VRIJE UNIVERSITEIT AMSTERDAM

FACULTY OF SCIENCE

BUSINESS ANALYTICS

RESEARCH PAPER

Automated Localization of Lung Nodules

Author:
Robin Vastenou

Supervisors:
Dr. M. Hoogendoorn
Dr. E. Haasdijk

September 30, 2018

Preface

This research paper has been written for the course Research Paper Business Analytics as part of the Business Analytics Master program at the Vrije Universiteit Amsterdam. The goal of this research is to demonstrate the student's ability to successfully go through the process of doing a research, write a report and present it.

I would like to thank my supervisors Dr. M. Hoogendoorn and Dr. E. Haasdijk for their introduction to the subject and assistance. Besides, I would like to thank the radiologists of the VUmc for providing me with extra information from their point of view.

Abstract

Lung cancer is the deadliest form of cancer for men and women. For people that are diagnosed with cancer it is of great importance that the disease is detected at an early and treatable stage. In order to improve radiologists' performance for detecting nodules on CT examinations a computer aided detection system is developed that predicts the location of lung tumors in computerized tomography scans. The goal of these methods is to reach a high true positive rate and a low true negative rate.

In this research paper, the requirement of using relatively low computational power is added. Therefore, a two-dimensional method is built that also includes three-dimensional features. Hereby, the steps of preprocessing, segmentation, candidate detection, feature extraction and classification are taken. The method that is created vertically connects the components of segmented data. Neighbourhood Component Analysis is then used to make an optimal subset of features. These features are used by a Neural Networks classifier and a Support Vector Machine classifier to classify the connected components. Finally, the results of these methods are compared and the Support Vector Machine classifier turns out to have the highest AUC.

Contents

1 Introduction

Lung cancer is by far the deadliest form of cancer for men and women (Association, 2014). The Dutch Cancer Registration registered 12758 incidents of lung cancer in 2006 in the Netherlands(NKR, 2016). In the same year, the institute noted that 10420 incidents had caused mortality. These numbers show the urgency of research in ways to fight this terrible disease.

For people that are diagnosed with cancer it is of great importance that the disease is detected at an early and treatable stage. Having the cancer nodule detected, the chance of surviving can be strongly enhanced with a proper treatment. The five-year survival rate for lung cancer is 55 percent for cases detected at an early stage according to the American Lung Association (2014). To be able to treat the disease well, for example by radiation, the nodules need to be localized very precisely. The detection of lung cancer nodules is done on computerized tomography scans (CT-scans). These are cross-sectional images of the body that are produced with X-rays. One CT-scan can consist of over 50 two-dimensional slices of the body. CT-scans are usually done on patients that have complaints which are directly linked to lung cancer. More than half of these patients die within one year of being diagnosed because the disease has spread (American Lung Association, 2014). However, the CT-scans of patients that do not have a spread disease require a thorough analysis as obviously no nodules can be missed. For instance, nodules can be smaller than 10 mm and very hard to find. Therefore they are often overlooked by visual inspection alone. Besides, determining the difference between blood vessels and actual tumors can be complex. These and similar issues, in combination with the possibly large amount of slices that a CT-scan can contain, makes the analysis very time-consuming for the radiologists.

Currently, radiologists use computer aided detection (CAD) to analyze the CT-scans. CAD is a technique that is designed to increase the detection of diseases by decreasing the false negative rate for the interpretation of medical images (Castellino, 2005), and in this case CT-scans. In their analyses, radiologists use it as an extra tool that automatically detects and localizes lung tumors. This way CAD becomes the second reader: The radiologist and the CAD system do the analysis independently, after which the results are put together. In many literature, like Sahiner (2010) and G. Rubin (2005), screening by visual inspection alone is compared to screening with the use of CAD and it is shown that CAD indeed improves radiologists' performance for detecting nodules on CT examinations.

The CAD-systems that we know can be divided into 2D-systems and 3D-systems. These 2D-systems can achieve a high sensitivity but they are generally too simple and return too many false positives. On the other hand, the 3D-systems require high computational power. Therefore, a method is proposed that requires relatively low computational power, but still achieves a high sensitivity with a low number of false positives. This method uses 2D-methods with 3D features.

This research starts with a literature review. Then, a description of the dataset is given in section 3. In section 4 the methods that are used are described. The experimental setup can be found in section 5 and the results in section 6. Finally, the assumptions and results are put into perspective in the conclusion in section 7 and the discussion (section 8).

2 Literature review

The typical set-up of a CAD system is preprocessing, segmentation of the structures of interest, detection of candidates, extraction of the candidates by features, and classification of the candidates (I. Sluimer, 2006):
Segmentation techniques convert gray-scale images into binary images. Tresholding techniques are very common in the field of lung nodule detection for segmentation. One may think of global tresholding (Messay, 2010), here the same threshold is applied to all pixels: All the pixels below the treshold get value 0 and above the treshold get value 1. For optimal tresholding (Dehmeshki, 2007), firstly a grey level histogram is approximated by a linear combination of Gaussians. Then the tresholds are computed by using a minimum classification error. Ye (2009) propose an algorithm that creates a five-dimensional vector that includes the position (x,y,z), the intensity and the shape index for every voxel. With this vector the intensity mode map and shape index mode map are computed. An expectation maximization algorithm is then used to merge the neighbouring modes based on the Bayesian probability theory that assigns a probability for each mode to which class it belongs. Another segmentation option is region growing. This approach examines neighboring voxels of initial seed points and determines whether the pixel neighbors should be added to the region. Bellotti (2007) adopt two inclusion rules for the neighboring voxels: Simple Bottom Threshold/Simple Top Threshold and Mean Bottom Threshold/Mean Top Threshold. The first checks per voxel and the latter checks for the average of all neighbouring voxels whether the voxels should be adopted or not. The seed point is chosen as the first voxel that satisfies the inclusion rules. Another example of a region growing method in lung nodule detection is using wavefront simulation and suitable stop conditions (Nunzio, 2011).
Having the lung segmented, the nodule candidates can be detected. Several techniques have been proposed to detect the candidates like mathematical morphology(Awai, 2004) and connected component analysis (Oda, 2002). Mathematical morphological operations tend to simplify image data preserving their essential shape characteristics and eliminating irrelevancies (Haralick, 1987). In connected component analysis a unique label is assigned to every maximal connected region of foreground pixels. With both methods a structure of possible candidates remains.
The third step that I. Sluimer (2006) proposes is feature extraction. In feature extraction mathematical, textural and geometrical properties are calculated from the segmented region(N. S. Lingayat, 2013). Features that N. S. Lingayat (2013) proposes are area, perimeter, irregularity index, equivalent diameter, convex are, solidity, gray-Level co-occurrence matrix properties, contrast, correlation, energy, homogeneity, statistical properties of an image, mean, variance and standard deviation. Singh (2016) use the area, perimeter, shape complexity, mean standard deviation and circularity as features. The basic characters of geometric feature are area, perimeter and compactness. A. Gajdhane (2014) restricts to the basic characters of geometric feature which are area, perimeter and compactness. P. Lambin (2008) and Ahmad (2015) distinguish three types of features: Shape based features, intensity based features and texture based features.
With the retained features the extracted volumes can be classified. D. Zinovev (2011) uses belief decision trees. Furthermore, the K-nearest neighbor algorithm is applied by Farag (2011) to classify the lung nodules. Another classification method is the usage of support vector machines (P. Lin, 2013). A more complex form of classification is deep learning. One may think of (convolutional) neural

