normally exhale after a full inhalation. KCO is the DLCO value compensated for the alveolar volume and hemoglobin concentration.

Following these criteria, the pulmonologists assess the application. This way of processing applications is currently considered the state-of-the-art and yields satisfactory results, but is both time-consuming and expensive. The inter-observer variability is high: there was unanimity between pulmonologists on the panel in only 76% out of the first 507 applications.

Automatic solutions such as convolutional neural networks (CNNs) can aid in this process. These CNNs have the ability to extract useful information from medical images such as CT scans. By training the CNNs on the CT scans of patients with the label as the verdict of the panel of pulmonologists, the CNN will recognize certain features in the CT scan that correspond to the diagnosis of asbestosis. These features will be used by the CNN to classify CT scans with asbestosis from CT scans without asbestosis. Here, we do not aim to classify the clinical diagnosis of asbestosis, but rather the legal diagnosis described above. After training, the CNN is tested on CT scans in the test set. If it performs sufficiently, it can possibly be deployed for automatic classification of CT scans of patients that file an application. The CNN will output a probability of the patient having asbestosis based on the CT scan.

Our aim is to improve the speed, consistency, and costs of the diagnosis, which is not only beneficial to the panel but also to the patients who file the application. By using the data that is available to the pulmonologist, and required to get the diagnosis, we can estimate the probability of the patient getting the legal diagnosis of asbestosis. Furthermore, we can test

TABLE 4.1: Table for converting lung function parameters to AMA class. FVC and DLCO values are the corrected percentages for age, length, and sex of the predicted normal value. VO2 max is given in ml/min/kg.

Class	**0**	**1**	**2**	**3**	**4**
FVC	$\geq$80%	70-79%	60-69%	50-59%	<50%
DLCO	$\geq$75%	65-74%	55-64%	45-54%	<45%
VO2 max	>25	22-25	18-21	15-17	<15

the correlation between the predicted score of the CT scan and the lung function tests of the patients to gain insights into the model decision-making process. For certain thresholds within this probability, we aim for automatic approval or denial of the application based on the CT scan and lung function test with the supervision of one pulmonologist. The cases that are less straightforward can consequently be processed by the panel. This automatic processing of the application can decrease the workload for the pulmonologists in the panel and help in the development of a more standardized method.

4.2 Material and Methods

4.2.1 Datasets

The study dataset consisted of 523 patients with suspicion of asbestosis, collected at the Netherlands Cancer Institute (NKI; Amsterdam). The mean age is 74.5$\pm$7.6 years and the dataset only contained 2 females. Exclusion criteria were: the absence of a CT-scan, CT slice thickness > 5 mm, and/or insufficient quality of the lung segmentation (see Chapter 3). To ensure the inclusion of the thoracic wall, lung segmentations were dilated through morphological operators with a kernel of 20 x 20 x 5 voxels. The anomaly heat maps, derived from Chapter 3, are used as an additional input to guide the model attention: they indicate regions that contain anomalies, and, therefore, could be used as extra information for the classification. Lung function tests for most patients in the study dataset were retrieved. When recorded, the parameters were: VC, FVC, DLCO, and KCO. FVC and DLCO were used to convert to an AMA class.

A total of 523 patients were retrieved for this study. Patients were excluded due to the absence of CT scans (N=74), slice thickness >5 mm (N=10), insufficient quality of the lung

segmentation (N=16), non-thorax-specific scan (N=13), no verdict yet of the panel (N=3), insufficient asbestos exposure (N=4). Special cases (N=4) presenting with seeming fibrosis but no asbestosis and vice versa were held back for evaluation purposes. This resulted in a total of 399 patients (and corresponding CT scans) from the study dataset.

4.2.2 Data Curation

To mitigate differences of imaging protocols, all CT density histograms were clipped between -1024 and 3072 Hounsfield Units (HU) and scaled on the interval [0, 1]. CT scans were cropped to 192 x 192 x 96 (x, y, slices) due to hardware constraints. Also due to hardware constraints, images were cropped at 192 x 192 around the lung segmentation. If the segmented lungs were larger than 192, then axial rescaling was applied after cropping.

4.2.3 Network Design

The 3D ResNet-18 architecture was implemented[28]. It learned features from the CT scan (and corresponding anomaly heatmap) from 192 x 192 x 96 x 2 through multiple convolutions with striding operations to 3 x 6 x 6 x 512. The global average pooling layer compresses the feature maps to a vector representation. These 512 features are subsequently fed to the logistic classifier, which results in a corresponding score of each class (e.g. asbestosis or no asbestosis). If the model is well calibrated, this score can be interpreted as a probability. An overview of the implemented ResNet architecture is shown in Figure 4.1. Due to results stated in the Results section, we implemented a lung function parameter as input of the ResNet model as well. For this model, an additional layer was implemented before the classification layer with four fully connected nodes to summarize the 512 pooled features of the CT image input. We implemented the lung function parameter value parallel to this layer and connected it to the classification layer.

4.2.4 Labels

Labels were implemented in two configurations: hard and soft. The hard labels are binary (i.e. asbestosis or not), whereas the soft label reflects the agreement of the diagnostic board with the number of positive pulmonologists divided by the total number of pulmonologists (e.g. one out of three is 0.33, two out of three positive is 0.67).

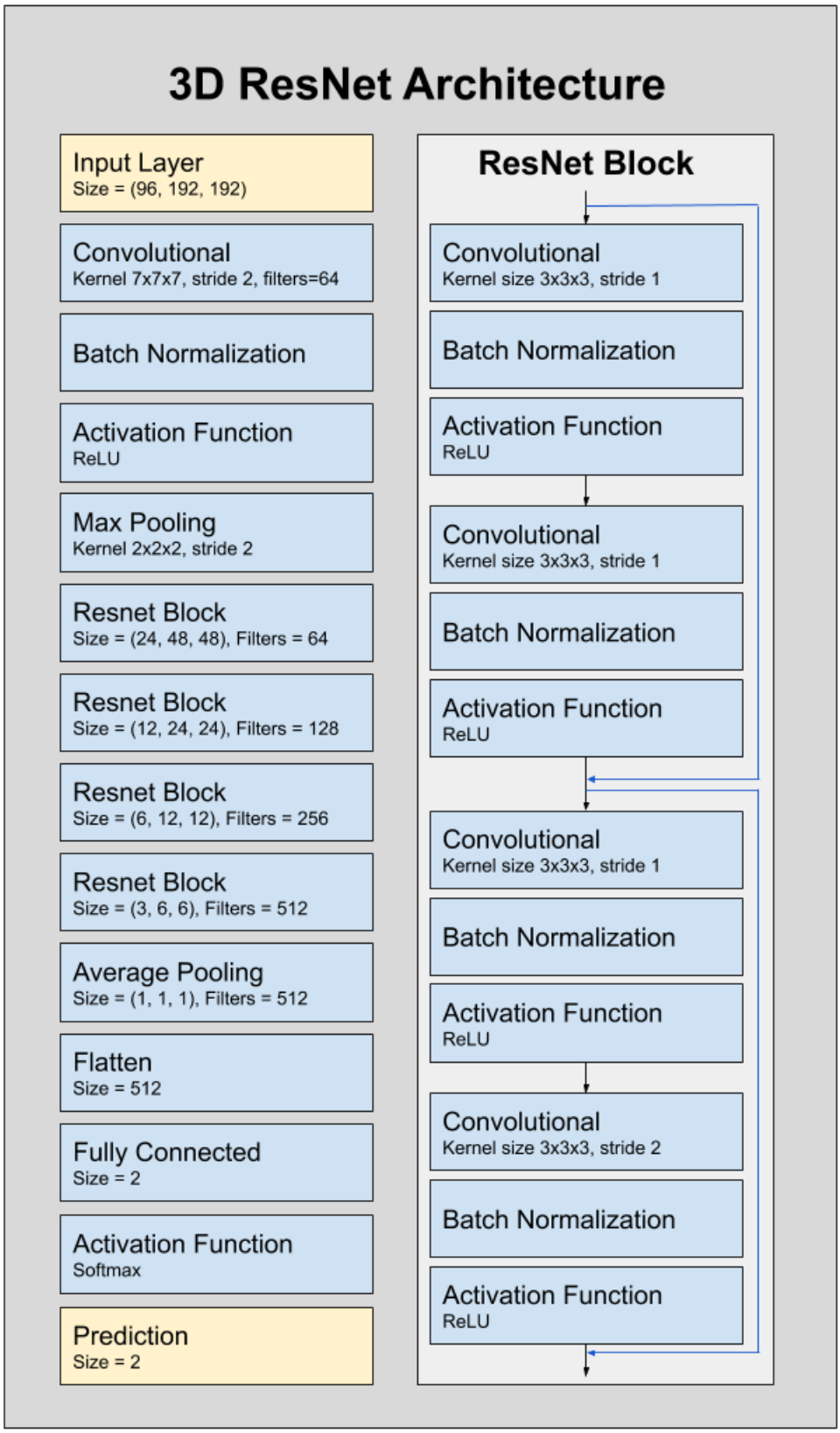

FIGURE 4.1: The architecture of the implemented 3D ResNet. The left column shows the encoder, where the image is downsampled through subsequent ResNet blocks to generate a prediction. The right column shows the ResNet block architecture. The black arrows represent the connections of the blocks. The blue arrows represent the identity connections, where the output of an activation layer is added to the input of another convolutional layer.

