VRIJE UNIVERSITEIT AMSTERDAM

RESEARCH PAPER

Lung nodule detection by using

Deep Learning

Author:
Thomas HEENEMAN

Supervisor:
Dr. Mark HOOGENDOORN

Msc. Business Analytics

Department of Mathematics

Faculty of Sciences

January 2018

Preface

This research paper is written as compulsory part of the Business Analytics Master's program at the "Vrije Universiteit" in Amsterdam. It is written with the aim of performing research in the field of business analytics and describe the problem and found solutions in a scientific manner.

I would like to thank my supervisor Dr. Mark Hoogendoorn for guiding me through this research.

Abstract

Lung cancer is one of the most common types of cancer worldwide. It is also one of the deadliest types of cancer. However, research indicates that early detection of lung cancer significantly improves chances of survival. Using computed tomography scans of the human lungs, radiologists can detect dangerous nodules in early stages. When more people adopting for these scans, the workload on the radiologists rises. Computer aided detection systems that automatically detect these nodules can support radiologists and reduce their workload. However, current CAD systems often have a low sensitivity and suffer from a high number of false positives. Therefore, the aim of this research is to develop a CAD system that is able to localize as many nodules as possible, while keeping the number of false positives low. To localize the nodules in the three-dimensional space of a scan, a quantile approach over the cross-sectional slices is applied to retrieve a two-dimensional slice. Next, a sliding window method on the slice is used to obtain the (x, y) co-ordinates of the nodule. A two-dimensional convolutional neural network is used as classifier of these windows. The developed CAD system is able to detect and localize 60.1% of all the nodules with an average number of 2.1 (1.5%) false positives per slice. The system is evaluated using all available examples and by simulating the scenario of a radiologist, as much as possible. In addition, three different types of neural network structures for this CAD system are tested and compared. With 95% confidence we can conclude that deeper neural networks decrease the false positives significantly. In this research the publicly available LIDC-IDRI database has been used. This database contains CT scans of human lungs with annotations from radiologists.

Keywords: lung, cancer, computed tomography, CT scan, neural networks, convolutional neural networks, CNN, deep learning, Computer Aided Detection, CAD, Watershed, LIDC-IDRI

Contents

1 Introduction and Related Work

In this research a Computer Aided Detection (CAD) system has been developed. This system is able to localize lung nodules in three-dimensional CT scans of the human lungs in order to detect possible cancer tissue in early stages. These systems are needed to support radiologists and to provide a second-opinion.

1.1 Lung Cancer

Worldwide, lung cancer is the most common type of cancer [1], contributing with 13% to the total number of cancer cases (skin-cancer excluded). Of the three most common types of cancer, lung-, breast- and prostate cancer, the death rate and probability of dying is the highest with lung cancer [2]. In The Netherlands lung cancer is in 2016 the fourth most common type of cancer, with a contribution of 12% for men and 11% for women [3]. Research indicates that early detection of lung cancer significantly increases the survival rate [4].

1.2 Detecting lung nodules using CAD systems

A lung nodule refers to a range of abnormalities considered as small, round opacity, roughly spherical, restricted on abnormal tissue. To detect lung nodules radiologists use chest computed tomography (CT) scans. CT scans can detect smaller nodules at earlier stages than chest radiography scans can [5]. Furthermore, according to Awai et al. [6], the detection rate of lung cancer is $2.6 - 10$ times bigger using a CT scan than with analogue radiography. The cross-sectional images of the lungs, generated during a CT scan, can be reformatted in multiple planes. This results in a three-dimensional representation of the lungs. Inspection of all these cross-sectional images is a time-consuming job for radiologists. With more people adopting for early CT scans, the workload on radiologists will increase considerable.

Computer Aided Detection (CAD) systems for detecting and localizing lung nodules within CT scans is a solution to reduce this increasing workload on radiologists. CAD systems aim to assist in the detection and classification of diseases through a second opinion for radiologists [7].

The goal of a CAD system is to achieve high sensitivity, i.e. detecting as many of the nodules present in a scan. However, current CAD systems produce many false positives, i.e. incorrectly detecting a spot as a nodule, when it's not. There is a trade-off between high sensitivity and a high number of false positives. The performance of these systems varies. In a comparative study [8], Jacobs et al. compares the performance of three commercial CAD systems for lung nodule detection. Their performances on sensitivity are 62%, 74% and 82%, while the number of false positives are 3.1, 3.8 and 4.0 per scan, respectively.

1.3 Related Work

Over the years many different types of CAD systems for detecting and localizing lung nodules have been developed and tested in this research field. Since techniques and computing power have improved over the years, so have the CAD systems.

2000 - 2010
Gurcan et al. [9] proposed a multistage CAD system that segments the lung regions of interest using k-means clustering and detect lung nodules using rule-based classifiers and linear discriminant analysis (LDA). They report a sensitivity of 84% and an average of 1.74 false positives per slice.

McCulloch [10] also proposed a multistage CAD system in which every stage captures different anatomic information and lung structures. Their system produces 8.3 false positives per scan. Ge et al. [11] extracted three-dimensional (3D) features from volumes of interest to detect nodules. Within these features they obtain a sensitivity of 80% with around 0.77 false positives per section. Localizing a nodule on an image or scan means getting the exact x, y- and z-coordinates. With that in mind, Lee et al. [12] proposed a sliding window method, in which moving windows across the image are classified. They report a sensitivity of 100% and 1.4 false positives per slice, tested on 20 selected slices.

2011 - now
With the breakthrough of deep convolutional neural networks (CNN) in image classification with the AlexNet [13], developed by Alex Krizhevsky, many computer vision tasks are being done by CNN's. Ever since, the technique is often referred to as "deep learning". They also have found great success in medical imaging applications as Kayalıbay proved in his research [14]. Therefore, more recent CAD systems for detecting lung nodules use CNN architectures. Since CNN's often need many examples to generalize well, Yang et al. [15] investigated whether data augmentation improved performance on a 2D CNN. Their main conclusion is that data augmentation does improve performance on their CNN. In the recent research of Hamidian et al. [16] a 3D CNN is investigated, which reaches a sensitivity of 80% and 22.4 false positives per scan.

CT scans are 3D volumes and are usually anisotropic. Deep convolutional networks have also been shown to perform well in 3D segmentation as shown by Milletari et al. [17]. They also have been successfully adapted from 2D to 3D by Cicek et al. [18].

In most related studies, CAD systems are evaluated using sensitivity and the number of false positives per slice or CT scan. False positive rate or –percentage is another variant that is often used.

1.4 What research is missing

In computer vision in general, and lung nodule detection specifically, the performances of CNN's are very promising. In related studies in this field, most CNN's perform a binary task: does the scan or slice contain a nodule yes/no. However, with this approach, no exact coordinate location in the x, y, z-plane is found. The sliding window method of Lee et al. [12] solves this problem partially, because a sliding window method can provide x, y coordinates of a nodule on a 2D slice. In his research, a Random Forest classifier was used and not a CNN. In addition, most research does not pay attention to retrieving the z-coordinate of a nodule. One could argue to investigate every slice, but with a different number of slices for different CT scan machines and the fact that subsequent slices in a CT scan contain very similar information, this is not a strong argument. Furthermore, many proposed CAD systems are evaluated using only a (selected) subset of the available data. This does not realistically reflect the scenario of the radiologist.

1.5 This research

This paper presents a CAD system that is a combination of a 2D CNN sliding window method and a quantile approach to localize the nodules in the x, y, and z-space. The proposed CAD system will be evaluated as realistically as possible, by evaluating on slice level using all available CT scans. Furthermore, a comparative study has been carried out to investigate the effect of deeper CNN's on the performance of this CAD system.

1.6   Structure of this paper

This paper is divided into seven parts. In this first section, an introduction to the problem has given along with an overview of related research in this field. Hereafter, the available data set is presented in section 2. In section 3 the overall CAD system and its methods are outlined. This section is followed by evaluation methods and the setting for the experiment to test this system. In the fifth part the results of this research are presented and these results are put into perspective and discussed in part 6. Finally, a conclusion on this research is given in section 7.

2 Data

To conduct this research and to develop a CAD system that detects and localizes lung nodules, examples of nodules are needed. Examples of nodules that have been used are from the Lung Image Database Consortium (LIDC-IDRI) image collection. This database consists of diagnostic- and lung cancer screening thoracic CT scans with annotated lesions [19]. It contains scans of high-risk patients with independent annotations of four experienced radiologists. A final annotation of a lesion is made when 3 of the 4 radiologists independently agree on that lesion. Lesions are classified into three categories: nodule > 3 mm, nodule <= 3 mm and non-nodule >= 3 mm. In this research only the nodules > 3 mm are considered, because many related studies have done the same. With this selection, the total data set contains scans of 799 patients. The LIDC-IDRI database is publicly available which makes this research reproducible.