networks and deep belief networks (K. Hua, 2015).

3 Data

In this section, a description of the dataset is given along with some background and terminology that are of hand further in this research.

3.1 LIDC dataset

The Lung Image Database Consortium image collection (LIDC-IDRI) is a collection of thoracic CT-scans and corresponding annotations that is available on the website of the cancer imaging archive. The data is gathered by seven academic centers and eight medical imaging companies and consists of 1018 scans. Each scan consists of all the slices of the CT-scan (DICOM images) and an XML-file with the annotations of four experienced radiologists.
The annotations are made in a so-called two phased image annotation process. In the first phase of the process, the four radiologists analyze the images independently and mark the lesions belonging to one of three categories: ”nodules of at least 3mm”, ”nodules smaller than 3mm” and ”non-nodules bigger than 3mm”. In the second phase, the radiologists compare their own analysis with the analyses of the other radiologists to adjust their own analysis if necessary and make a final annotation. As a result, the CT-scan for a patient is analyzed as completely as possible by four radiologists.
In total, the LIDC-IDRI database contains 7371 lesions marked as ”nodule” by at least one radiologist. Of these nodules, 2669 were marked as ”nodules of at least 3mm” by at least one radiologist. Thus, on average, a scan has 7,24 lesions annotated as nodule and 2,62 lesions annotated as nodules of at least 3mm.

3.2 Data background

The images in the scans are composed of 512 rows, each of 512 pixels, i.e., a square matrix of $512X512 = 262144$ pixels. An example of such a DICOM image can be found in figure 1. This is an average cross-section of the thorax.

Figure 1: Example of a CT-scan slice

To be able to make well-substantiated choices in building the model, it is important to understand more about the thorax that is visible in the CT-scan slices. The two dark regions inside the thorax are obviously the lungs. In figure 1, the bright part around the lungs is the chest wall, consisting mostly of bones and skin. In the bright part in between the two lungs, you can find the heart, the aorta, the pericardium, the esophagus and other organs. For this research paper only the lungs are of interest, including what happen at the edges of the lung. The edges of the lungs are called the pleurae. A pleura includes two thin layers of fluid-filled space that protects and cushions the lung. The small bright regions inside the lungs are blood vessels and air.
Generally, three types of lung nodules can be distinguished based on external attachment (Mukhopadhyay, 2016):

- Well-circumscribed nodules
- Juxta-vascular nodules
- Juxta-pleural nodules

The well-circumscribed nodules are detached, the juxta-vascular nodules are attached to blood vessels, while the juxta-pleural nodules are attached to the pleura. This terminology will be of hand in the candidate detection process in section 4.3.

4 Methodology

In this section, the methods that are used in this research paper are described. Firstly, the preprocessing stage is described. Secondly, the structures of interest are segmented. Then the candidate are detected from the segmented images, after which the feature extraction takes place. The section ends with the classification of the candidates.

4.1 Preprocessing

Due to a limited computation power, only a part of the in total 1018 scans is used in this research paper. The scans with a slice thickness larger than 2,5 are deleted. This is most convenient for the connected components algorithm that is used in section 4.3. It is better to reduce the slice thickness to 1 thickness, as P. Monnin (2016) does, however this is not within the scope of this research.
The slice thickness can be found in the DICOM metadata. By deleting the scans with a thickness larger than 2,5 mm, 888 scans remain:

Slice thickness (in mm)	**Amount**
0,6	7
0,75	30
0,9	2
1	58
1,25	343
1,5	5
2	123
2,5	320

Table 1: Division data in slice thickness

Of the remaining 888 scans, randomly 488 scans were removed. This also had to deal with limited computation power. This measure goes at the cost of the model accuracy unfortunately. The second step is to turn the annotations and the slices into the same filetype. TIFF images are used because they are easy to work with in MATLAB. This is done with a MATLAB function that extracts the annotations of each individual radiologist from the XML file into TIFF images. If no annotations are made for a slide, then an empty TIFF image is generated. Besides, the function converts the DICOM images of the slices into TIFF images for convenience.
The third step is to process the annotations. As explained in section 3.1, every CT-scan is analyzed by four different radiologists. As an empty TIFF image is generated when no annotations are made for a certain slice, every slice is linked to four TIFF images, which are either empty or annotated. To have only one reference instead of four, for every slice the unit of the annotations is taken. The reason for this is that I prefer to have a (part of a) lesion annotated as tumor that is not a tumor, over missing a (part of a) tumor. By doing this, no information gets lost. An example of 4 annotations in figures 2, 3, 4 and 5 being united as one annotation in figure 6 can be found below.