4.3 Experiment

4.3.1 Network Implementation

The 3D ResNet network was implemented and trained using Tensorflow (v1.15.0) and Keras (v2.3.1) libraries on two NVIDIA GeForce RTX 2080Tis. The dataset was randomly split in a training (N=240), validation (N=64), and a test set (N=88). The batch size was set at sixteen total, eight per GPU. Adam was used as optimizer, with an initial learning rate of 1e-3. The CNN was trained for a maximum of 200 epochs, where early stopping was used to stop the training if the validation loss did not improve over 30 epochs. The best model checkpoint at the end of every epoch was performed. Data augmentation with rotation (up to 10°) around the longitudinal axis, and flipping over the sagittal plane of the image was implemented at runtime during training. We tested different setups according to different label formats (i.e. soft and hard) and inputs (i.e. with and without the anomaly heatmap). The performance of the best scoring setup was further developed, which means it was retrained with that specific setup and the performance was subsequently analyzed. We performed McNemar's test to test for significant differences between the performances of different setups.

4.3.2 Analysis

After training, we used the model to make a prediction on the CT scans in the test set. To visualize the areas in the CT scan where the model is 'looking at' to make a prediction, we traced the activations of the prediction back to the input, creating saliency maps. These saliency maps contain higher values on areas of the CT scan that contribute more towards the final prediction. The performance of the trained models were evaluated through area under the curve (AUC) of the receiver operating characteristic (ROC), accuracy, sensitivity, specificity, positive- and negative predictive value.

4.4 Results

We trained the different setups for our CNN on the training set, retrieved the best scoring model on the validation set, of which the results on the test set are presented in Table

TABLE 4.2: The results of the different tested setups of CNN models. The bold number show the maximal performance in terms of the metric of that column. The p value was computed with respect to the best performing model, in this case the Soft + Anomaly Heatmap setup. Other combinations yielded no significant differences.

Label	Anomaly Heatmap	ACC	SENS	SPEC	PPV	NPV	AUC	p
Hard	Yes	0.65	0.46	0.81	0.68	0.63	0.77	0.017
Hard	No	0.65	**0.93**	0.40	0.58	**0.86**	0.66	0.042
Soft	Yes	**0.78**	0.71	**0.85**	**0.81**	0.77	**0.78**	
Soft	No	0.66	0.78	0.55	0.60	0.74	0.66	0.043

4.2. Overall, soft labels in combination with the anomaly heatmap yielded the best performance in terms of accuracy, specificity, and positive predictive value. Following the McNemar test, it yielded significant differences to the other setups ($p<0.05$). All other combinations yielded no significant differences ($p>0.05$) and were omitted from the table. More specifically, the model with the combination of soft labels + anomaly heatmap, denoted as CNN(CT), yielded an AUC of 0.87 (CI: 0.78 — 0.94, $p<0.001$). Setting a cut-off probability value of 0.5, yielded an accuracy of 0.82 (CI: 0.74 — 0.90), with a sensitivity of 0.76 (CI: 0.62 — 0.88), and a specificity of 0.87 (CI: 0.77 — 0.96). Positive and negative predictive values were 0.84 (CI: 0.71 — 0.95) and 0.81 (CI: 0.69 — 0.91), respectively. In Figure 4.2, we present two cases with the saliency maps, where the asbestosis positive case shows more activations than the asbestosis negative case. Figure 4.3A shows the box plot of the probability distributions the CNN(CT) model gives to each CT scan, according to the number of pulmonologists that gave a positive assessment.

Outlier analysis was carried out in those cases where three out of three pulmonologists were positive, but the model prediction was negative (N=5). Most of those patients (N=4) had the highest AMA class, indicating a severe reduction in lung function. After testing the separate lung function parameter as a predictor of asbestosis, the DLCO yielded about the same performance as the model (0.85 AUC, CI: 0.80—0.89, p = <0.001). All other lung function parameters yielded lower results: 0.67 AUC for VC (CI: 0.60—0.72, $p<0.001$), 0.63 AUC for FVC (CI: 0.56—0.68, $p<0.001$), 0.75 AUC for KCO (CI: 0.69—0.81, $p<0.001$), and 0.83 AUC for AMA (CI: 0.78—0.87, $p<0.001$). For completeness, our model prediction is plotted against all lung function parameters in Figure 4.4. Interestingly, the correlation between the ResNet prediction and the DLCO was weak to moderate negative (r=-0.47), indicating that

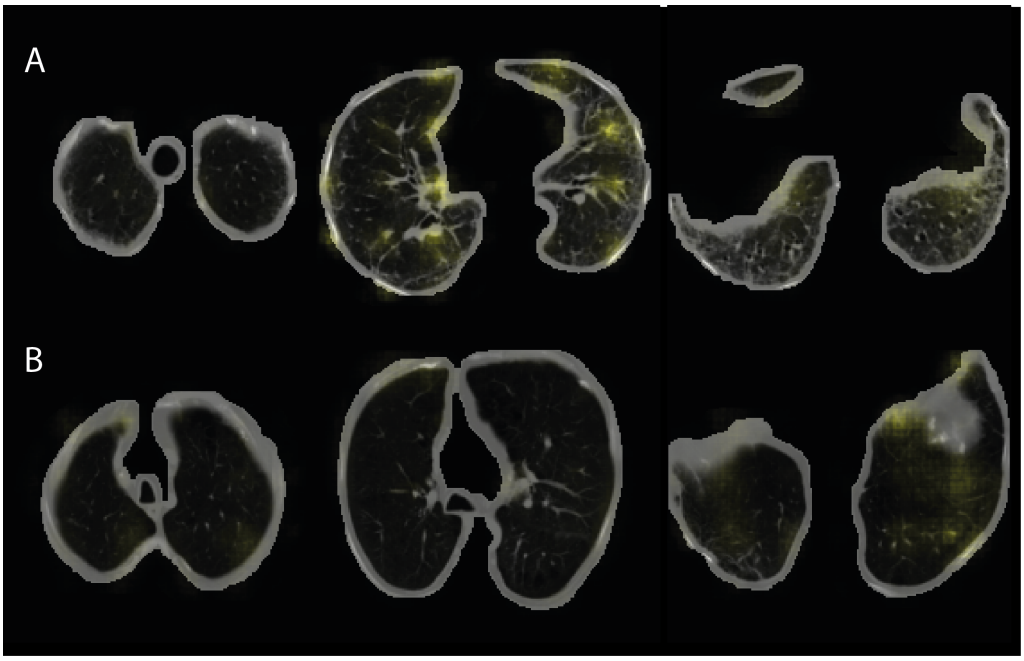

FIGURE 4.2: Saliency map yielded by the CNN(CT) of two CT scans in the test set. The areas in yellow represent the attention of the model. The left column shows a slice in the top of the lungs, the second column a slice in the middle of the lungs, and the last column a slice at the bottom of the lungs. (A) CT scan where 3/3 pulmonologists were positive and the model yielded a high probability of asbestosis (0.81). (B) CT scan where 0/3 pulmonologists were positive and the model yielded a low probability of asbestosis (0.19)

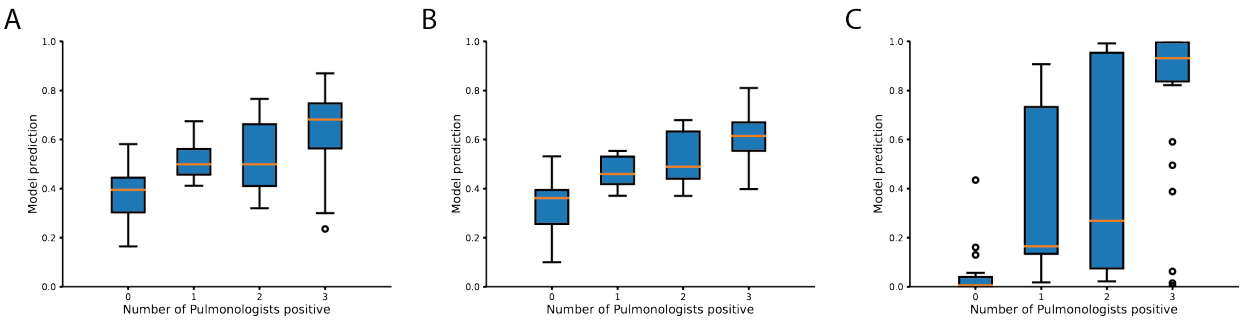

FIGURE 4.3: Boxplots on different setups of prediction. The x-axis shows the agreement of the panel of pulmonologists. The y-axis shows the predicted probability of asbestosis. (A) The prediction of the CNN(CT) model. (B) The score of the CNN(CT) + DLCO. (C) The prediction of the CNN(CT, DLCO) model.