2.1 CT Scans

All CT scans are in the MetaImage (mhd/raw) format, which is a text-based tagged file format for medical images. Each .mhd file is stored with a separate .raw file that contains all the pixel-/ voxel data. Each CT scan consists of a number of cross-sectional slices of the chest. In the data set used for this research, the number of slices ranges from 95 to 733. In this research this is called the z-dimension of a scan. The number of slices varies with the slice thickness of the scan. The thicker each slice, the less slices in a scan, and vice-versa. Every cross-sectional slice is a two-dimensional image of 512 by 512 pixels and these are called the x- and y-dimensions respectively. Every pixel on a slice contains a Hounsefield Unit (HU) value [20]. HU values are a measure of radio density and are commonly used in CT scans to express values in standardized and convenient form. HU values range from [-3024; 3071]. Different substances in a human body produce different HU values. Furthermore, each CT scan has meta data: slice thickness, number of slices, the spacing between the voxels and the origin.

2.2 Annotations

Every scan was inspected by thoracic radiologists in a two-phase annotation process. In the initial phase, each radiologist independently marked lesions belonging to one of the following three categories: nodule >= 3 mm, nodule <= 3 mm and non-nodule >= 3 mm. In the second phase, each radiologist reviewed their own marks along with the marks of the other anonymized marks by the other three radiologists. A final annotation is marked when 3 of the 4 radiologists agree on a lesion.

For a complete description of the annotation process, please read the corresponding paper [21].

3 Methods

In this section, the proposed CAD system is presented. First, a general overview of the system with its components is given. This is followed by a detailed description of each of the components and clarifications for design choices. In the end of this section, the other two neural network structures are presented for the comparative study.

3.1 Overall CAD system

The goal of this CAD system is to localize lung nodules in the three-dimensional space of a CT scan. To go from a raw scan to the localization, the following pipeline of steps have been used:

1. Choose cross-sectional slices at quantiles

2. Segment the lungs from the surrounding parts of the body using Watershed algorithm

3. For each slice with segmented lungs obtain slice grid using sliding window method

4. Binary classification of each window using a 2D convolutional neural network

5. Final prediction and localization

3.2 Quantile approach

Although the number of slices in all the CT scans varies, human lungs do not vary much in size. Data analysis performed on this data set showed that the average size of the lungs in this data set is 315.89 mm, so 31.59 cm. Due to the earlier mentioned difference in slice thickness of the scans, 2D slices are obtained at quantiles of the scan. Each quantile of a scan, represents roughly the same area as the same quantile in another scan. This can be explained by the following example. If scan A has a slice thickness of 2mm and 150 slices and scan B has a slice thickness of 1 mm and 300 slices, then the 10'th quantile of scan A is at slice 15 at an height of 30 mm. The 10'th quantile of scan B is slice 30, also at an height of 30 mm.

Since every nodule to be considered is larger than 3 mm and every quantile represents on average 3.1589 mm, every quantile a cross-sectional slice is obtained. Quantiles range from 0 to 1, and 1 quantile means 0.01 in this [0,1] range. The smallest nodule in this data set has a diameter of 3.2532 mm. This diameter is measured while inspecting the 2D slices during the annotation process and is confirmed by at least 3 of the 4 radiologists. Annotated nodules are visible on multiple subsequent slices and have circular shapes. This suggests, and it is assumed in this research, that most nodules are spherical in shape. This implies that the diameter of a nodule in the 2D x, y –plane is the same in the z-plane. Obtaining a 2D slice at every 3.1589 mm implies that every nodule in this data set could be found.

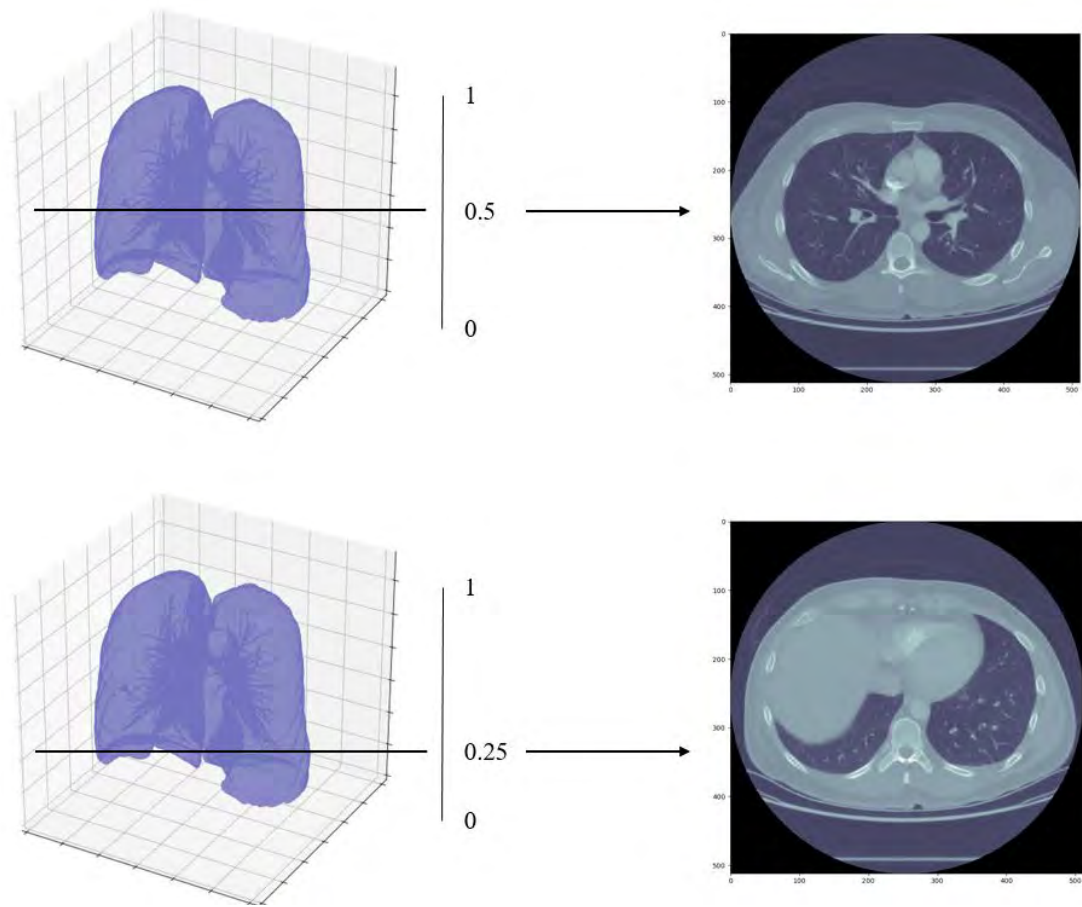Please see Fig. 1 for a graphical representation of the proposed quantile approach.

Figure 1: Extracting 2D slices at every 0.01 quantile

3.3 Segmentation of the lungs using Watershed

At the obtained 2D slices, the lungs are surrounded by other parts of the human body. As all the nodules are only inside the lungs, these other parts can be filtered out. Without these other parts a possible image classifier can only focus on the lungs and cannot be distracted by the other, irrelevant parts. During this research, three different approaches have been tested to segment the lungs from its surroundings: Open Source Computer Vision (OpenCV) [22], Python Imaging Library (PIL) [23] and the Watershed algorithm [24]. It was found out that the Watershed algorithm performed best on this lung segmentation task.

3.3.1 Watershed algorithm in general

The name of this algorithm refers to a geological watershed or ridge that divides areas by different water systems. Any gray scale image, where each pixel has one value, can be seen as a topographic surface where high values represent peaks or hills and low values represent valleys. The algorithm starts by 'flooding water' around the topographic surface of the image. When the water merges the peaks, barriers are built to prevent this merging, until all peaks are under water. The resulting barriers give the segmentation of peaks and valleys in the image. This is illustrated in Fig.2.
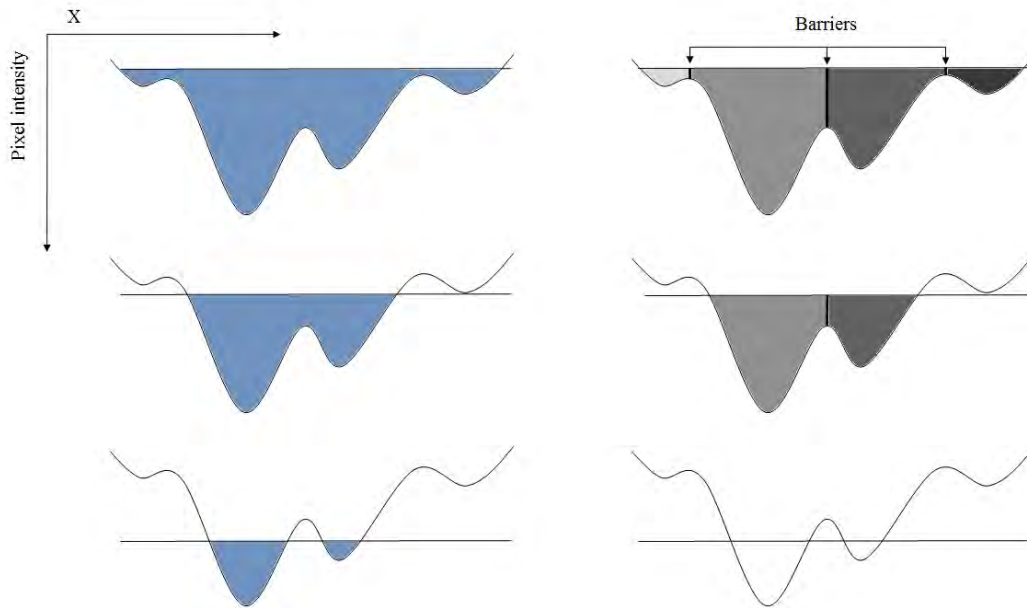