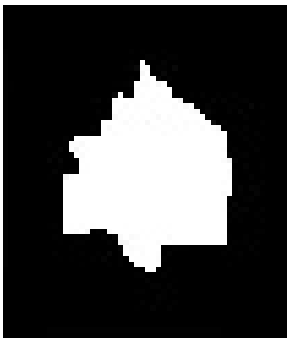

Figure 2: Annotation 1

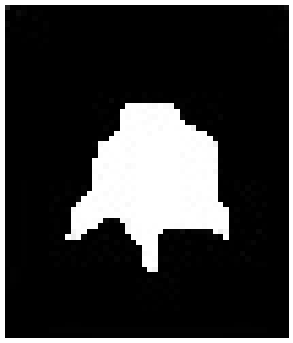

Figure 3: Annotation 2

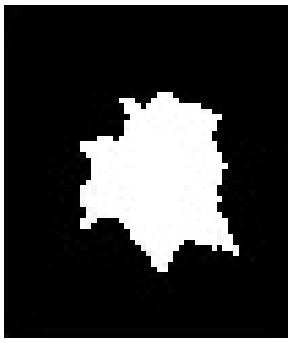

Figure 4: Annotation 3

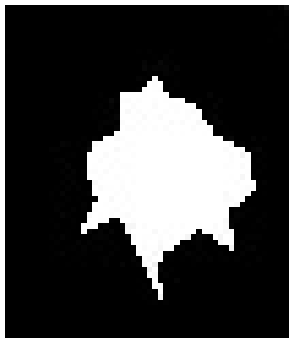

Figure 5: Annotation 4

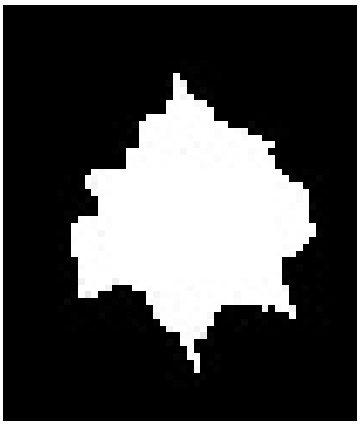

Figure 6: Unit annotations

4.2 Segmentation

To be able to differentiate the nodules from other parts in the CT scan, the gray scale image is converted into a binary image. Because the difference between air and other parts in the body are very obvious, a simple segmentation technique like tresholding is sufficient. To find the right treshold, the Houndfield scale is used. This scale is defined in HU (Hounsfield units) and is the unit of measurement in CT-scans. The scale encompasses the body's composition from air designated as -1000 units, running through water at 0 units up to bone, which is the densest human tissue, at $+1000$ units (Bhattacharyya, 2016).
To convert the normal values that are found in the CT-scan into HU, a linear transformation is needed:

$$HU = pixelvalue * slope + intercept.$$

The slope and the intercept of each slice can be found in the CT scan's data. A histogram of the transformed slice in figure 1 into HU looks as follows:

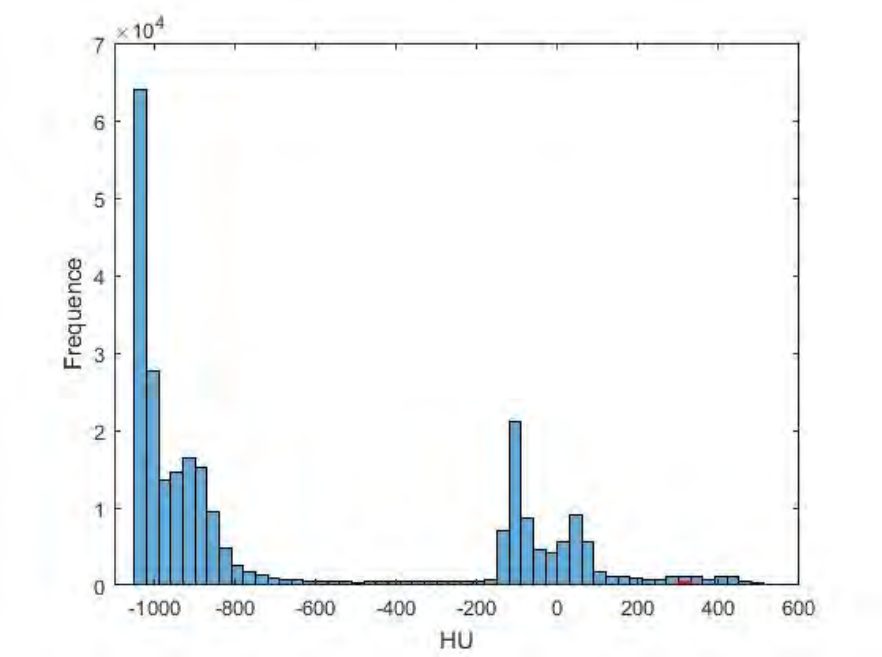

Figure 7: Histogram of the HU scale of a CT-scan

In figure 7 two separate peaks are visible. To have a better understanding of this histogram an oversight of the HU scale is given in table 2 by T. Buzug (2008). With this table it is easy to distinguish air and tissue in figure 7: the left peak is air and the right peak is tissue. Because the difference between air and tissue is that clear, a global treshold at -400 HU suffices to separate the air and tissue.

CT parts	**HU**
Bone	1000
Liver	40-60
White matter	20-30
Grey matter	37-45
Blood	40
Muscle	10-40
Kidney	30
Cerebrospinal fluid	15
Water	0
Fat	-50–100
Air	-1000

Table 2: HU corresponding to different tissues in a CT-scan

The result of a segmented image can be found in figure 8

Figure 8: Segmented CT-scan

4.3 Candidate detection

Having the CT-scan segmented, the lung candidates need to be detected. In a strictly 2D CAD system, one can simply use the foreground segments in the segmented image as nodule candidates. However, in this research paper the focus lies on the 3D features that can improve the nodule localization. Therefore, a method is constructed that creates candidates that are situated over multiple slices. It can be seen as a variant of connected component analysis. Thus, instead of only horizontal candidates that remain from a segmented slice, vertical candidates are looked for over multiple slice. Hereby, it is important to note that we are only interested in possibly malignant tumors. Therefore, the assumption is included that a malignant nodule occurs in at least 3 CT slices. This property was given by the radiologists of the VUmc. Besides, in table 3 of A. Leung (2007) we see that nodules smaller than 8 mm are hardly malignant. As the largest distance between slices is 2,5 mm (figure 1), all the possibly malignant nodules should be included by using this assumption.