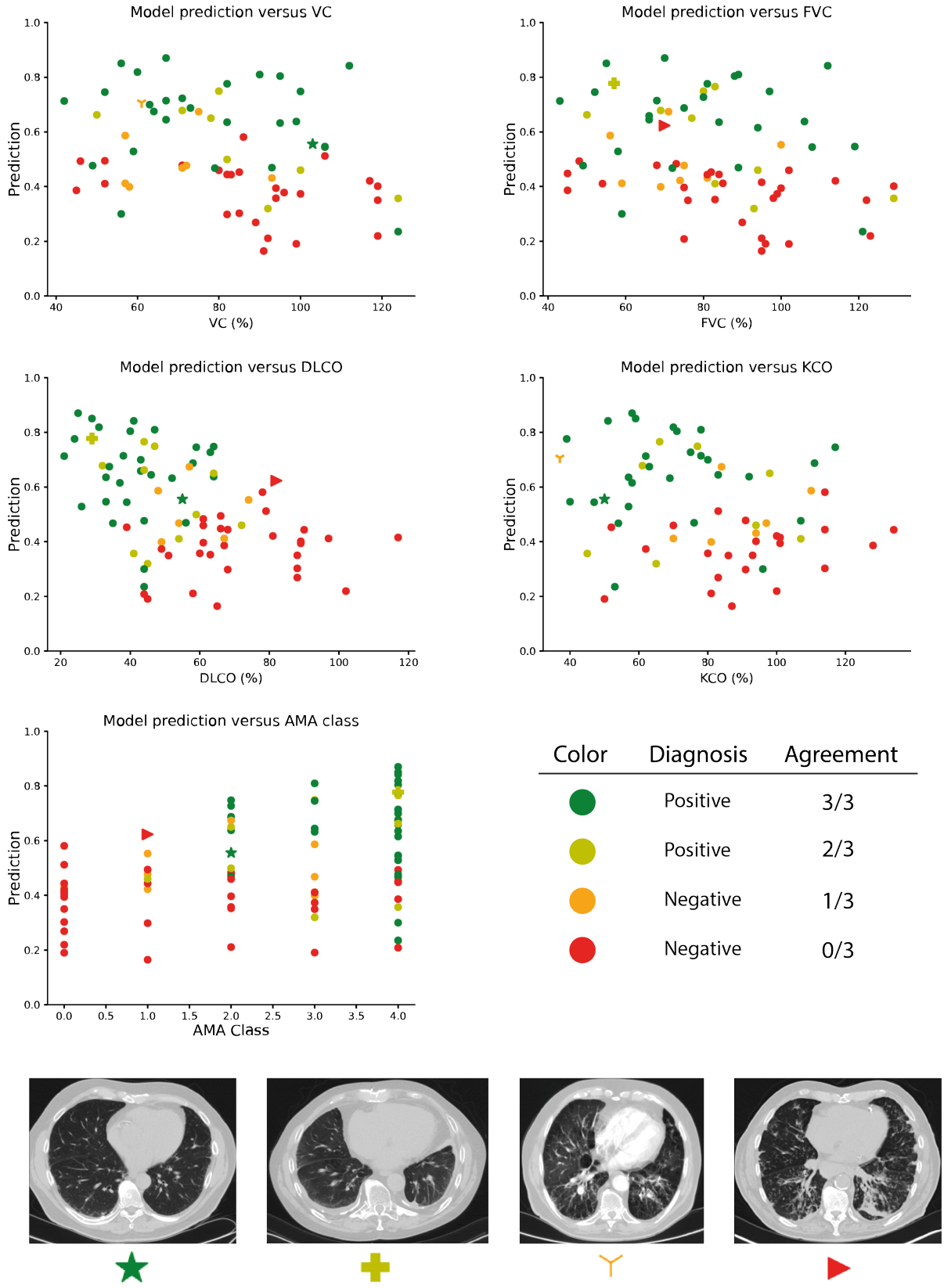

FIGURE 4.4: Probability of asbestosis predicted by the CNN(CT) model versus the lung function parameters in the percentage of the predicted score of that patient. The bottom row shows several cases where the amount of fibrotic tissue does not reflect the verdict of the pulmonologists. The symbols of each example are visualized in the figures when the respective lung function parameter of the patient is known. The colors reflect the agreement of the panel of pulmonologists, from asbestosis positive (green dots) to asbestosis negative (red dots).

both variables are independent predictors of asbestosis. Based on the approximately equal AUC of the CNN(CT) and the DLCO, and to prevent overfitting, we combined the two scores by $score = \frac{1-DLCO+CNN(CT)}{2}$ with a minimum value of 0 and a maximum of 1. This score, denoted by CNN(CT)+DLCO, yielded a ROC-AUC of 0.95 (CI: 0.89—0.98, p<0.001). Setting a cut-off probability value of 0.5, yielded an accuracy of 0.84 (CI: 0.76—0.92), with a sensitivity of 0.77 (CI: 0.63—0.89), and a specificity of 0.91 (CI: 0.81—1.00). Positive and negative predictive values were 0.91 (CI: 0.80—1.00) and 0.78 (CI: 0.65—0.90), respectively. The prediction scores for this setup do not yield any false negative and positive under 0.35 and above 0.60, respectively (Figure 4.3B). We trained a ResNet model which took the CT and the DLCO as input as well, denoted by CT(CNN, DLCO), which yielded a ROC-AUC of 0.92 (CI: 0.86—0.97, p=<0.001). Setting a cut-off probability value of 0.5, yielded an accuracy of 0.84 (CI: 0.76—0.92), with a sensitivity of 0.74 (CI: 0.60-0.87), and a specificity of 0.94 (CI: 0.85—1.0). Positive and negative predictive values were 0.94 (CI: 0.83—1.0) and 0.77 (CI: 0.64—0.89), respectively. The spread of predictions in agreement with the pulmonologists was wider, as shown in Figure 4.3C. More specifically, the CNN(CT, DLCO) predicts more CT scans closer to either zero or a hundred percent probability than both the CNN(CT) and the CNN(CT)+DLCO do.

4.5 Discussion

Medical assessment for financial compensation for asbestosis patients is a laborious process. In this study, we aimed to automate this process by means of artificial intelligence, namely use AI for automatic classification of patients that apply for compensation. More specifically, we implemented a convolutional neural network based on CT-scans of the chest. Predictive performances were evaluated with respect to the outcome, as well as several lung function parameters, when available.

Our findings show significant predictive performance of our CNN(CT) model, comparable to the best performing lung function parameter (i.e. DLCO). The combination of the AI classification and DLCO values, the CNN(CT, DLCO) model, seems particularly stronger, yielding a superior diagnostic performance than CNN(CT) and DLCO alone. Apart from the DLCO, the other lung function tests did not seem to possess any predicting value of

asbestosis. The outliers where all pulmonologists were positive about asbestosis but the CNN(CT) yielded a low probability, could suggest that the diagnosis for these patients was largely based on lung function tests. This indicates that the CNN(CT) model was not able to detect a worsening lung function based on the CT scan of these patients.

Medical experts do usually not assess CT scans without any prior information about the patient. Here, an assessment from another physician provides the medical expert the information needed to locate the anomalies in imaging scans. The same can apply for machine learning models, especially in binary classification problems. If we deliver an anomaly heatmap that indicates what the possible anomalies in a CT scan are, other algorithms (e.g. classification and segmentation neural networks) could benefit from this prior knowledge. This concept is applied to our models in the form of the anomaly heatmap created by the variational autoencoder in Chapter 3. We showed that the concatenation of the anomaly heatmap to the input of the ResNet model increased the classification results.

The implementation of soft labels further increased the classification result. The soft labels indicate the uncertainty in the assessment of patients that are borderline asbestosis. This enabled the model to learn the distribution of the probability and uncertainty of asbestosis better. Medical image analysis is notorious for label noise[29]. Especially for the patients where the panel of pulmonologists did not reach unanimity, there is a plausible chance that the verdict of the panel might not always be the true underlying biological factor. Cross entropy is an unbounded loss function — i.e. the range is [0, inf] instead of [0, 1] — which makes it more vulnerable to the influence of outliers[30]. When applying the soft labels of 0.33 and 0.67, the model gets penalized less by wrong predictions of cases that are already dubious in the eyes of specialists. This could be an explanation for the better results of the soft labels in comparison to the hard labels.

The observation that DLCO seems to have a superior diagnostic accuracy compared to AMA class for asbestosis is noteworthy. The pulmonologists refer to the AMA classification to determine whether the lung function reaches a certain threshold. However, based on our findings, the inclusion of the FVC only deteriorates the ability to distinguish between asbestosis positive and negative patients.

The study contains limitations. The CT scans have to be downsampled to a lower resolution due to hardware limitations. This won't have implications for larger structures, but

it does for finer-grained structures like fibrosis. By comparing the original CT scan to the input of the ResNet model, it was evident that the detailed structures of fibrosis in the original CT scan were lost in the downsampled input. This could lead to an inability of the ResNet model to learn the correlation between the fibrosis and the likelihood of asbestosis. Higher-resolution models that can process the segmentation of the lung at full scale could, therefore, yield better results.

4.6 Conclusion

We developed an AI model for the classification of asbestosis in patients. Classification models based on only the CT scan and on a combination of the CT scan and the lung function test were quantitatively and qualitatively assessed. The model that made classifications based on the CT scan and the DLCO was superior to the other model and reached excellent diagnostic accuracy. The results suggest that the implementation of this model in the clinical setting could have benefits for the patient in terms of reproducibility, consistency, and speed of the assessment of asbestosis.

In preparation for submission for publication to Thorax

5

Pleural Plaque Quantification and Correlation to Lung Function

5.1 Introduction

Pleural plaques are specific manifestations of asbestos exposure. They present as local areas of hyalinized collagen fibers and may vary in calcified or noncalcified form[2–8]. The exact mechanism of pleural plaque formation is not well understood[3,9]. The likelihood of the plaques is correlated to the time since first exposure to asbestos and the cumulative exposure[6], but they can even form after minimal exposure to asbestos[31]. They usually occur 20 to 30 years after asbestos exposure and are typically located in both lungs[3]. Current scientific evidence[8] suggests that patients with pleural plaques are most often asymptomatic, with no evident association between pleural plaques and lung function test results. However, it is found that pleural plaques, depending on the extension of the pleura, can lead to a restrictive lung function[7]. The extension of pleural plaques can be measured by volume, and the volume is correlated to the cumulative asbestos exposure[6]. Pleural plaque volume (PPV) assessment is found to have excellent intraobserver reproducibility (ICC: 0.98) and a very good interobserver variability (ICC:0.93)[32]. For practical assessments, one study included 26 patients. It divided those into three groups of <10 mL, 10-20 mL, and >20 mL, where no significant differences were found between the groups in terms of lung function values[33]. They quantified PPV by the outline of the plaques. Another study measured PPV of 75 patients in three-axis and correlated the measured volume to lung function parameters[34]. They did not find any significant correlations between lung functions, exercise capacity, cumulative asbestos exposure, and PPV. There is financial compensation for patients with proven mesothelioma and asbestosis in the Netherlands, but not for pleural plaques, since no correlation to impairment of the lung function has been proven quantitatively on large cohorts. The lack of such studies is partially due to the time required for human readers to segment pleural plaques on CT scans manually. The feasibility of studies with larger cohorts requires developing an alternative method for segmentation/volume quantification that is fast, accurate, and reproducible. This study aims to employ artificial intelligence for the automatic volumetric segmentation of pleural plaques in CT scans of patients exposed to asbestos. The resulting algorithm will enable us to investigate the correlation between PPV and lung functions for patients that applied for financial compensation after asbestos exposure in the Netherlands, as commissioned by the Dutch Institution for Asbestos Victims.