Figure 2: Watershed illustrated by floating water and creating barriers

10

3.3.2 Watershed applied to human lungs

To segment the lungs from its surrounding, a slight variant of the Watershed algorithm is used. In his paper [25], R. Shojaii presents a method to segment human lungs by using internal and external markers. The markers are combined with the image gradient and the Watershed algorithm, to find the lung borders. A rolling ball filter is used to smooth the contour and fill the cavities, while preserving the original borders. A little margin around the borders is kept to make sure no juxtapleural nodules are cut away from the edge. Fig. 3 illustrates how these steps are applied to the lungs. Example code to use this algorithm on CT scans is found on the website of Kaggle [26]. I would hereby thank the author of the article, Ankasor, for sharing the code.
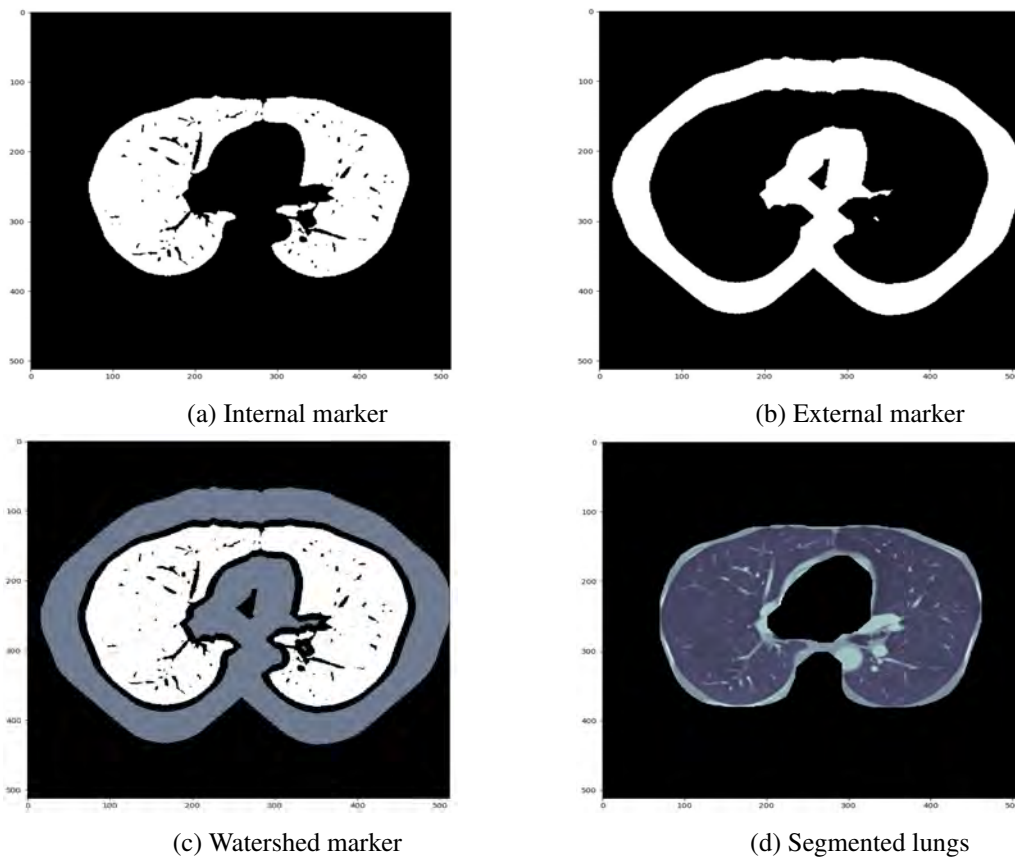

(a) Internal marker

(b) External marker

(c) Watershed marker

(d) Segmented lungs

Figure 3: Multiple stages of applying Watershed algorithm to human lungs

3.4 Sliding window method

Localizing a nodule on a slice means that the exact location of a possible nodule needs to be found. A location in a 2D plane of a slice is a set of (x, y) coordinates or a range of x- and y-coordinates in which the nodules lies. To get to coordinates, a sliding window method is applied on each of the 2D slices. A sliding window is a region, smaller than the concerning slice, which is moved across the slice. To be precise, on the 512 by 512 pixels slice, a sliding window of 64 by 64 pixels is used. Furthermore, overlay of windows is used to deal with the problem of having nodules partly on a window. Please see Fig. 4 for a graphical representation of this method.
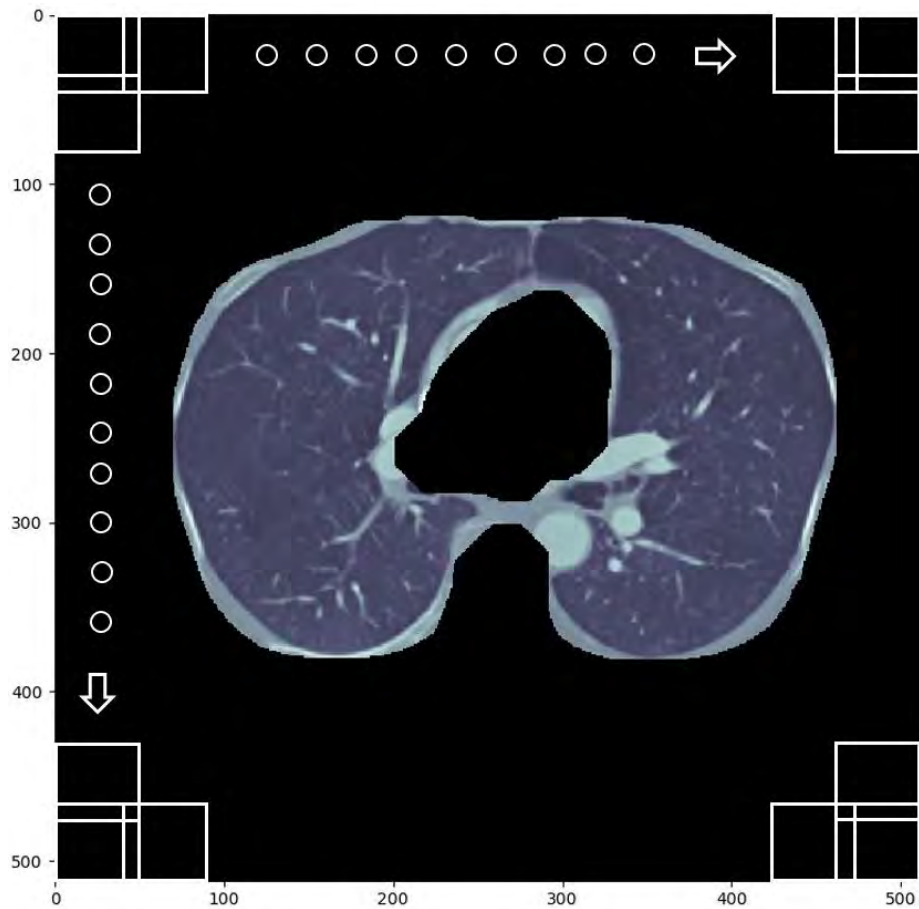

Figure 4: Sliding window method

3.4.1 Window size

Data analysis showed that the median and mean size of a nodule are 16 and 20 pixels in diameter respectively. The minimum diameter size of a nodule in this data set is 3.2532 mm or 8 pixels and the maximum size of a nodule is 32 mm or 79 pixels. With a window size of 64, 99.58% of all the nodules fit inside this window. Even 95% of all nodules are less than 19 mm or 46.5 pixels. Furthermore, 64 fits exactly 8 times in the [0, 512] range. However, using a sliding window method, the nodules won't always be perfectly located in the centre of the window. This relates mainly to the larger nodules. There can be windows with half a nodule, or even a quarter of a nodule. During this research it was found out that these nodules where hard to detect. Therefore, overlay of the windows is used, so that these nodules are more likely to fit in a whole window, which makes it easier to detect them.