Diameter	**Malignancy**
$<$4 mm	0%
4-7 mm	1%
8-20 mm	15%
$>$20 mm	75%

Table 3: Relationship between diameter and chance of malignancy

The algorithm starts by randomly choosing a segment, say segment s that is not yet examined in the CT-scan. In order to make the algorithm require low computational power, these segments are

picked only from every 3rd slice from the center of the scan. Say, a scan contains 14 slices, firstly the middle slice (slice 7) is picked. Then, from this middle slice, upwards and downwards, every third slice is used. Downwards that is slice 4 and 1, upwards that is slice 10 and 13. Thus, in this case the algorithm would only randomly select a segment of slices 1, 4, 7, 10 and 13. As we make the assumption that a nodule occurs in at least 3 slices, this does not make us miss any tumors.
Of segment s, the algorithm seeks whether other segments in the 2 neighbouring scans have a similar location in the CT-scan. The location of segment s is denoted as similar to the location of another segment when their centroids, which are (x, y)-coordinates, lay within a certain Euclidean distance from each other. This Euclidean distance is yet to be determined.
At this point, two cases are distinguished: Segment s has segments with similar locations in at least one neighbouring slice, or it does not. If one or both of the neighbouring scans indeed have a segment with a similar location as segment s, the segments are connected and seen as one possible candidate. Subsequently, the same process is repeated for the neighbouring slice of the segment that is connected to segment s. In the second case, one gap is 'used' and in one slice further again a segment with the same location is searched. If all the gaps are used and still no similar segment has been found, the segments are deleted and a new segment is chosen randomly from every 3rd slice from the middle. Reason for this is the assumption that a malignant nodule occurs in at least 3 CT slices.
The algorithm allows a certain amount of gaps within the connected components to also include the juxta-vascular nodules (section 3.2). The amount of gaps is yet to be determined. The blood vessels with which the juxta-vascular nodules are connected are usually thinner and not visible on all slices the nodule is situated at.

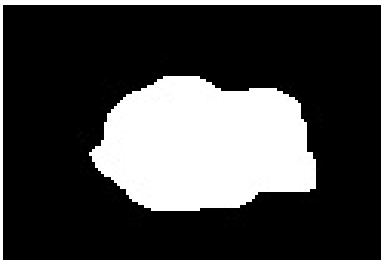

Figure 9: Nodule from lower slice

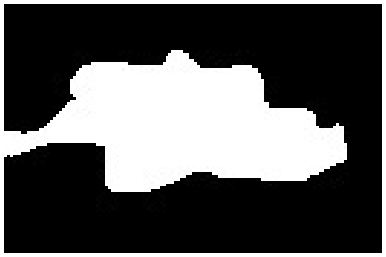

Figure 10: Juxta-vascular nodule

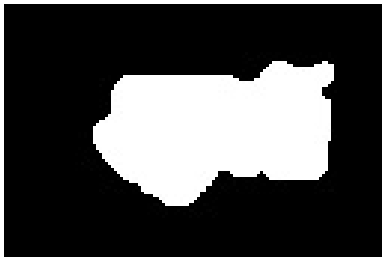

Figure 11: Nodule from upper slice

As a result that the blood vessels can strongly influence the centroid of the segment. In the example in figures 9, 10 and 11 such an occurrence is illustrated. Here, the centroid of the middle slice is completely changed due to the blood vessels. Thus, in this case allowing the gaps means that a component can be spread out over all three slices. The gap is counted as an occurrence if it has at least one segment in the slice above and at least one segment in the slice below in the same component.
The algorithm continues for segment s until the allowed amount of gaps is passed, resulting in one connected component. Then, the process is repeated for another randomly chosen segment that is not yet examined, until all segments have been examined.
As part of detecting the candidates, some segments are eliminated:

- The bones, which are the segments that have a HU of around 1000.

- The chest wall, which is the largest segment. Removing the chest wall also means that the juxta-pleural nodules are removed. These segments are saved in a different dataset.
- The segments in the vertically connected components that occur in less than 3 slices.

In this section a few variables remained that need to be tweaked in section 5:

- The Euclidean distance denoting the segments as being similar.
- The amount of gaps allowed within a connected component.

4.4 Feature extraction

Of the vertically connected components that remained from section 4.3, the features can be extracted. Because the components are situated on multiple slices, 3D features can be used. For this research, shape based features, intensity based features and texture based features are used. With these types of features sufficient properties of the nodules are included.
An oversight follows of the 2D features with their definition. The 3D features are retained by taking the average of the 2D features over all components.
The shape based features were retained from Mathworks (2016b):

- Eccentricity, which is the circularity.
- Area, which is the amount of pixels of the segment.
- Centroid's x-coordinate, which is the x-coordinate of the centroid.
- Centroid's y-coordinate, which is the y-coordinate of the centroid.
- Major Axis Length, which is the length of the major axis of the ellipse that has the same normalized second central moments as the segment.
- Minor Axis Length, which is the length of the minor axis of the ellipse that has the same normalized second central moments as the segment.
- Orientation, which is the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the segment.
- Perimeter, which is the length of the line forming the boundary around the segment.
- Solidity, wich is the ratio between the area and the convex area of the segment.

The intensity based features are computed from the original pixel values and are pretty straightforward:

- Mean.
- Median.

- Skewness.
- Standard deviation.
- Kurtosis.

The texture based features are computed by normalizing the gray-level co-occurrence matrix and retained from Mathworks (2016a):

- Contrast, which is the measure of the intensity contrast between a pixel and its neighbor over the whole image.
- Correlation, which is a measure of how correlated a pixel is to its neighbor over the whole image.
- Energy, which is the sum of squared elements in the GLCM.
- Homogeneity, which is the value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

In section 5 the right combination of features is determined.

4.5 Classification

To classify the vertically connected components two well-known methods are used: a neural network and a support vector machine. These are supervised learning techniques that learn the mapping function from input to output. The input is the vertically connected component and the output denotes whether the vertically connected component is a tumor or not. Supervised learning can be done because the data is labeled.