5.2 Material and Methods

5.2.1 Datasets

The study dataset consisted of 523 patients with suspicion of asbestosis, collected at the Netherlands Cancer Institute (NKI; Amsterdam). The mean age is 74.5±7.6 years and the dataset contained only 2 females. CT scans were derived from hospitals over the Netherlands with different imaging protocols. Patients were excluded due to the absence of CT scans (N=74) and absence of any pleural plaques (N=27), yielding a total dataset of N=422 CT scans. To test for systematic segmentation errors, we collected a control dataset (N=76) of patients and their CT scans of healthy lungs collected at the NKI.

Lung function tests for most patients in the study dataset were performed. When recorded, the parameters were: Vital Capacity (VC), Forced Vital Capacity (FVC), Diffusing capacity of Lung for Carbon Monoxide (DLCO), and Carbon Monoxide Transfer Coefficient (KCO). FVC and DLCO were used to convert to an American Medical Association (AMA) class (see General Background).

5.2.2 Data Curation

To mitigate differences of imaging protocols, all CT density histograms were clipped between -1024 and 3072 Hounsfield Units (HU) and scaled on the interval [0, 1].

5.2.3 Network Design

The UNet architecture was implemented[35]. It filtered, downsampled, and pooled features from the CT scan (and corresponding anomaly heat map). We tested both 2D and 3D architectures. The 2D architecture downsampled from 512x512x1 through multiple convolutions with striding operations to 32x32x512, while the 3D counterpart downsampled from 192x192x96x1 to 12x12x6x512. The feature representation was upsampled to a segmentation through deconvolution layers and skip connections from the encoding path. Batch normalization was implemented after each (de)convolutional layer. An overview of the implemented 2D UNet architecture is visualized in Figure 5.1. The same concept holds for the 3D counterpart.

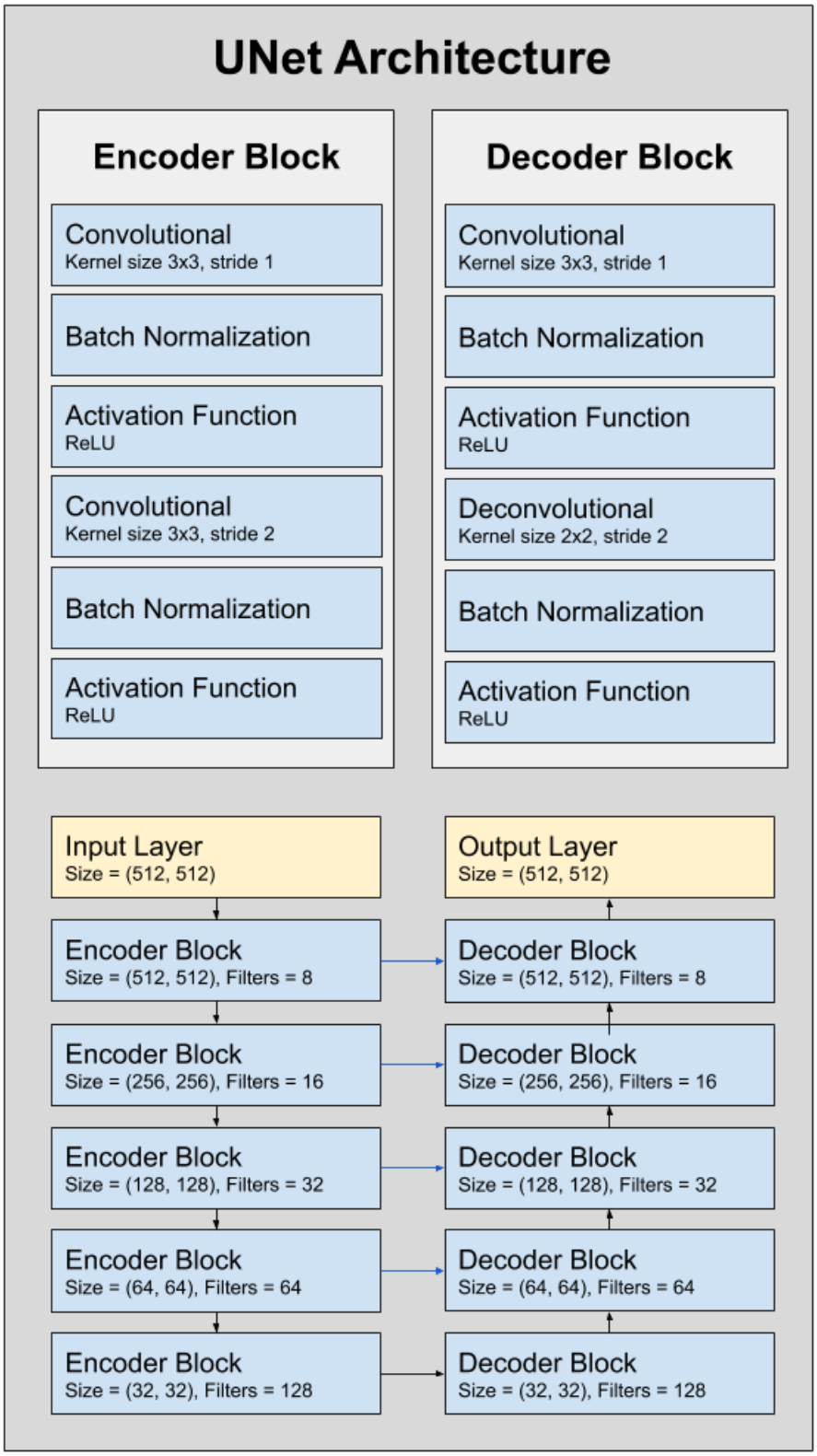

FIGURE 5.1: The architecture of the implemented UNet. (Top) The encoder and the decoder blocks used to build the AI model. (Bottom) The left column shows the encoder, where the image is downsampled through subsequent encoder blocks. The right column shows the decoder, where upsampling of the feature representation to the segmentation is performed through subsequent decoder blocks with deconvolutions. The black arrows represent the connections of the blocks. The blue arrows represent the skip connections, where the output of an encoder block is concatenated to the input of the equally sized decoder block.

5.2.4 Labels

Labels were acquired by seven radiologists. One radiologists has more than five years of experience, the rest of the radiologists less. The dataset was split equally between them and they labeled each pleural plaque voxel in each CT scan in the dataset in the program 3D Slicer. Hard cases could be reported and the most experienced radiologists would review the segmentation. Moreover, the segmentation of the most experienced radiologists could be used as examples.

5.3 Experiment

5.3.1 Network Implementation

The UNet networks were designed and trained using Tensorflow (v1.15.0) and Keras (v2.3.1) libraries on two NVIDIA GeForce RTX 2080Tis. The training set consisted of N=242 patients, the validation set of N=80, and the test set of N=100 CT scans for the 3D architecture. For the 2D architecture, only CT slices with labeled pleural plaques were retrieved. This yielded a training set of N=19496, a validation set of N=6751, and a test set of N=7743 slices, where slices of the same patient were always part of the same set. To compare performances, both the 2D and the 3D architecture were tested on the whole CT volumes of patients in the test set. The batch size was set at 24 total, twelve per GPU. Adam was used as optimizer, with an initial learning rate of 1e-3. The loss function was the Tversky loss[36], an adaptation to the Dice Similarity Coefficient (DSC) loss. It weighed the false positives and false negatives by a factor alpha and beta, respectively. The clinicians stated that the model should not underestimate the pleural plaque volume. Therefore, we opted for alpha=0.3 and beta=0.7, penalizing the missed pleural plaques by the AI model more than the falsely positive segmented voxels. This setup yielded the best results at the authors' experiment[36]. The CNN was trained for a maximum of 200 epochs, where early stopping was used to stop the training if the validation loss did not improve over 30 epochs. The best model checkpoint at the end of every epoch was performed. Data augmentation with rotation (up to 20°) around the longitudinal axis and flipping over the sagittal plane of the image was implemented during training.

5.3.2 Analysis

The performance of the trained CNNs were evaluated through DSC, which is an objective measure to determine the overlap between the predicted volume by the CNN and the radiologists (see General Technical Background). The higher the overlap between the two segmentations, the better the performance of the CNN. However, DSC does not cover the clinical aim of a reliable volume estimation completely, since it is sensitive to the total volume of pleural plaques. Therefore, the best architecture was further analyzed by (1) determining the correlation between segmented volume by the CNN and the radiologists, (2) comparison of the mean difference between the groups and how this difference compares to the mean absolute error, and (3) the error on the control set, to get an indication what the systemic error is on healthy CT scans with no plaques. With these performance measures, we evaluate the model for the intended clinical goal.