3.4.2 Overlay

Overlay is used to deal with the problem of nodules that are half or quarterly located on the sliding windows. Again, it was desirable to fit at least 95% of the nodules in a window. That means all nodules equal or smaller than 46.5 pixels. In the worst case scenario, a nodule can be precisely on half of two windows. Then, an overlay of 23 or 24 pixels would be sufficient to get the whole nodule in a window. Therefore an overlay of 23 pixels is used. By creating 12 windows in total, 4 new windows are added and so are the 4 times 64 pixels. This means 11 times overlay of which 8 of them use 23 pixels and 3 overlays use 24 pixels. In previous research of Wan [27], it was found that nodules attached to the thoraric wall, the so called juxtapleural nodules, are the hardest to detect. Therefore, overlays of 24 pixels are used in the 4th, 6th and 8th overlay moments – the moments which cover most of the thoraric walls. See Fig.5 and Fig. 6.
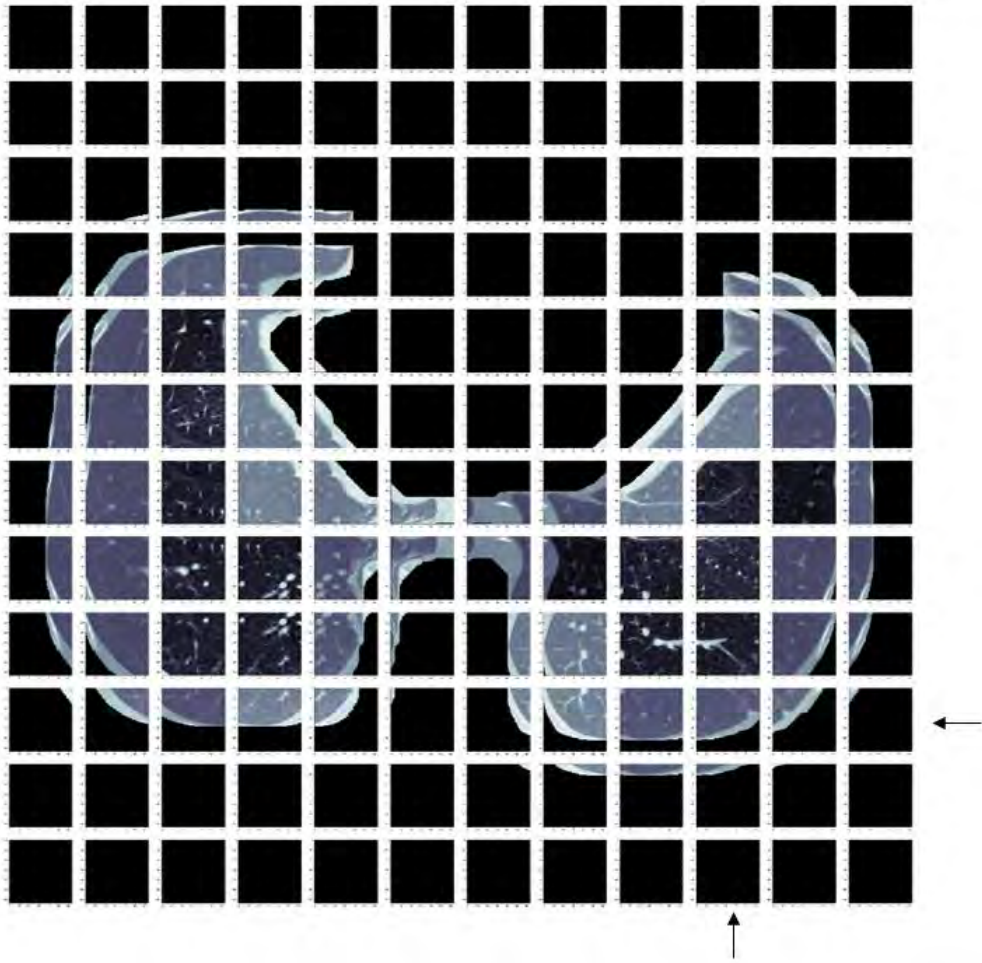
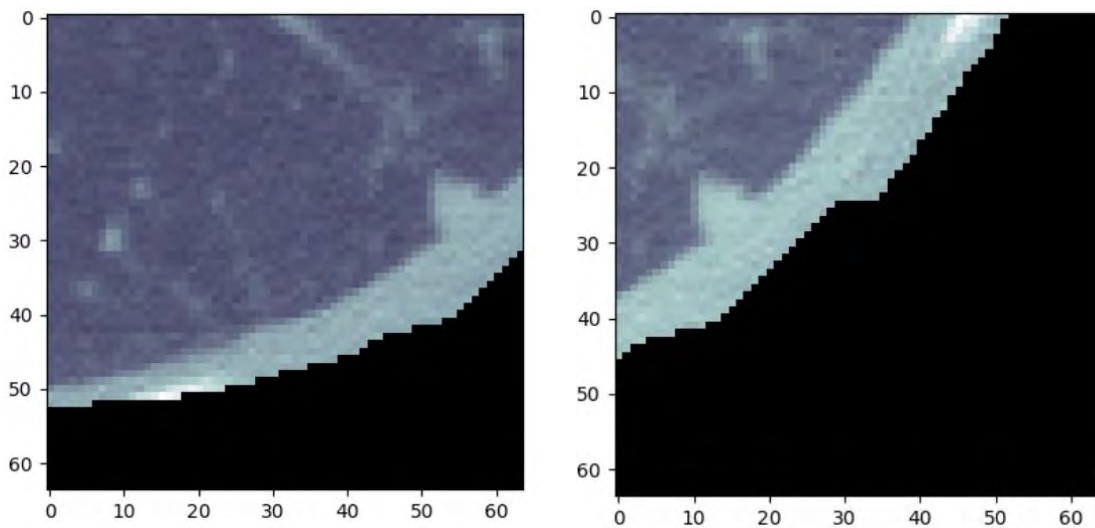Figure 5: Sliding window method with overlay

Figure 6: Nodule fully visible on both adjacent windows

3.5 CNN classifier

For this binary classification task there has been chosen for a two-dimensional convolutional neural network (CNN). The CNN model used in this research receives 64 by 64 pixel windows as input and predicts a binary 1/0 variable, indicating a positive or negative classification of the window. Its architecture is as follows:

1. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a Rectified Linear Unit (ReLu) as activation function

2. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

3. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

4. Max pooling with pool size [2, 2]

5. Dropout with a probability of 0.25

6. Convolutional layer of 64 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

7. Convolutional layer of 64 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

8. Max pooling with pool size [2, 2]

9. Dropout with probability of 0.25

10. Dense layer with 256 units of neurons and L2 regularization

11. Dropout with a probability of 0.5

12. Fully connected layer with a Sigmoid activation function

Training this CNN is done in Python by using the Keras package with Tensorflow as back-end. Every model was trained in 25 epochs after which loss was converged. Updating the weights was done after each epoch using batch-wise training with a batch size of 64 examples. For the loss function binary cross-entropy is used since it is a binary classification problem.

Furthermore, the adaptive gradient descent algorithm Adam [28] was used to update the weights of the network during the training process. Metrics to evaluate the model during the training process are accuracy and sensitivity. After each epoch, the two evaluation metrics are calculated on the whole validation set, so not batch-wise.

In this architecture the same filter size is used in each convolutional layer. Previous research of Yang et al. [15] showed that the same filter size of 3 is successful for this lung detection task. Since many of the training examples are highly correlated, heavy dropout is used to prevent overfitting. In dropout, the output of a hidden neuron is set to zero with a specified probability. Therefore, it does not contribute any more to forward pass and in back-propagation. The Sigmoid activation function is used in the fully connected layer, because this worked best for this binary classification task.

3.6 Other architectures

After numerous tests, the above stated architecture performed best for the detection task. To compare the performance of this CNN and investigate whether a deeper network structure has influence on the performance, two other networks are trained and tested. These two networks, indicated as "small" and "intermediate", have both a few layers stripped of the proposed network, the "deep" network.

The "intermediate" network:

1. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a Rectified Linear Unit (ReLu) as activation function

2. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

3. Max pooling with pool size [2, 2]

4. Dropout with a probability of 0.25

5. Convolutional layer of 64 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

6. Max pooling with pool size [2, 2]

7. Dropout with probability of 0.25

8. Dense layer with 256 units of neurons and L2 regularization

9. Dropout with a probability of 0.5

10. Fully connected layer with a Sigmoid activation function

The "small" network:

1. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a Rectified Linear Unit (ReLu) as activation function

2. Convolutional layer of 32 filters with size [3, 3], stride of 1, no padding and a ReLu as activation function

3. Max pooling with pool size [2, 2]

4. Dropout with probability of 0.25

5. Dense layer with 256 units of neurons and L2 regularization

6. Dropout with a probability of 0.5

7. Fully connected layer with a Sigmoid activation function

3.7 Model input

For training the classifier, positive and negative examples are needed for the training process. In this research an example is a window of 64 by 64 pixels. There are various ways to extract these examples from the raw data, which will be outlined in this section. In addition, since there is not much data and previous research indicates it improves a CAD system, data augmentation is used to enrich the data set even more.