4.5.1 Neural Network

A neural network is a form of machine learning that is based on the functioning of the brain. As in figure 12, there is an input vector X which consists of all n input variables $x_1, x_2, ..., x_n$. Recall that n is determined in the feature selection in section 5. Linear combinations, called activations, are made of the input variables for all hidden neurons. In figure 12 the structure of a Neural Network can be found. The Neural Network has k hidden layers, and every hidden layer has m hidden neurons. The amount of hidden layers and hidden neurons is yet to be determined.

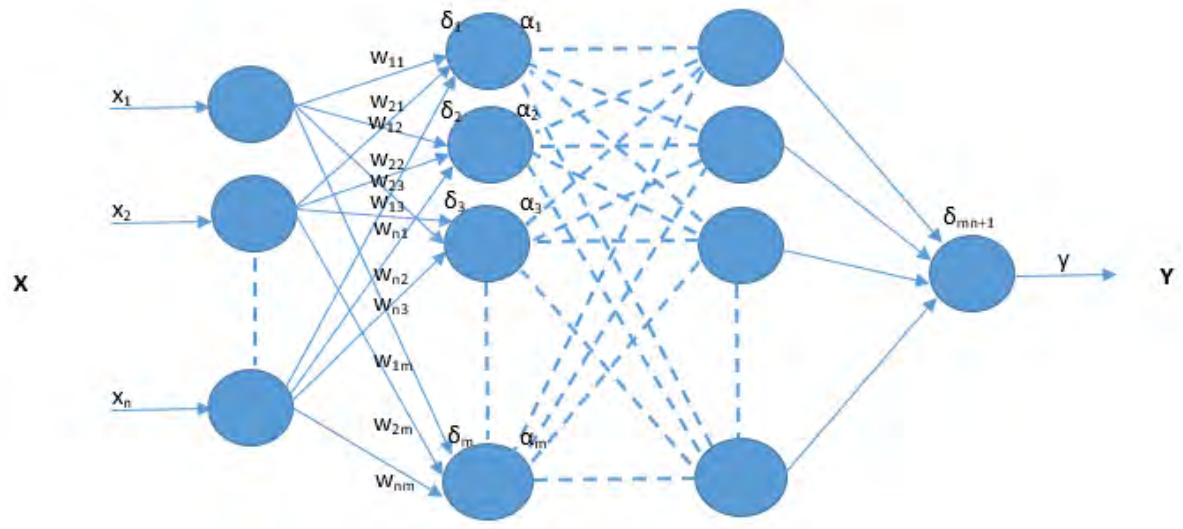

Figure 12: A possible Neural Network structure

The weights of these combinations are denoted by vector W with w_{ij} being the weight of the connection between input variable of the previous layer i and neuron j. The summed products of the weights and the input of the previous layer are called activations $\delta_1, \delta_2, ..., \delta_m$. For every neuron i that is

$$\delta_i = \sum_{j=1}^{n} w_{ij} \cdot x_j + bias. \tag{1}$$

To check the δ value produced by the neurons and decide whether outside connections should consider this neuron as activated or not, these activations are passed through a neuron's so-called activation function, say f:

$$a_i = f(\delta_i). \tag{2}$$

A well-known activation function is the step-function:

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leqslant 0 \end{cases}$$

Another option is the logistic sigmoid function, which is the smoothed version of the step-function.

$$f(x) = \frac{1}{1 + \exp(-x)}.$$

The advantage of smoothing the step-function is that small changes in the weights and in the bias result in small changes in the output α_i. Besides, the logistic sigmoid function takes values between 0 and 1.

The hyperbolic tangent activation function is easier for optimization however because it is zero centered. Therefore, this function is used as activation function for most hidden layers:

$$f(x) = 1 - \frac{\exp(-2x)}{1 + \exp(-2x)}.$$

In every hidden layer these steps are repeated until the output layer is reached. The output layer is determined by applying a similar combination as in (1) and (2) for the j-th output variable:

$$y_j = g\left(\sum_{i=1}^{m} w_{ij} \cdot a_i\right).$$

The activation function $g(x)$ that is used for the output layer is the logistic sigmoid function, because it returns values between 0 and 1, which is easy for finding the optimal treshold.
To train the weights of the model, back-propagation can be used. Back-propagation is a repetitive process to update the weights of the neural network. Firstly, the output is computed with random initial weights. With the sum of the squared error,

$$E(y,t) = \sum_{j}(y_j - t_j)^2,$$

the loss with respect to the target values t_j is computed. This error is then propagated back through the network by updating the weights. This process is repeated until the error is minimized. This back-propagation algorithm is known as a computationally efficient algorithm (Bengio, 2012).
To find the set of weights that minimizes the error, gradient descent is commonly used. With this method the gradient of the error with respect to the weights is computed:

$$\triangle w_{ij} = \eta \frac{\delta E(y,t)}{\delta w_{ij}}. \tag{3}$$

An important requirement for gradient descent is that its input, weight and activation functions have derivative functions, which is indeed the case. With (3) the direction with the steepest descent is chosen. Hossain (2015) shows that scaled conjugate gradient descent is a better option however. Scaled conjugate gradient descent does not only look for the steepest descent, but searches along conjugate directions. This makes the scaled gradient descent algorithm a faster algorithm faster than normal gradient descent.
In this section a few variables remained that need to be tweaked in section 5:

- The amount of hidden layers and hidden neurons.

4.5.2 Support Vector Machine

To have multiple options for comparison, support vector machine (SVM) is also used as a classification method. An SVM is a rather straightforward classifier that divides the data into nodules and non-nodules with a hyperplane. This hyperplane is chosen such that the distance between the nearest data points on each side is maximized. When the distance is maximized, the hyperplane is called maximum-margin hyperplane. In figure 13 an example of a maximum-margin hyperplane can be found. The dotted lines are the support vectors.