We want to avoid underestimating the PPV, but do want to remain the same order in volumes over the patients (i.e. the systemic error where the CNN segments more than the radiologist should be as constant as possible). In this way, we can reliably estimate if the total PPV of patients is exceeding a certain threshold, for example for financial compensation. The correlation gives an indication of whether the prediction of the CNN does remain the same order in volumes as to the segmentation of the radiologists. However, the pleural plaque volume should be normally distributed before determining the correlation. To convert the distribution of the PPV to a normal distribution, we apply the Box-Cox power transformation. This calculates the variable lambda and converts the distribution following $y = \frac{x^{\lambda}-1}{\lambda}$, where y is the output distribution, x the input distribution and lambda the calculated conversion variable. If the lambda value is close to a typical distribution (like 0.5 for square root or 0 for log), we showed those as well. We aim for a negative result for the differences between the true and the predicted volume, which means the predicted volume is larger than the true volume. Ideally, the mean absolute error equals the mean error, indicating no predicted volume smaller than the corresponding true volume. A mean error of zero and a large mean absolute error would indicate a random error in the segmentation of the CNN.

There were occurrences of insufficient segmentation quality of the radiologists, which means another radiologist confirmed pleural plaques were missed, or tissue was incorrectly

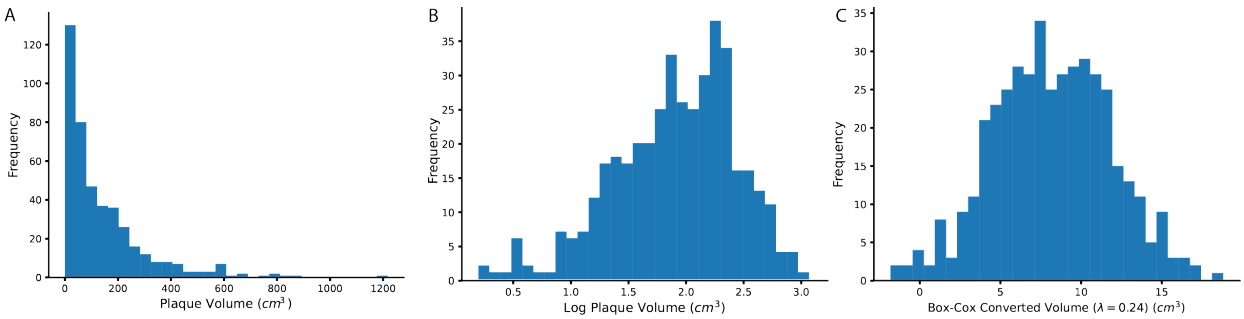

FIGURE 5.2: Visualization of the distribution of the pleural plaques. The x-axis shows the volume of the pleural plaques in cm^3. The y-axis shows the incidence. (A) The distribution with normal x-axis. (B) The distribution with logarithmic x-axis with base 10. (C) The Box-Cox converted volume, which is normally distributed.

segmented as pleural plaque. Due to time constraints, we have not been able to readjust these segmentations yet, but this will be arranged in the near future. In these cases, the segmentations will be temporarily excluded to highlight the results without the erroneous segmentations interfering.

5.4 Results

The 2D network yielded a mean DSC of 0.59 and a median DSC of 0.70 in the test set, where the 3D counterpart yielded a mean DSC of 0.49 and a median DSC of 0.56. Therefore, the predictions of the 2D network were further explored. The UNet network segmented consistently more voxels in CT scans than the experts (median: 119.3 cm^3 vs 80.9 cm^3). The mean absolute error was 43.7 cm^3 and the mean error yielded 29.9cm^3 more volume in the segmentation of the CNN on average. The Box-Cox power transformation yielded $\lambda = 0.24$ to convert all pleural plaque volumes to a normal distribution (Figure 5.2). This value indicates the distribution shares characteristics of a log distribution (λ=0) and a square-root distribution(λ=0.5). After conversion to the normal distribution of the PPV in the test set, we found a strong correlation between the predicted volume of the plaques and the segmented volume of the plaques (r=0.88) (Figure 5.3A). It is visualized on the log-log scale as well for better volume interpretation (Figure 5.3B). The regression line is plotted to indicate the consistent higher volume prediction of the CNN compared to the radiologists. The formula of the regression line is shown for completeness, not to state the relation of how one can map the one to the other.

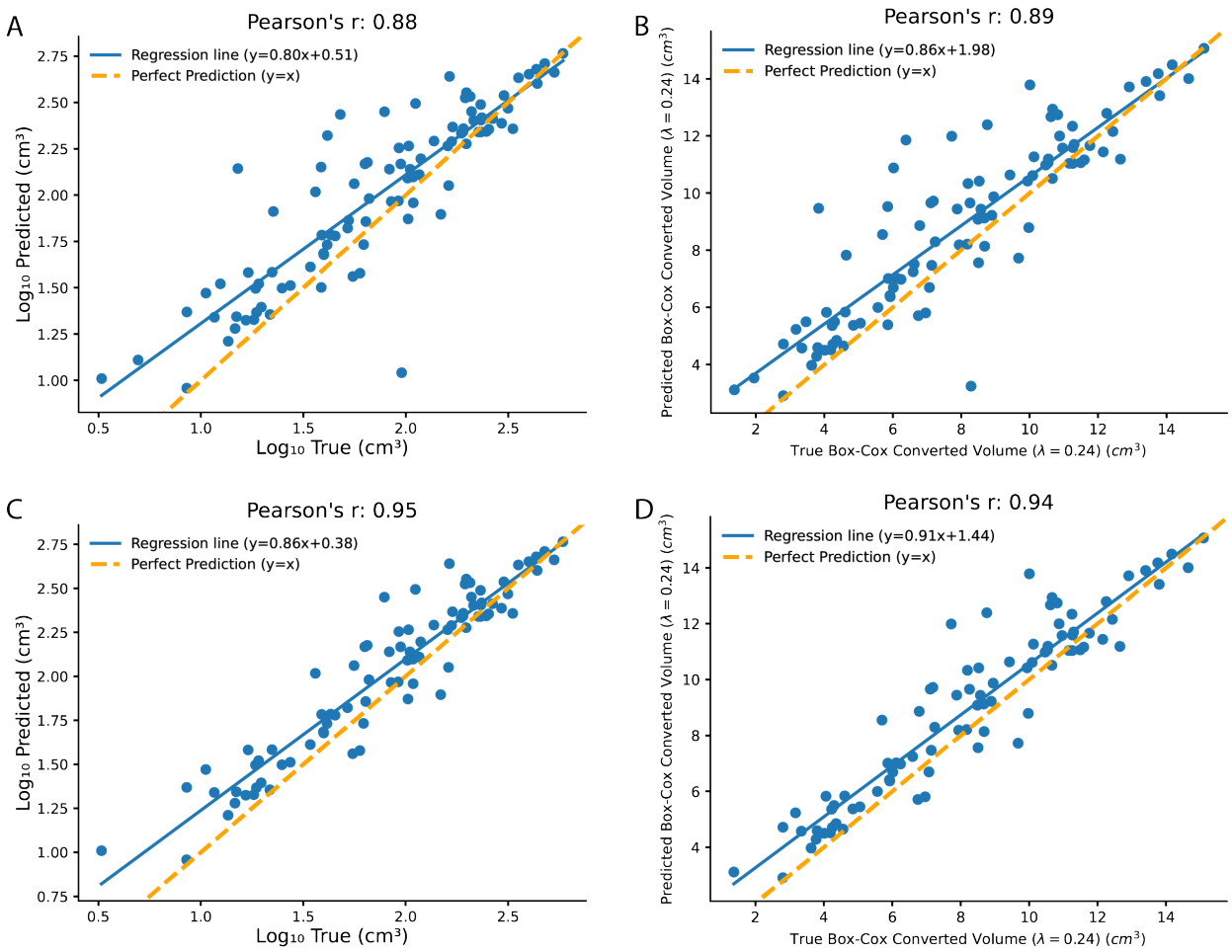

FIGURE 5.3: Visualization of the correlation between the predicted volume on the y-axis (Predicted) versus the segmented volume by the radiologists on the x-axis, both on logarithmic scales. The blue line represents the linear regression equation. The orange line represents perfect volume prediction. (A) Visualization of the volume of the N=100 cases in the test set on log log scale. (B) The cases in the test set converted to a normal distribution via the Box-Cox method. (C) Visualization after excluding the N=16 segmentations of insufficient quality. (D) The result of excluding the segmentations of insufficient quality and converting the test set to the normal distribution by using the lambda variable.