A balanced training set was created with approximately the same amount of positive and negative examples. In this research, a positive example is a window that completely contains a nodule. A negative example is a window that does not contain a nodule or any part of it.

3.7.1 Training examples

Positive examples

Positive examples are created in two ways: artificially placing the nodule at a fixed position in a window and naturally, in which a nodule is located in a window using the sliding window method. This second method creates more natural variance in the position of the nodule on a window. Since each nodule in this data set is larger than 3 mm in spherical diameter, and the maximum slice thickness in this data set is 2.5mm, a nodule is always visible on multiple subsequent slices. Therefore, for each nodule the centred z-slice is taken along with the slice above and below, where the nodule is also visible. In this process of taken extra slices, the size of the nodule and slice thickness of the scan are taken into account to make sure each example slice really contains the nodule. By using different slices for the same nodule, a bit of variance in the appearance of the nodule is extracted. Besides, after the first model was trained, the false negatives serve as positive input for the later models. These FN's are only extracted from the training data.

Fixed positions

Artificial created windows with positive examples have the same nodule at fixed positions of the window. The fixed positions are: the centre, North, South, West, East, North-West, North-East, South-West and South-East position of the window. Since these 9 examples of the same nodule are highly correlated, this might result in overfitting. However, during this research, it was found out that including these highly correlated examples as an enrichment of the data, the results on the test set improved. Please see Fig. 7 for a graphical representation of the positive nodules examples and the difference between a natural- and a fixed position.
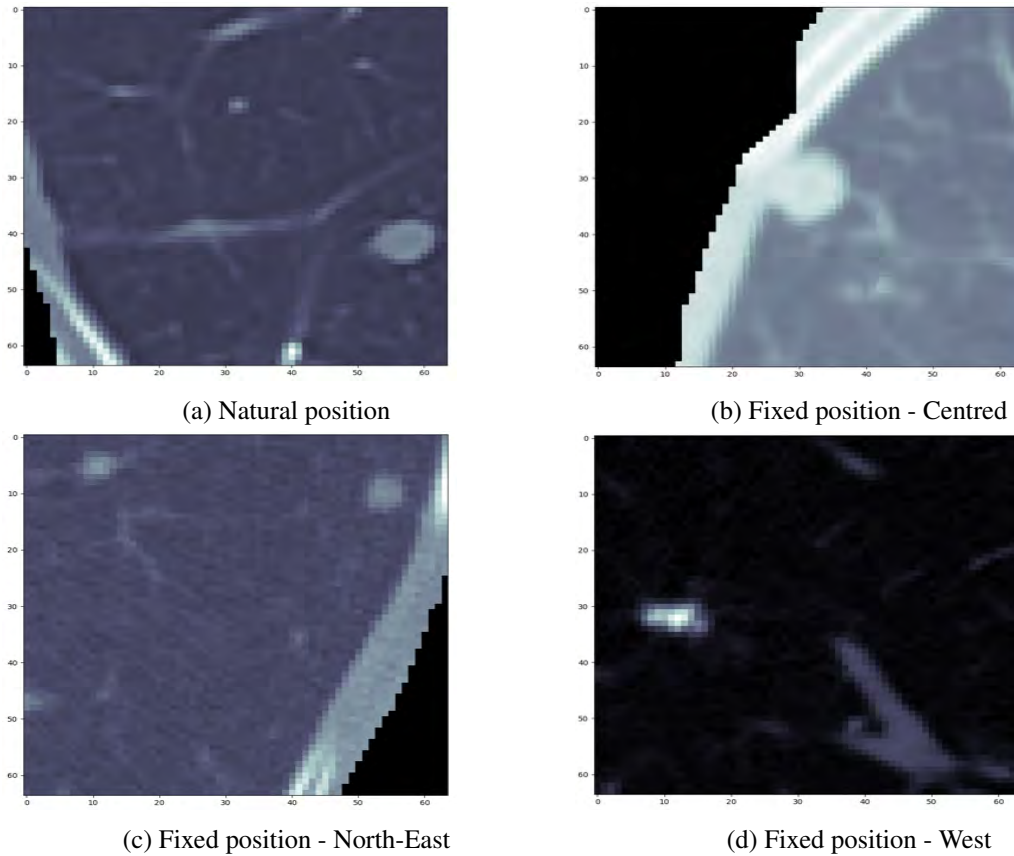
(a) Natural position   (b) Fixed position - Centred

(c) Fixed position - North-East   (d) Fixed position - West

Figure 7: Different types of positive nodule examples

Negative examples

Negative examples are created in multiple phases. For the first model negative examples are randomly are chosen windows from negative patient scans. After the first model was successfully trained, the false positives (FP) of this model served as negative examples for the next models. FP windows were extracted from both the positive and negative patient scans. In the final model a small portion of these randomly chosen negative windows was kept as training examples. The reason for this is to learn the model that areas outside the human lungs are negative as well. If only FP's are included in the negative training set – areas inside or at the border of the lungs - the model will not be able to know what the areas outside the lungs represent.

3.7.2 Data Augmentation

In addition to the above mentioned methods to create training examples, data augmentation was applied to further enrich the data set. It was applied only applied on the natural extracted windows and applied in three different ways:

1. Flipping the nodule horizontally

2. Flipping the nodule vertically

3. Rotating the nodule 90 degrees

These 3 methods are randomly applied with one of the methods, two combined or all together. Rotating less than 90 degrees would not add much variance. Other angles larger than 90 degrees would not result in new variance, because of the flipping in both ways. Please see Fig. 8 for a graphical repsentation of this process.
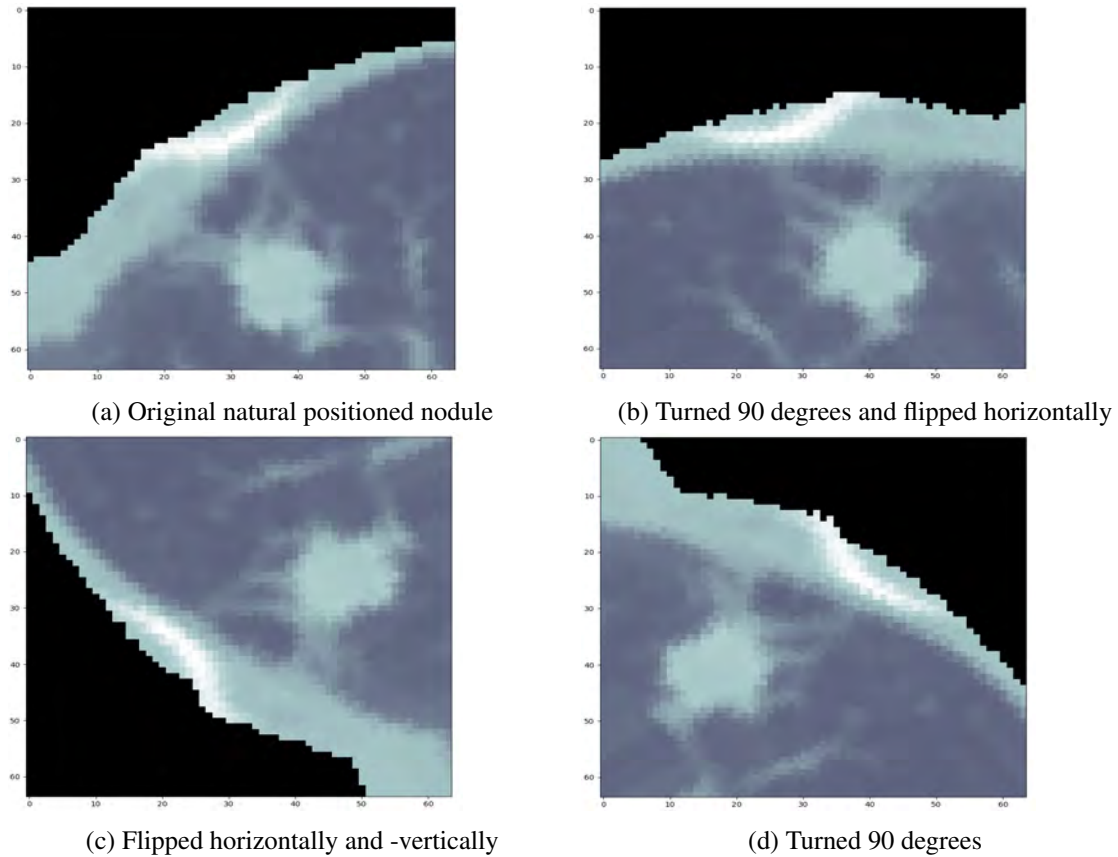

(a) Original natural positioned nodule

(b) Turned 90 degrees and flipped horizontally

(c) Flipped horizontally and -vertically

(d) Turned 90 degrees

Figure 8: Three types of data augmentation applied to the same nodule

3.7.3 Distribution of training types

With the above mentioned methods, many types of training examples are created. If all the examples are used, some types – e.g. fixed positions – are over represented in the training set. Other types – e.g. natural class – are underrepresented in the data. Therefore a ratio between these types is used. It was found out during this research that using this ratio for the training set, the results on the test data were the highest. In Table 1 and Table 2 the distribution of types is presented. The percentages are based on around 16.000 training examples per class (positive or negative).