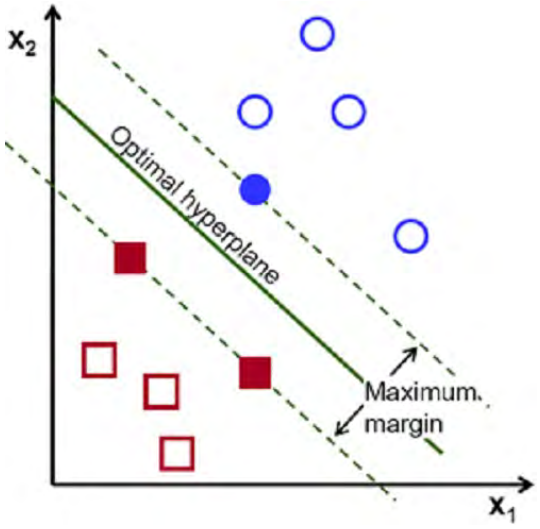

Figure 13: An SVM that divides the data with a maximum-margin hyperplane

To determine (the weights of) the optimal hyperplane, the following LP needs to be solved.

$$W(\alpha) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(X_i X_j)$$

subject to

$$\sum_{i=1}^{n} y_i \alpha_i = 0,$$
$$\alpha_i \leqslant 0.$$

In this LP the Lagrangian constant α is used to include the constraints. Furthermore, $x_1, x_2, ..., x_n$ is the input and $y_1, y_2, ..., y_n$ is the output which returns either a 0 (no tumor) or a 1 (tumor). $K(X_i X_j)$ is the kernel function of the dot product of input data points that maps the data points into the higher dimensional feature space by transformation. The kernels that are examined in this research paper are the linear function, the polynomial function and the radial basis function (RBF):

$$K(X_i X_j) = \begin{cases} X_i \cdot X_j & \text{Linear} \\ (\gamma X_i \cdot X_j + C)^d & \text{Polynomial} \\ exp(-\gamma |X_i - X_j|^2) & \text{RBF} \end{cases}$$

In this section a few variables remained that need to be tweaked in section 5:

- The kernel function.

5 Experiments

In the section 4.3, section 4.4 and section 4.5 some variables remained that need to be tweaked. An oversight of these variables is given:

- The Euclidean distance denoting the segments as being similar.
- The amount of gaps allowed within a connected component.
- The input features.
- Support Vector Machine: Kernel function.
- Neural Network: The amount of hidden layers and hidden neurons.
- Classifier.

The data is divided into a train-set and a test-set. The train-set is used to determine the optimal variables and parameters (section 5.1) and train the data on. The test-set is used to compare the different models. The performance measure can be found in section 5.2.

5.1 Experimental setup

Firstly, the Euclidean distance denoting the segments as being similar and the amount of gaps allowed within a connected component are determined. The values are determined by computing and 3D-plotting the amount of connected components that are nodules over the first 25 scans for different combinations of distances and amount of gaps.
The second step is to determine the optimal subset of features. Wrapper methods and embedded methods are well-known approaches. Wrapper methods can be computationally expensive however for large-scaled data sets, as for every subset a new model needs to be trained. In embedded methods the feature selection is built into the classifier construction and gradient descent method is used to optimize the feature weights. However, this method is rather complex and time-consuming because different models are used. A method that is not only simpler and more efficient, but also yields competitive results with the previous methods is Neighbourhood Component Analysis (W. Yang, 2012). Neighbourhood Component Analysis is a method that works as follows:
Denote the weighted distance between two samples x_i and x_j as

$$d_w(x_i, x_j) = \sum_{r=1}^{p} w_r^2 |x_{ir} - x_{jr}|.$$

A reference point is determined by a probability distribution. The probability that x_j is picked as the reference point for x_i is

$$p_{ij} = \frac{k(d_w(x_i, x_j))}{\sum_{j=1, j\neq i}^{n} k(d_w(x_i, x_j))}.$$

In this equation,

$$k(z) = exp(-\frac{-z}{\sigma}),$$

is a kernel function and the kernel width σ is an input parameter that influences the probability of each points being selected as the reference point.
The chance that the classifier correctly classifies observation i is

$$p_i = \sum_{j=1, j\neq i}^{n} p_{ij} y_{ij}.$$

Here, it holds that

$$y_{ij} = \begin{cases} 1 & y_i = y_j, \\ 0 & \text{other.} \end{cases}$$

The approximate leave-one-out classification accuracy is

$$\begin{aligned} F(w) &= \sum_{i=1}^{n} p_i \\ &= \sum_{i=1}^{n} p_i - \lambda \sum_{r=1}^{p} w_r^2 \end{aligned}$$

Here, a regularization term is added to perform feature selection and alleviate overfitting. The regularization term is λ and is tuned through cross validation. Finding the weight vector can be expressed into a minimization problem. To do this, first assume the following relations:

$$f(w) = -F(w),$$

$$\sum_{i=1}^{n} \sum_{j=1, j=\neq i}^{n} p_{ij} = 1$$

Note that the minimum does not change if you add a constant to $f(w)$.

$$\begin{aligned} \hat{w} &= \arg\min_{w} \{f(w)\} \\ &= \arg\min_{w} \{1 + f(w)\} \\ &= \arg\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} p_{ij} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} p_{ij} y_{ij} + \lambda \sum_{r=1}^{p} w_r^2 \right\} \\ &= \arg\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} p_{ij}(1 - y_{ij}) + \lambda \sum_{r=1}^{p} w_r^2 \right\} \\ &= \arg\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} p_{ij} l(y_i, y_j) + \lambda \sum_{r=1}^{p} w_r^2 \right\} \end{aligned}$$

The loss function $l(y_i, y_j)$ is defined as follows:

$$l(y_i, y_j) = \begin{cases} 1 & y_i \neq y_j \\ 0 & \text{other} \end{cases}$$

The third step of the experiments is to determine the amount of hidden layers and hidden neurons for the Neural Network. Hereby, we used a 3D plot for different amounts of hidden neurons and hidden layers to find the optimal accuracy. The fourth step is to determine the best Kernel function for the Support Vector Machine. 5-fold crossvalidation is applied to avoid overfitting. Here, the train-set is divided into 5 parts. 1 part is used as validation-set and the rest is train-set. This process is repeated until all parts have been a validation-set once. Of the results the mean is taken. In the final step both methods are compared on the test-set. The performance measure that is used is explained in section 5.2.