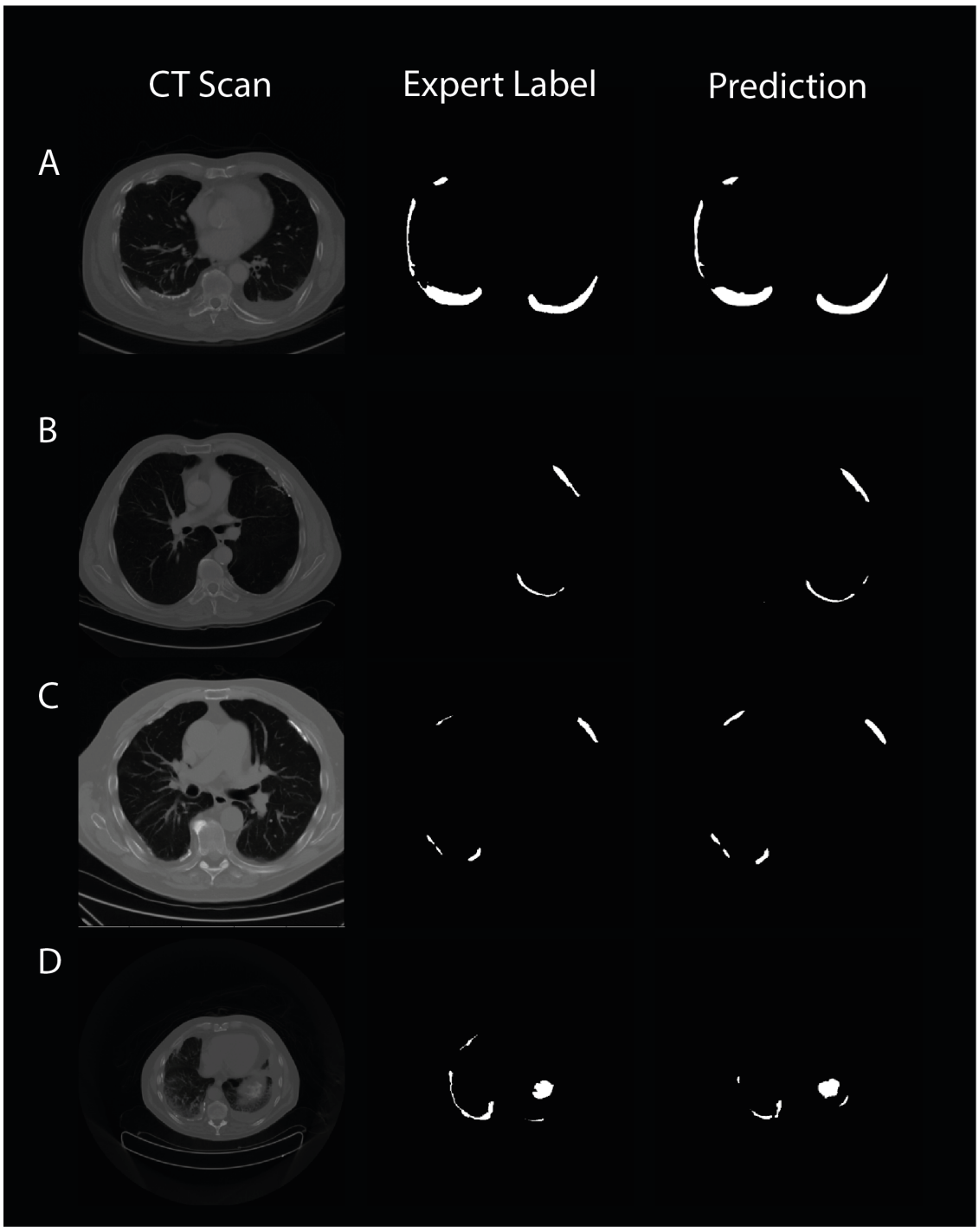

FIGURE 5.4: Visualization of cases segmented by the radiologist with >5 years of experience, with the input (CT Scan), segmentation of the radiologist (Expert Label), and the CNN segmentation (Prediction). (A) Case with a large pleural plaque volume, which is well segmented by the AI model. (B, C) Cases where the CNN model segments the same pleural plaques as the radiologist, but with more voxels in the vicinity. (D) Case where the CNN was not able to segment all the pleural plaques.

Visual results of the segmentation of the most experienced radiologist and CNN prediction are shown in Figure 5.4. On visual inspection, the majority of the segmentations by the CNN looked close to equal to the segmentation of the radiologist. Only pleural plaques around the diaphragm seem prone to segmentation errors by the CNN. The extra voxels are often in the vicinity of a pleural plaque, creating thicker segmentations.

Outlier analysis was performed on N=53 CT-scans that yielded a DSC score of <0.40. In this group, it segmented pleural plaques in N=35 CT scans that were not segmented by the radiologist, but later confirmed by a radiologist to be pleural plaque. In some cases, only the calcified parts of the pleural plaque were segmented by the radiologists (N=5) (Figure 5.5A). In other, radiologists missed the pleural plaques completely (N=27). In N=2 cases, random voxels were segmented due to interpolation problems. In one case, a patient contained malignant tissue, which was segmented by the radiologist but correctly marked as non pleural plaque by the CNN model (Figure 5.5B). These cases will be adjusted by the most experienced radiologist. Of these cases, N=18 CT scans were of bad quality and further segmentation quality could be low. N=16 cases of insufficient quality of segmentation belonged to the test set. After (temporarily) excluding these cases from the test set, the main outliers were removed, as visualized on the log-log plot in Figure 5.3C. Consequently, the performance measures of the CNN model improved; The model now yielded a mean DSC of 0.69, median DSC of 0.73, and a stronger correlation (r=0.94) (Figure 5.3D). However, on visual inspection, it became apparent that the CNN model was by no means perfect; it segmented often foreign bodies, such as a pacemaker (Figure 5.5C).

The CNN model performed segmentation on the CT scans in the control dataset to test for systemic errors on healthy CT scans. Within these N=76 CT scans, it segmented a mean of 2.8 cm^3 and a median of 1.8 cm^3 PPV. After outlier analysis was performed on four cases (12.9-15.5 cm^3), it became apparent that the ribs were segmented in these cases. For one case, the CNN misclassified normal tissue as pleural plaque between the high density line in the breast and the rib (Figure 5.5D).

Patients that had diffuse fibrosis were excluded for correlation between the lung function and the volume of pleural plaques, since fibrosis is a confounding variable[37]. Diffuse fibrosis was defined as 5% or more fibrosis of the pleural surface, assessed by three independent pulmonologists. This yielded a dataset of N=203 patients with no fibrosis. The lung function

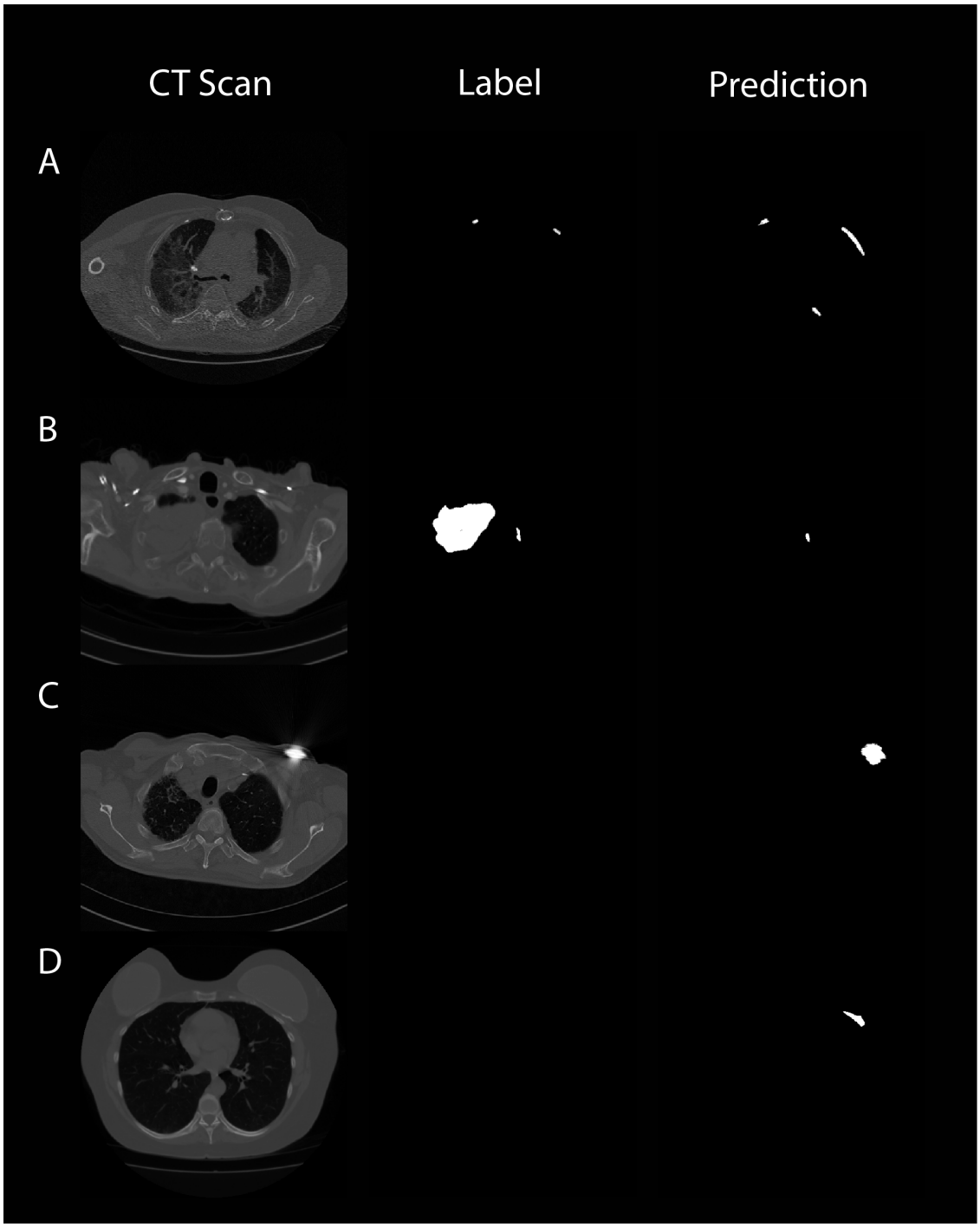

FIGURE 5.5: Visualization of interesting cases of the input (CT Scan), segmentation of the radiologist (Label), and the CNN segmentation (Prediction). (A) Case which shows that the segmentation of the radiologists only contains the calcified parts, where the CNN model segments more, but still not all the plaques. (B) Lung cancer case, where the malignant tissue was segmented by the radiologist, but not by the CNN model. (C) Pacemaker in the patient's body gets segmented by the CNN model as pleural plaque. (D) Case out of the control set, where the CNN model segments the tissue between the rib and the high density line in the breast.