Table 1: Distribution positive example types

Positive Type	%
Natural	6.64
Fixed positions	19.99
FN	26.92
Augmented data	46.45

Table 2: Distribution negative example types

Negative Type	%
FP	79.38
Randomly chosen windows	20.62

3.7.4 Normalization and zero-centring

As final pre-processing steps, the model input data is normalized and zero-centred. According to Andrew Ng [29], normalization of input data can speed up the training process. In addition, the input data is centred at 0. This means that for every pixel value, the mean of all pixel values is subtracted.

4 Experimental set-up

To test the proposed CAD system scientifically, an experiment has been carried out. The experiment consists of testing the proposed CAD system, along with comparing the three different neural network architectures. In this section the experimental set-up and evaluation methods are described.

4.1 Train-, validation- and test set

From the obtained data set as described in section 3.7, 80% was used as training set and 20% was kept as test set. This test set is completely kept out of the training process. The choice of this 80:20 ratio is based on the little amount of data available during this research. Data analysis showed that there is much variance in the size and location of the nodules. Therefore a model needs enough training examples to generalize well. On the other hand, there need to be enough test data to evaluate the models after training. Therefore, the 80:20 ratio is a reasonable choice. To validate the model during the training phase, the same 80:20 split is made to divide the training set into a training- and validation set. The same arguments apply for this choice. Please see Fig. 9 for a graphical view.
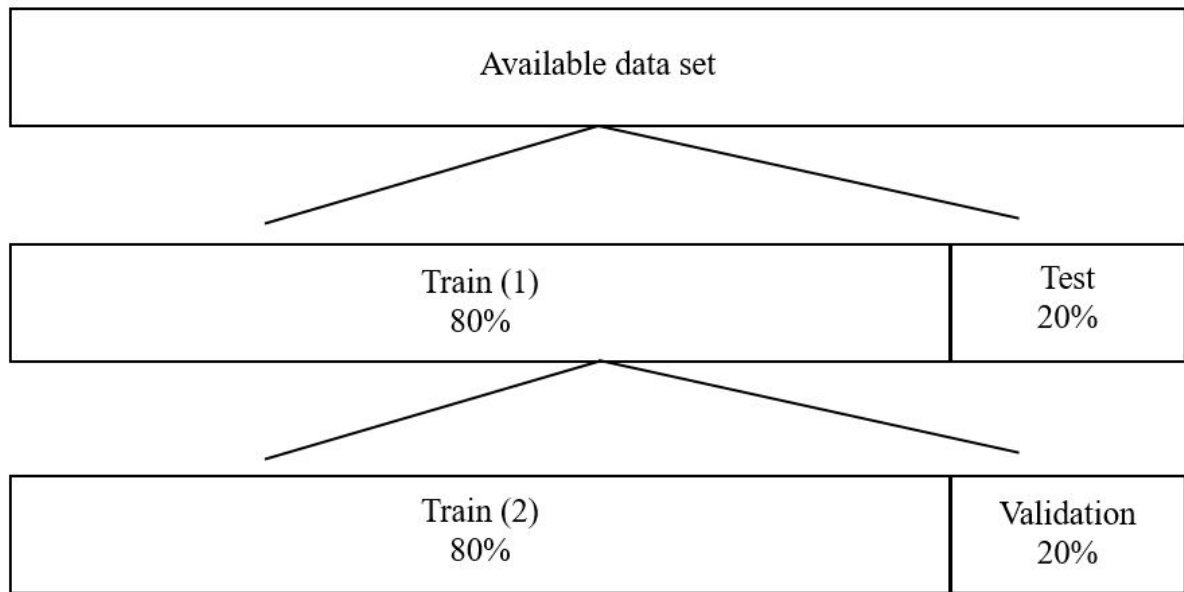

Figure 9: Splitting the obtained data set into different parts

During training, validation and testing of the CNN classifiers, each of these subsets has been used with another purpose.

- Train (1): used in the training process. Contains training- and validation part.

- Test: used to test classifier. This subset is completely kept out of the training process.

- Train (2): used to train the classifier. Every classifier is only trained on this subset. Reason to not let the CNN train again on Train(1) is computational effort and time.

- Validation: used to validate the classifier during the training process. Furthermore, used for stop- or improvement criteria.

21

4.2 Evaluation methods

Evaluation of the model is done at multiple stages. For each stage, different evaluations metrics have been used. In this section, each of these stages and concerning metrics are specified and choices are clarified.

4.2.1 Evaluation during training with validation set

During training of a machine learning classifier, in general, two types of errors can be found: the training- and validation error. The training error is the error made during training. The error is often called loss and needs to be minimized. After training a model, the validation error can be found by testing the model with held out data not used during training.

Besides loss, other metrics are used: accuracy and sensitivity. Accuracy is the amount of true positive (TP) classified examples and true negative (TN) classified examples relative to the total number of cases. Sensitivity, often called sensitivity, is the number of TP's relative to all true positive cases, including missed examples, which are called false negatives (FN).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

To use the above mentioned metrics a threshold must be set for a prediction, which is a probability that an example belongs to one of the classes. During training and validation, a threshold of 0.5 is used.

The choice for a threshold of 0.5 is quite common in science. However, an ideal threshold for classification tasks varies and cannot be stated up front. To be able to evaluate the model without setting a threshold a Receiver Operating Characteristic (ROC) can be plotted and its area under the curve (AUC) can be used as single value for evaluating the model. An ROC curves plots the performance of a binary classification model for many different thresholds. An AUC score of 1 is the maximum score and indicates a perfect model.

4.2.2 Evaluation after training using test set

The 2D CNN performs its classification task on a window from a single slice of the whole scan. Each slice is split into 144 smaller windows and each of these windows can be either a TP, FP, TN or FN, depending on the classification and its true class. It is desirable for the CAD system to keep the FP and FN errors as low as possible. However, for the medical diagnostic purpose it is better to have more FP's than to have more FN's. A radiologist rather has more false positives from his system than missed nodules, the false negatives.

On slice level, each of the 144 windows is evaluated. Positive slices are chosen by using the annotations in which the radiologists' denote the slices that contain nodules. Negative slices are chosen from the negative scans and the slices are chosen random-ly from the same quantile range in which all the positive slices fall. The amount of positive / negative slices is equal.

After the model has seen all 144 windows on a slice, the values for TP, FP, TN and FN are found and the accuracy and sensitivity on the slice can be calculated.

When all scans in the test set are evaluated the average scores of the evaluation metrics can be calculated. These scores serve as final ones for the model.

4.3   5-fold Cross-Validation

In this research 5-fold cross-validation (CV) is used to evaluate a model. CV is a commonly used technique in machine learning problems and gives a better indication of how the model performs. By using this technique all the data is used during training, validation and testing. Since its computational expensive to create a models' input data for this research, including the augmented examples, there has been chosen to use 5-fold CV instead of the commonly used 10-fold CV. Furthermore, there has been chosen to use stratified 5-fold CV in order to preserve the same amount of negative and positive examples in each fold. In the end, the scores of each fold are averaged and serve a final performance scores for the model.

4.4   Confidence intervals

Confidence intervals are used to compare the performance of the proposed CAD system with the performance of the two smaller networks. For both Sensitivity and FP rate - as these two are our most important evaluation metrics - 95% confidence intervals are derived. However, before deriving these confidence intervals, a normal distribution is needed. After normality tests, it was found out that normality could not be assumed. Therefore, bootstrapping with thousand times sampling is used to define the confidence intervals.

5 Results

In this section the results of the proposed model are presented. The results are presented in the order of evaluating. In addition, in section 5.4, the results of the three different network structures is presented.

5.1 Results during training

Table 3 presents the evaluation results of the training process. These scores are averaged from the 5-fold CV results. To calculate these metrics a threshold of 0.5 is used.

Table 3: Results of the 5 CV folds during training

Metric	Loss	AUC (ROC)	Accuracy	Sensitivity
Score	0.07	0.99	97.4%	96.8%

5.2 Results after training

On slice level other evaluation metrics are used. Just like related studies, evaluation metrics used are FP's per slice, Sensitivity and Accuracy.

Again, to go from 5-fold CV score to single scores, averaging is used and the results are presented in Table 4. The FP score is shown in percentage and absolute value for the 144 windows of one slice.