5.2 Performance Measure

To determine the best variables, a Relative Operating Characteristic curve is used, also known as the ROC-curve. An ROC-curve plots the true positive rate (TPR) against the true negative rate (TNR) for different tresholds. To understand these rates, see figure 14.

		True Class		
		Hit	Non-hit	
Predicted Class	Hit	True Positive (**TP**)	False Positive (**FP**)	PPV= **TP**/(**TP**+**FP**)
	Non-hit	False Negative (**FN**)	True Negative (**TN**)	NPV= **TN**/(**TN**+**FN**)
		TPR= **TP**/(**TP**+**FN**)	TNR= **FP**/(**FP**+**TN**)	

Figure 14: Confusion matrix

In words, the TPR is the proportion of positives that have been assigned as positives and the TNR is the proportion of negatives that have been assigned as positives. The goal is to have a high TPR and a low TNR. With the ROC-curve, the Area Under the ROC Curve value (AUC) can be computed. AUC is the two-dimensional area underneath the ROC-curve. AUC values ranges between 0.5 and 1, where 0.5 is worthless and 1 is excellent. In figure 15 a different ROC's are illustrated. Here, yellow has a good AUC value and blue has a bad AUC value.

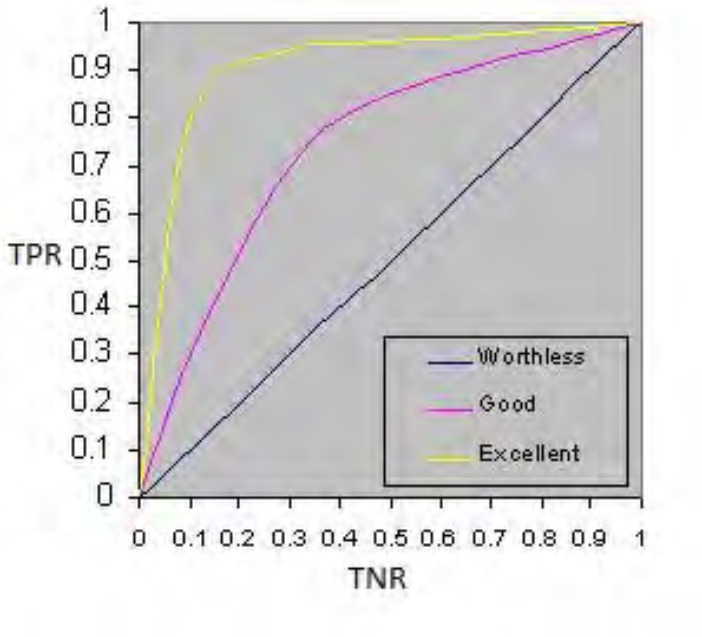

Figure 15: Comparing ROC curves

After the AUC values have been computed, a (paired) t-test is performed to see whether the difference between the models is significant.

6 Results

In this section, (the performance of) the optimal model is determined.

6.1 Connected Component parameters

Firstly, the amount of gaps that are allowed on each side of the random starting point of the connected component algorithm and the distance between two nodules denoting them as similar, is determined. The results can be found in figures 16, 17 and 18.

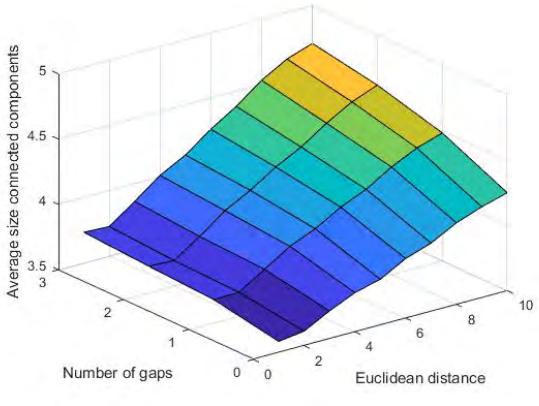

Figure 16: Average number of segments in connected components per scan

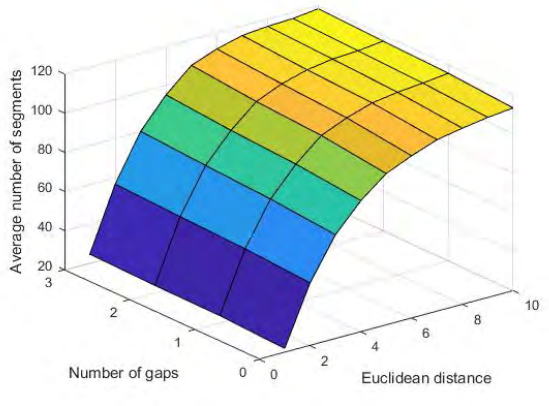

Figure 17: Average number of connected components per scan

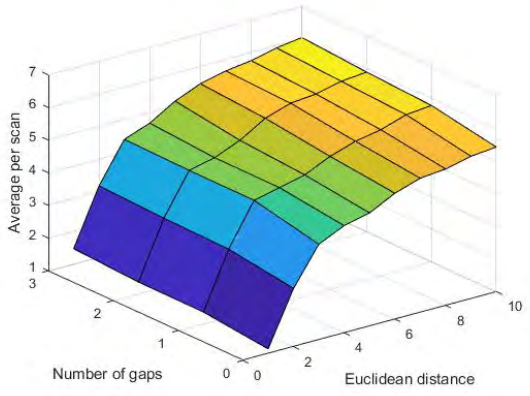

Figure 18: Average amount of segments annotated as nodules in connected component per scan

Figure 16 shows that a large Euclidean distance in combination with a large number of gaps leads to a large number of segments in the connected components. The reason for this is that the algorithm adds segments to the connected component that belong to different lesions. Thus, in order to avoid this, the number of gaps and the Euclidean distance should be kept low.
From figure 17 a similar conclusion can be drawn for the Euclidean distance. A large Euclidean distance results in a large number of connected components because the minimum number of occurrences, which is three, is more easily reached for the small segments that are connected with each other. Changing the number of gaps has little influence in this case.
From section 3.1 we know that on average, a scan has 7,24 lesions annotated as nodule and 2,62 lesions annotated as nodules of at least 3mm. However, the juxta-pleural nodules are included in these average numbers. Therefore, the juxta-pleural nodules are also added in the averages of figure 18.
We see that 7,24 nodules are almost reached by increasing the Euclidean distance and the number of gaps. The remaining must be the small nodules that cannot be connected to at least 2 other segments.