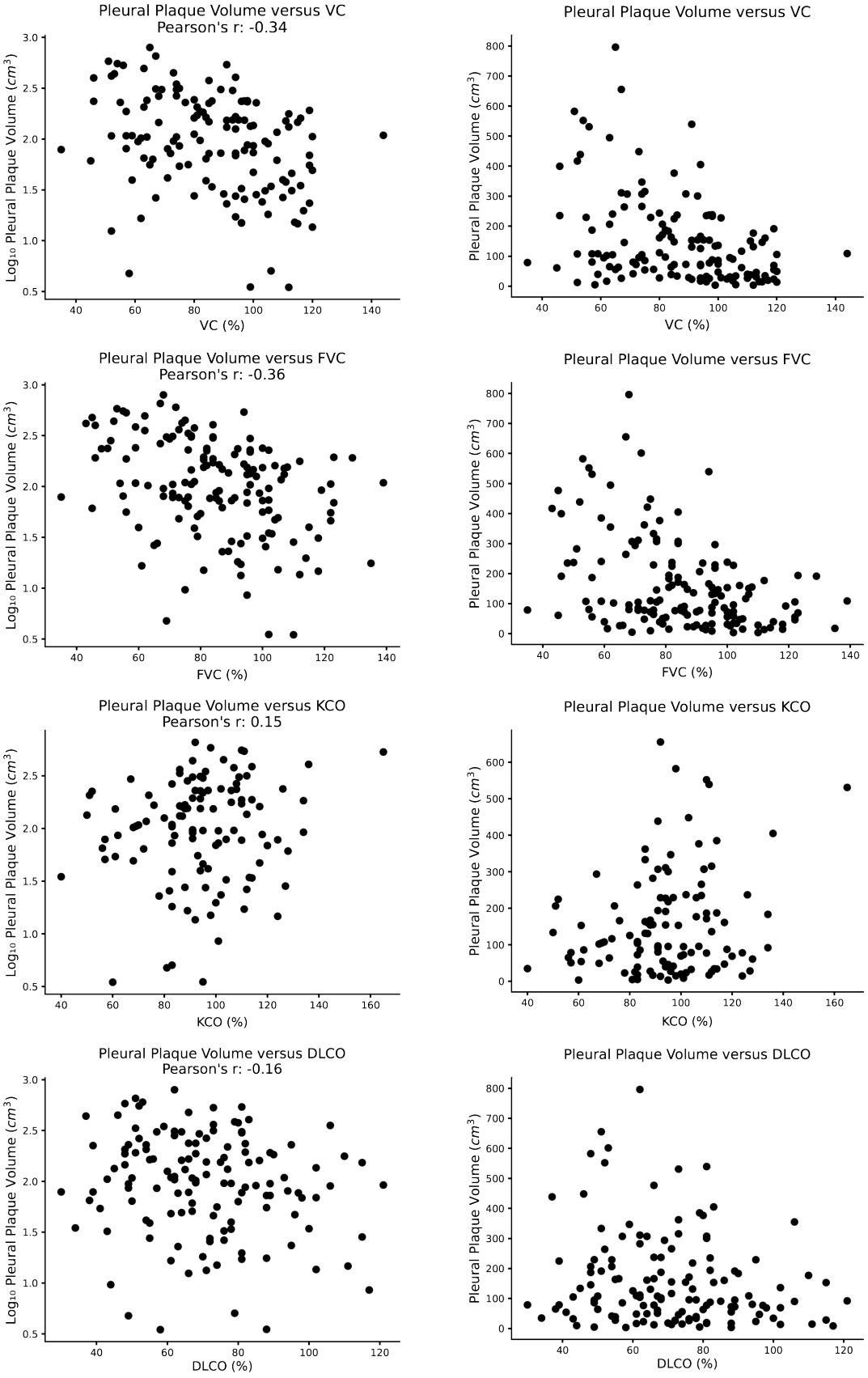

FIGURE 5.6: Visualization of the correlation between pleural plaque volume on the y-axis versus the recorded lung function parameters on the x-axis. The left column shows the relation on a logarithmic scale, while the right column shows the normal scale.

TABLE 5.1: The results of the statistical test performed on the VC and PPV.

Cut off (percentile)	PPV (cm^3)	N upper	N lower	Mean upper	Mean lower	p
12.5	25.2	118	17	84.6	97.6	0.018
25	40.6	101	34	82.7	96.6	<0.001
50	104.7	67	67	82.3	90.5	0.024
75	215.4	34	101	74.0	90.3	<0.001
87.5	307.0	17	118	68.0	88.8	<0.001

TABLE 5.2: The results of the statistical test performed on the FVC and PPV.

Cut off(percentile)	PPV (cm^3)	N upper	N lower	Mean upper	Mean lower	p
12.5	26.9	130	19	83.1	94.6	0.025
25	54.0	111	37	81.4	94.3	0.001
50	107.4	74	74	80.1	89.5	0.006
75	228.9	37	111	69.4	89.7	<0.001
87.5	343.9	19	130	64.1	87.6	<0.001

parameters available of these patients were VC (N=138), FVC (N=152), DLCO (N=137), and KCO (N=119).

We found a weak negative correlation between the VC and the plaque volume (r=-0.34), and between the FVC and the plaque volume (r=-0.36). After excluding the N=35 cases with insufficient quality of segmentation, the correlation got stronger with VC (r=-0.41) and FVC (r=-0.44). No correlation was found between the DLCO and the plaque volume (r=-0.16), and between the KCO and the plaque volume (r=0.15). The PPV versus each lung function parameter on normal and logarithmic scales are plotted in Figure 5.6. For the VC and FVC, we tested different cut off values in the distribution of the PPV (Table 5.1 and 5.2, respectively). All differences were statistically significant ($p<0.05$). However, when Bonferroni is applied, the threshold of significance drops to $p<0.01$, resulting in the non statistically significant differences of the 12.5th percentile cut-off of VC, and 12.5th and 50th percentile of FVC.

5.5 Discussion

In this study, we proposed an AI algorithm for the automatic segmentation of pleural plaques in patients with asbestos exposure. The resulting algorithm enabled us to study the correlation between pleural plaques volumes and lung functions in a large cohort of patients,

where we assessed the correlation between lung function and pleural plaques.

The 2D UNet architecture outperformed the 3D counterpart. While the 3D architecture enabled the AI model to learn spatial correlations over the longitudinal axis of the CT scan, it reduces the number of training examples by two orders of magnitude. Moreover, due to hardware limitations, CT scans had to be downsampled for the 3D architecture, leading to a loss of resolution. Both could be an explanation for the better results of the 2D architecture. The spatial correlation problem could be resolved with voxel resampling, where each voxel would get the same size. While the results suggest a merely moderate overlapping measure between the segmentations of the AI model and the one from the expert reader, the high correlation between the segmented volume of the CNN model and the expert readers indicates that the proposed CNN model can estimate the volume moderately well, and that systematic segmentation error might have been present. This systematic segmentation error is most likely due to the implementation of the Tversky loss, which favours overestimation of the volume over underestimation. This is also supported by the results found in the mean absolute error compared to the mean error. The model can segment a large portion of the existing plaques in the patient and can be used to propose a segmentation that the expert reader has to adjust, reducing the workload of the readers.

The correlation was weak for the VC and FVC, but the student t-test for these parameters between patients with a high and low volume of pleural plaques showed significant differences, probably generated by the large sample size. The DLCO and KCO showed no correlation and were not further assessed. One possible explanation could be that the pleural plaques impede the ability of patients to fully in- and exhale, reducing the VC and FVC. DLCO and KCO are gas exchange parameters and less affected by total air inhalation, where pleural plaques do not inhibit gas exchange itself[38]. To the best of our knowledge, our study is the first one that segmented pleural plaques by expert readers on a voxel basis in a cohort of over 100 patients and determined the correlation to lung functions. Compared to similar studies that found no correlation between PPV and lung function parameters[33,34], we decided on another methodology. First, we labeled the pleural plaques on a voxel basis instead of measuring the longest diameter in three axes. Second, the thresholds of PPV to test significance was higher in our study than in other studies. Our lowest cut off of the 12.5th percentile was 25.2 cm^3 or mL for VC, where a similar study defined the highest volume

group of plaques as >20 cm^3.

There are limitations to this study, namely that we could not correct for confounders in the correlation between the lung function parameters and the volume. The lung function parameters are already corrected for height, weight, age, gender, and race. While we did exclude fibrotic patients, lung function parameters for other confounders (e.g. smoking) could not be corrected since that information was not known. Furthermore, not all labels were of sufficient quality. In some cases, pleural plaques were missed by the radiologists, or only the calcified parts of the pleural plaques were segmented. In other scans, pleural plaques were visible to the radiologist but hardly possible to segment due to the bad quality of the CT scan. At last, certain functions to aid the segmentation process yielded segmentation artifacts, which resulted in a positive label for non pleural plaque tissue. Currently, N=35 cases were found, but there are probably more cases in the CT scans with > 0.40 DSC that contain missing segmentations as well. Revision of all insufficient segmentations is necessary to yield reliable results. Through exclusion, we introduce a bias where all hard to segment cases are removed. A consequence could be that the CNN model seems to achieve high performance on the test set, but if a hard to segment case is filed during a hypothetical compensation process, it might not achieve the shown performance.

5.6 Conclusion

In this study, we showed a correlation between pleural plaque and lung function parameters. We developed an AI model for the automatic segmentation of the pleural plaques in CT scans to estimate the volume. The segmentations were quantitatively and qualitatively assessed and showed a high correlation to the segmentation of expert readers. The AI model can be used to decrease the workload for the expert readers and to continue to expand the dataset to get a larger sample size. The statistical tests between the lung function and the pleural plaque volume suggest that patients with a higher volume of plaques have a worse lung function.