Table 4: Averaged results for the evaluation metrics on slice level

Metric	FP rate	Accuracy	Sensitivity
Score	1.5% (2.1)	98.2%	60.1%

An optimal threshold can be chosen if weight is added to either FP or Sensitivity. With weight to each score, an optimal trade-off point between these scores can be found. This is up to the experts, the radiologists, and is not done during this research. However, the Fig. 10 is generated to compare the relationship of both evaluation metrics, with different thresholds. In this figure both evaluation metrics are plotted against 10.000 different threshold values ranging from 0 to 1. These FP and Sensitivity scores are combined from all 5 CV folds, so all the CT scans.
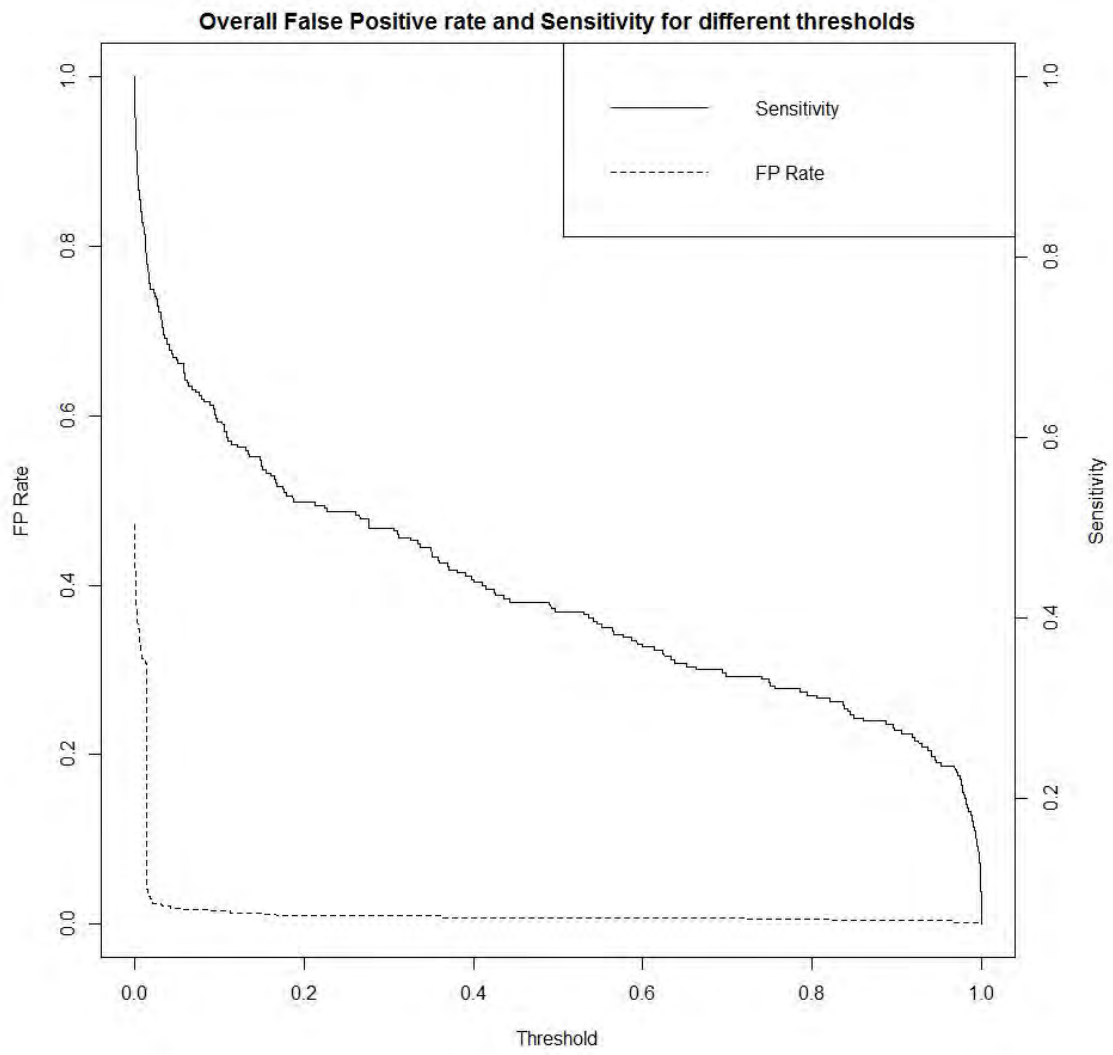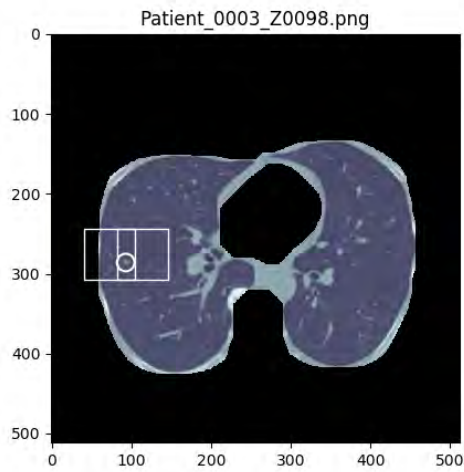
Figure 10: Overall False Positive rate and Sensitivity for different thresholds
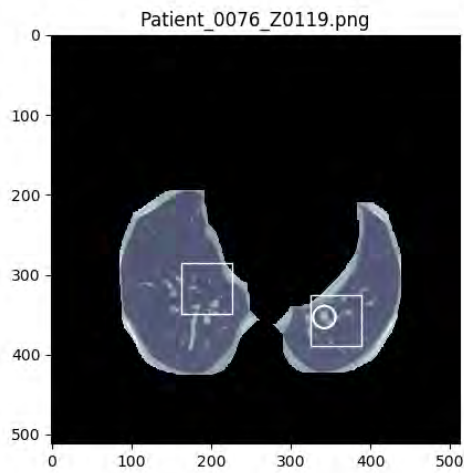
5.3 Final predictions

The following examples are final predictions of this CAD system. An annotated nodule is marked by a circle, while a positive prediction from the CAD system is marked by a square.

(a) Slice 98 of patient 3
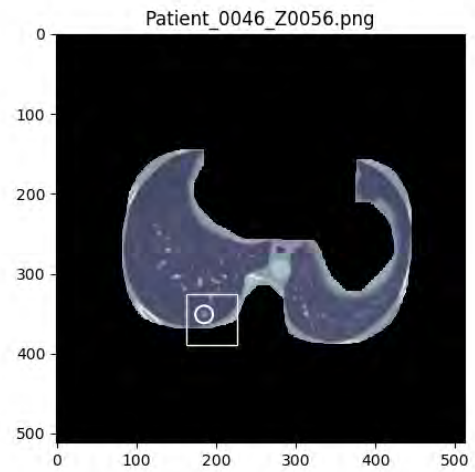
(b) Slice 56 of patient 46

(c) Slice 119 of patient 76

(d) Slice 59 of patient 150

Figure 11: Example predictions of CAD system

5.4 Comparison of network structures

Results for the three different network structures.

Table 5: Averaged results for the evaluation metrics on slice level

Network	FP rate	Accuracy	Sensitivity
Small	10.3% (14.8)	89.4%	49.8%
Intermediate	6.8% (9.8)	92.9%	58.1%
Deep	1.5% (2.1)	98.2%	60.1%

The 95% confidence intervals for the three different network structures are shown in Table 6.

Table 6: Averaged results for the evaluation metrics on slice level

Network	FP rate	Sensitivity
Small	[9.7% , 10.8%]	[45.8% , 53.1%]
Intermediate	[6.4% , 7.2%]	[53.6% , 61.2%]
Deep	[1.4% , 1.6%]	[57.6% , 64.2%]

6 Discussion

In this paper the development and evaluation of a whole CAD system for detecting lung nodules in 3D space is presented. To qualify the system, its performance and results are put into different perspectives and compared to previous research.

6.1 Putting results CAD system into perspective

When comparing the evaluation scores between the validation set and test set on a threshold of 0.5, one can see that there is a big difference in the FP and Sensitivity scores. This can be qualified as overfitting: fitting the training data too well, but fails to generalize to new, unseen examples. An explanation for this might be the large number of highly correlated training examples, created in the data augmentation phase. However, in this research it was found out that without these extra training examples, validation scores and test scores were even lower, and the difference larger. Therefore, this research confirms previous research in the sense that data augmentation is needed for CNN's to perform better on this detection task with this relative small data set.

However, there is still a difference in the way the validation scores and test scores are calculated. During training, the training- and validation set consisted of an equal amount of negative and positive examples. During testing, the evaluation results were on slice level. This has been done to simulate the scenario of a radiologist as much as possible. On each slice there are much more negative windows, containing no nodule, than positive windows, containing a nodule. Data analysis on 100 randomly chosen positive slices from the data set showed an average of 2.4 (1.67%) positive windows on each slice versus 141.6 (98.3%) negative windows.

In addition, the accuracy measure is also calculated in this research. Looking only to the accuracy one could say that the model performs well. However, since there is a large unbalance in the number of positive and negative examples, this measure is not reliable and is not used during in model comparison.