6

General Discussion

6.1 Clinical Relevance

Currently, AI models often lack clinical usefulness and validation[39]. In this thesis, we focussed on the usability of our models in the clinical environment to assist the clinicians. Therefore, we did not strive for the best score of chosen metrics, but rather aimed for results which contain explainability and potential to be clinically implemented. In the classification of asbestosis, we developed a model that yielded tight distributions of predictions compared to the number of pulmonologists without any outliers. By adding the DLCO manually, we retain the explainability of the lung function test and prevent the model from becoming overconfident. This enables us to choose cut-offs of probabilities where we can be confident the predictions are right, and are, therefore, able to clinically implement such models. Moreover, it reduces the 'black box' part of the algorithm where we do not know exactly why certain scans are labeled as positive or negative for asbestosis. For the automatic segmentation of the pleural plaques, the clinicians stated that the model should not underestimate the pleural plaques in the patient, where the predicted volume is much lower than the real volume. Therefore, we adapted the model with a certain loss function that penalized missed pleural plaques harder than the segmentation of voxels that are not pleural plaques. The result was a systematic overestimation, and the model would rarely underestimate the volume of the pleural plaques. Furthermore, the AI model segmented pleural plaques that were missed by the radiologists, but confirmed to be pleural plaque after the prediction of the AI model. We believe that these cases show the potential of strengthening clinical decision making if the model is adapted to assist the clinicians in their needs. The current preparation of the clinical validation of the asbestosis classification model supports this vision.

6.2 Limitations

The work in this thesis contains limitations. First, the labels on which the models are trained are not the real ground truth. For the asbestosis labels, pulmonologists had to diagnose the patients without invasive methods and even seeing the patient. 24% of the patients were not diagnosed with unanimity, where the 2 out of 3 pulmonologists out of the total pool were decisive in the diagnosis. One could argue that if other pulmonologists were selected for

the application of that patient, the diagnosis would be different. For the segmentation of the pleural plaques, the segmentations of the different fellow radiologists were only reviewed if they marked it as a hard case. Due to inexperience in estimating consistency of segmentation work and reported findings in literature, it was assumed that the inter-observer variability was low. However, after outlier analysis was performed on the CT scans that yielded a low DSC score, it became apparent that there was inconsistency between segmentations. This was partly due to the bad quality of several CT-scans, which is the next limitation. Some CT-scans date back to the early 2000s, when imaging protocols were different and hardware was not of the same quality as of today. This leads to lower resolution, making it harder to distinguish structure for both radiologists as well as AI. Third, the automatic segmentation of the lungs performs well if the lungs are healthy, but the software can fail when fibrosis is present within the lungs. Therefore, visual inspection after automatic lung segmentation is always needed, with the additional manual segmentation if the software failed to correctly segment the lungs. Finally, the general limitation of AI, which is the incapability of explaining why it predicts certain outcomes. We were able to estimate why certain predictions were made with knowledge about the specific model, in combination with the saliency maps, but an AI model will always lack explainability comparing to a human.

6.3 Recommendations

For future work, we advocate standardization of the segmentation process. Currently, only one radiologist is involved in processing one CT scan. We think that a revision pipeline could be more robust, where several radiologists can segment cases, and one or two experienced radiologists check the segmentations for sufficient quality. Furthermore, a spreadsheet where each reader ranks the quality of the CT scan and the perceived quality of his or her segmentation could be beneficial.

Retrieving missing lung function parameters of patients could lead to more reliable results, and would be a relatively straightforward method to extend this research. Acquiring additional lung function parameters would be interesting as well, as they could lead to new insights into asbestos-related diseases. Furthermore, the department of Pulmonology at the

NKI is currently retrieving metadata of the patient that filed an application, which could possibly lead to new patterns to discover.

To counter the mandatory downsampling of 3D CT scans to fit into the memory, and consequently the loss of resolution, higher capacity GPUs and CPUs were needed to process the full resolution CT scan. The department of Pulmonology and Radiology of the NKI have been so generous to invest in state-of-the-art GPUs and CPUs to realize this. Currently, models are being trained on higher resolution than have been described in this thesis, and the developments of these models should continue to achieve the maximal possible performance.

Bibliography

1. Świątkowska, B., Szeszenia-Dąbrowska, N. & Wilczyńska, U. Medical monitoring of asbestos-exposed workers: experience from Poland. en. *Bull. World Health Organ.* **94,** 599–604 (Aug. 2016).
2. Peacock, C, Copley, S. J. & Hansell, D. M. Asbestos-related benign pleural disease. en. *Clin. Radiol.* **55,** 422–432 (June 2000).
3. Roach, H. D. *et al.* Asbestos: When the Dust Settles—An Imaging Review of Asbestos-related Disease. *Radiographics* **22,** S167–S184 (Oct. 2002).
4. Norbet, C., Joseph, A., Rossi, S. S., Bhalla, S. & Gutierrez, F. R. Asbestos-related lung disease: a pictorial review. en. *Curr. Probl. Diagn. Radiol.* **44,** 371–382 (July 2015).
5. Greillier, L. & Astoul, P. Mesothelioma and asbestos-related pleural diseases. en. *Respiration* **76,** 1–15 (May 2008).
6. Paris, C *et al.* Pleural plaques and asbestosis: dose- and time-response relationships based on HRCT data. en. *Eur. Respir. J.* **34,** 72–79 (July 2009).
7. Kerper, L. E. *et al.* Systematic review of pleural plaques and lung function. en. *Inhal. Toxicol.* **27,** 15–44 (Jan. 2015).
8. Maxim, L. D., Niebo, R. & Utell, M. J. Are pleural plaques an appropriate endpoint for risk analyses? en. *Inhal. Toxicol.* **27,** 321–334 (June 2015).
9. Kamp, D. W. & Weitzman, S. A. The molecular basis of asbestos induced lung injury. en. *Thorax* **54,** 638–652 (July 1999).
10. Kamp, D. W. Asbestos-induced lung diseases: an update. en. *Transl. Res.* **153,** 143–152 (Apr. 2009).
11. Liu, G., Cheresh, P. & Kamp, D. W. Molecular basis of asbestos-induced lung disease. en. *Annu. Rev. Pathol.* **8,** 161–187 (Jan. 2013).
12. Raeve, H. d. *et al. Observer variation in computed tomography of pleural lesions in subjects exposed to indoor asbestos* 2001.
13. Miller, A., Warshaw, R. & Nezamis, J. Diffusing capacity and forced vital capacity in 5,003 asbestos-exposed workers: Relationships to interstitial fibrosis (ILO profusion score) and pleural thickening. *Am. J. Ind. Med.* **56,** 1383–1393 (2013).
14. Di Muzio, B. *HRCT chest, Radiopedia*
15. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. en. *Nat. Rev. Cancer* **18,** 500–510 (Aug. 2018).
16. Nair, V. & Hinton, G. E. *Rectified Linear Units Improve Restricted Boltzmann Machines* Jan. 2010.
17. Maas, A. L., Hannun, A. Y. & Ng, A. Y. *Rectifier nonlinearities improve neural network acoustic models* in *Proc. icml* **30** (2013), 3.
18. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv: `1502.01852 [cs.CV]` (Feb. 2015).

19. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv: 1502.03167 [cs.LG] (Feb. 2015).
20. Shi, W. *et al.* Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. arXiv: 1609.05158 [cs.CV] (Sept. 2016).
21. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. arXiv: 1312.6114v10 [stat.ML] (Dec. 2013).
22. Armato 3rd, S. G. *et al.* The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. en. *Med. Phys.* **38,** 915–931 (Feb. 2011).
23. LaLonde, R. & Bagci, U. Capsules for Object Segmentation. arXiv: 1804.04241 [stat.ML] (Apr. 2018).
24. Wang, Z., Yuan, H. & Ji, S. Spatial Variational Auto-Encoding via Matrix-Variate Normal Distributions. arXiv: 1705.06821 [cs.LG] (May 2017).
25. Zhao, S., Song, J. & Ermon, S. Towards Deeper Understanding of Variational Autoencoding Models. arXiv: 1702.08658 [cs.LG] (Feb. 2017).
26. Akrami, H., Joshi, A. A., Li, J., Aydore, S. & Leahy, R. M. Robust Variational Autoencoder. arXiv: 1905.09961 [stat.ML] (May 2019).
27. Asbestprotocollen, G. C. *Advies Protocollen Asbestziekten Asbestose* nl. 1999.
28. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. arXiv: 1512.03385 [cs.CV] (Dec. 2015).
29. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. arXiv: 1912.02911 [cs.CV] (Dec. 2019).
30. Martinez, M. & Stiefelhagen, R. *Taming the Cross Entropy Loss* in *Pattern Recognition* (Springer International Publishing, 2019), 628–637.
31. Myers, R. Asbestos-related pleural disease. en. *Curr. Opin. Pulm. Med.* **18,** 377–381 (July 2012).
32. Dournes, G. *et al.* 3-Dimensional Quantification of Composite Pleural Plaque Volume in Patients Exposed to Asbestos Using High-resolution Computed Tomography: A Validation Study. en. *J. Thorac. Imaging* **34,** 320–325 (Sept. 2019).
33. Cha, Y. K., Kim, J. S. & Kwon, J. H. Quantification of pleural plaques by computed tomography and correlations with pulmonary function: preliminary study. en. *J. Thorac. Dis.* **10,** 2118–2124 (Apr. 2018).
34. Çoşğun, I. G., Evyapan, F. & Karabulut, N. Environmental asbestos disease: pleural plaque volume measurement with Chest Tomography is there a correlation between pulmonary function? en. *Sarcoidosis Vasc. Diffuse Lung Dis.* **34,** 336–342 (Apr. 2017).
35. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv: 1505.04597 [cs.CV] (May 2015).
36. Salehi, S. S. M., Erdogmus, D. & Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. arXiv: 1706.05721 [cs.CV] (June 2017).
37. Plantier, L. *et al.* Physiology of the lung in idiopathic pulmonary fibrosis. en. *Eur. Respir. Rev.* **27** (Mar. 2018).
38. Shih, J.-F. *et al. Asbestos-induced Pleural Fibrosis and Impaired Exercise Physiology* 1994.

39. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. en. *BMC Med.* **17,** 195 (Oct. 2019).