Fig. 10 shows that is hard to define an optimal threshold. However, the FP Rate curve signifies a possible point at 0.02. It would result in a possible sensitivity of 80% and a FP rate of 0.02. Using this threshold, averaged over all the 5 CV folds, results in a Sensitivity of 83% and 11 FP's per slice. This large number of FP's make the threshold of 0.02 not a reasonable choice for radiologists.

Last but not least, in this research, only nodules larger than 3 mm are considered as real nodules. There is a chance that nodules smaller than 3 mm are found by the classifier but received the label 'false positive', because these small nodules are not considered as real nodules in this research.

6.2 Compared to related research

To evaluate this CAD system even further, the performance of this system needs to be put into perspective of other CAD systems from previous research. Since the CNN of this CAD system is trained on a CPU computer, it makes no sense to compare these results with very deep CNN's trained on (multiple) GPU computer(s). The following CAD systems were trained with CPU computers and serve as good reference systems.

Lee et al. [12] also used a sliding window method approach and the same data set. There system achieved a sensitivity of 1 and FP rate of 1.4 on slice level, which is better than our CAD system.

However, the scores from Lee et al. are obtained from just 19 slices in contrast to the 1600 slices in this research.

Yang et al. [15] also proposed a CAD system with sliding windows and a 2D CNN as classifier. They use a much larger test set 11.000 examples, but evaluate their system only with the validation loss. Their validation loss is somewhere between 0.02 and 0.48, depending on a (artificially) created data set of nodules. The validation loss in this research is 0.07, which makes it perform worse if the loss of 0.02 is chosen from Yang et al and better if the loss of 0.48 from Yang et al. is chosen and most of the values in between.

6.3 Comparison between the three different network architectures

Looking at Table 5 in the Results section, the network of the proposed CAD system (Deep) performs best at all evaluation metrics. Especially the FP rate decreases with deeper network structures.

To compare these different networks statistically we can look for overlap in the confidence intervals. If we compare the Small and Intermediate network, we can assume with 95% confidence that the difference in FP rate and Sensitivity level is significant. If we compare the Intermediate and Deep network we can assume that the FP rate is significantly different, but the Sensitivity level is not significantly different.

6.4 Limitations of this research

In many situations where machine learning techniques are applied, the goal is to find a model that generalizes well, so it is able to make good predictions on unseen data. During this research, it became clear that it was hard for the model to generalize well. More data with more variance in the nodule examples would in theory lead to better models. However, the data set used in this research is relatively small and although data augmentation improved performance a little bit, the difference between training- and test error was still large.

In this research, the CNN's are trained on a 2.4 GHz dual-core CPU computer. If training was done using a GPU computer, training would be much faster and deeper networks could be trained. Deeper 2D CNN's might be able to able to capture more distinguishing features of the nodules. The winning architectures over the last few years of the ImageNet challenge, all use 5 convolutional layers or more [30]. With the 3D characteristics of CT scans, a 3D CNN would in theory be a good choice and this is proved in previous research. However, these networks come also at a high computational price.

Regarding the three-dimensional characteristic of CT scans and nodules in general, no such features are derived from images in this research. Capturing three-dimensional anatomic information could possibly lead to better classification of tissue.

Besides sensitivity, the average number of FP's is an important measure for CAD systems. It's desirable to have these as low as possible. In this research, the number of FP's is measured directly after prediction. In other studies [31], subsequent classifiers are used to reduce the number of FP's. That's not done in this research, mainly because of the limited available time.

6.5   Recommendations for further research

This research indicates that it's hard to detect all the nodules, without making too many false classifications (FP's). It's hard to generalize well and this problem might be partly solved by adding more data. Another possible solution might be deeper 2D CNN's or 3D CNN's that capture three-dimensional information.

Just as other proposed CAD systems in similar research, the creation of a model that is able to detect nodules is done by a machine learning model, trained with many examples. It might be a better idea to combine this type of learning, with domain specific knowledge from real experts, the radiologists. Perhaps rule-based classifiers can be leveraged with machine learning ones to create even better CAD systems.

In the convolutional layers of a CNN architecture, a filter is stridden across the input image, searching for patterns. In this research, the default Tensorflow / Keras filters are used, but specific filters can be created that are more sensitive to e.g. round objects. In these filters, the underlying convolutional mathematical function is different. Since most nodules are rounded or spherical objects, such a specific filter might be able to detect more nodules.

7 Conclusion

The goal of this research was to develop a CAD system that is able to detect and localize lung nodules in three-dimensional CT scans of humans. The system proposed in this paper is able to do so with the stated sensitivity and false positive scores. Since every study, including this one, has its own pre-processing methods and experimental set-up, it's hard to compare the false positives and sensitivity scores with related studies. Therefore, different network architectures for the CAD classifier have been tested during this research. With a confidence of 95% we can conclude that deeper networks improve performance on the false positive scores significantly. Sensitivity scores did improve, but not significantly with deeper networks.

Current CAD systems serve as second-opinion for radiologists and reduce some of their workload. The results of this research confirm previous research in the sense that it's very hard to create a perfect CAD system for detecting lung nodules. Therefore, the domain specific knowledge and experience of radiologists will always be needed in this field, at least in the near future.

References

[1] "World Cancer Research Fund International". *"Worldwide data"*. URL: `http://www.wcrf.org/int/cancer-facts-figures/worldwide-data`.

[2] "American cancer society". *"Cancer statistics US"*. URL: `https://cancerstatisticscenter.cancer.org/#/`.

[3] "Integraal kanker centrum Nederland". *"Cijfers over kankker"*. URL: `http://www.cijfersoverkanker.nl/meest-voorkomende-soorten-52.html`.

[4] "Cancer research UK". *"Survival - Lung cancer"*. URL: `http://www.cancerresearchuk.org/about-cancer/lung-cancer/survival`.

[5] "American Journal of Respiratory and Critical Care Medicine". *"Screening for Lung Cancer with Low-Dose Spiral Computed Tomography"*. URL: `http://www.atsjournals.org/doi/abs/10.1164/ajrccm.165.4.2107006`.

[6] "Awai K. et al. ""Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance"". In: 2 ("2003").

[7] Kenji Suzuki. "A review of computer-aided diagnosis in thoracic and colonic imaging". In: *Quantitative Imaging in Medicine and Surgery* 2.3 (2012).

[8] Jacobs C et al. "Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database". In: (2016).

[9] "Gurcan MN. et al. ""Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system"". In: ("2001").

[10] "McCulloch CC et al. ""Model-based detection of lung nodules in computed tomography exams. Thoracic computer-aided diagnosis"". In: ("2004").

[11] Zhanyu Ge et al. "Computer-aided detection of lung nodules: False positive reduction using a 3D gradient field method and 3D ellipsoid fitting". In: (2005).

[12] "SLA Lee, AZ Kouzani, and EJ Hu". ""Automated identification of lung nodules"". In: ("2008").

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: (2012), pp. 1097–1105.

[14] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. "CNN-based Segmentation of Medical Imaging Data". In: *CoRR* abs/1701.03056 (2017).

[15] He Yang, Hengyong Yu, and Ge Wang. "Deep Learning for the Classification of Lung Nodules". In: (2016).

[16] Hamidian S et al. "3D Convolutional Neural Network for Automatic Detection of Lung Nodules in Chest CT". In: (2017).

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *CoRR* abs/1606.04797 (2016).

[18] Özgün Çiçek et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *CoRR* abs/1606.06650 (2016).

[19] "Cancer Imaging Archive". *"Lung Image Database Consortium image collection (LIDC-IDRI)"*. URL: `https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI#4364e357960c43638f9574657c822e8e`.

[20] "Dr Yuranga Weerakkody and Kyle Greenway et al.". *"Hounsfield Unit - Radiopaedia"*. URL: `https://radiopaedia.org/articles/hounsfield-unit`.

[21] Armato SG, McLennan G, and Bidaut L. "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans." In: (2011).

[22] "The OpenCV team". *"Open Source Computer Vision library"*. URL: `https://opencv.org/`.

[23] "Alex Clark and Contributors". *"Python Imaging Library"*. URL: `https://pillow.readthedocs.io/en/latest/`.

[24] "Jos B.T.M. Roerdink and Arnold Meijster". ""The Watershed Transform: Definitions, Algorithms and Parallelization Strategies"". In: ("2001").

[25] R. Shojaii, J. Alirezaie, and P. Babyn. "Automatic lung segmentation in CT images using watershed transform". In: (2005).

[26] "Ankasor". *"Lung segmentation using Watershed"*. URL: `https://www.kaggle.com/ankasor/improved-lung-segmentation-using-watershed`.

[27] CC Wan. "Lung nodules detection in CT scans". In: (2017).

[28] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014).

[29] "Andrew Ng". *"Normalization inputs - Deep Neural Networks"*. URL: `https://www.coursera.org/learn/deep-neural-network/lecture/lXv6U/normalizing-inputs`.

[30] Rajat Vikram Singh. "ImageNet Winning CNN Architectures – A Review". In: (2016).

[31] Firmino M et al. "Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy". In: (2016).