However, in this research paper we are interested in the possibly malignant tumors. The combination that returns slightly more tumors (which is more preferable than a lower amount) on average than 2,62 is using 1 gap on each side and by applying an Euclidean distance of 2. This combination matches the average amount of tumors in the connected components, while the number of components and the size of the components are kept small. When a larger Euclidean distance is used, the small and non-malignant tumors are connected to other lesions, such that the minimum number of occurrences is reached. This is undesirable because we do not wish to classify these non-malignant nodules as a nodule. Besides, a lower Euclidean distance would not include all nodules in to the components.
Because the ratio between components and components with tumors is around 54 to 3, a balanced data set is used for the training of the classifiers. The new balanced ratio is around 9 to 3.

6.2 Feature selection

Neighbourhood Component Analysis is applied to obtain an optimal subset of features. In table 6.2 the results of this method are presented.

Feature	**Weight**
Area	6,063241
Centroid's y-coordinate	3,845947
Standard Deviation	2,964714
Centroid's x-coordinate	2,071998
Mean	1,439018
Orientation	1,120594
Median	0,970196
Skewness	0,970196

Table 4: Optimal subset of features along with feature weights.

The resulting set of relevant features consists of the area, centroid's x-coordinate, centroid's y-coordinate, orientation, mean, median, skewness and standard deviation. The texture based features are thus not included in the model.

6.3 Neural Network

For the Neural network, the amount of hidden layers and hidden neurons need to be determined. In figure 19 a 3D plot is illustrated of the average AUC that is computed with 5-fold cross-validation on the train-set for different hidden layers and hidden neurons.

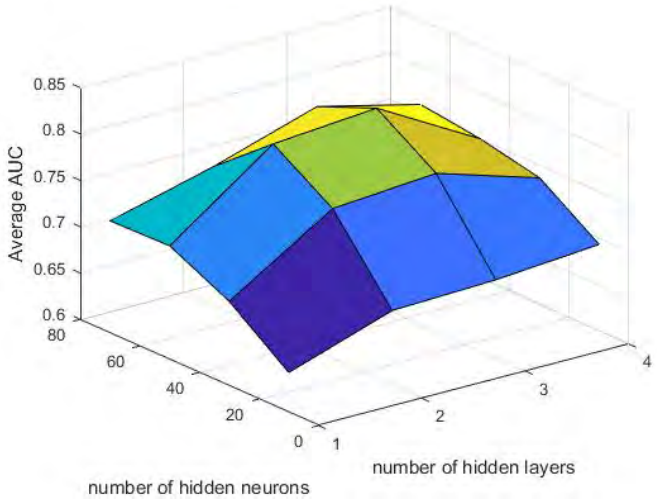

Figure 19: Average AUC for different amounts of hidden layers and hidden neurons

In this figure we see that an increaseing number of hidden layers leads to an increasing average AUC up to 3 hidden layers for all numbers of hidden neurons. From that point, the AUC starts decreasing, which is probably because of overfitting. The same holds for the numbers of hidden neurons: There is an increase up to 50 hidden neurons and the average AUC decreases for a larger amount than 50 hidden neurons. Thus, using 3 hidden layers with each 50 hidden neurons yields the highest accuracy.
To check whether the difference is significant with the results of other combinations, t-tests were performed on the results obtained with the 5-fold cross validation. The difference appeared to be significant as the p-values are smaller than 0.05.

6.4 Support Vector Machine

For the Support Vector Machine the Kernel function needs to be determined. 5-fold cross validation is used and the average AUC is displayed in table 6.4.

Kernel	**Average AUC**
Linear	0.8358
Polynomial	0.5362
RBF	0,7798

Table 5: Average AUC of SVM's with different kernels.

Just like in section 6.3, t-tests were performed on the results obtained with the 5-fold cross validation. The difference appeared to be significant as the p-values are smaller than 0.05.

6.5 Comparison methods

Finally, the two models are trained on the train-set and compared on the test-set:

Kernel	**AUC**
Support Vector Machine	0.8469
Neural Network	0.8192

Table 6: AUC of the methods.

To determine whether this difference is significant, bootstrapping with hundred times sampling is used. On the results a t-test is applied which returned a p-value smaller than 0.05. The difference is thus significant.

7 Conclusion

The goal of this research paper was to create a 2D method with 3D features that reaches a high true positive rate and a low true negative rate. This goal has been met as the best method has an AUC of 0.8469 on the test-set. It is hard to compare the result of this method with other methods because the assumptions that were made in this research are unique.

8 Discussion

This research paper has shown that a 2D method can have a good accuracy score. However, some considerations have to be made before such a method is put into practice.
Only 400 of the 1018 scans were used. The accuracy would have probably increased when more scans would have been used. Besides, the scans with a slice thickness greater than 2,5 have been deleted. In practice this cannot be done obviously. It would have been better to normalize the slice thickness for example. This is left to other research.
The classification of the juxta-pleural nodules cannot be done by the method that is provided in this research because the segmentation process that is chosen is rather simple and cannot separate the juxta-pleural nodules from the pleura. Therefore, in this research the juxta-pleural nodules are separated by placing the segment with the biggest area in a different dataset. Applying classification on this dataset did not return satisfying results.
In this research the strong assumption is made that a possibly malignant tumor occurs in at least 3 slices. In some exceptions the possibly malignant tumors could be missed in the candidate detection because the assumption is not satisfied. For example due to high variation in the centroids' location.

References

A. Gajdhane, L.M. Deshpande (2014). *Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques*.

A. Leung, R. Smithuis (2007). *Solitary pulmonary nodule: benign versus malignant*.

Ahmad, W. (2015). *Radiomics: Extracting more information from medical images using advanced feature analysis*.

Association, American Lung (2014). *Trends in Lung Cancer - Morbidity and Mortality*. URL: http://www.lung.org/assets/documents/research/lc-trend-report.pdf.

Awai, K. (2004). *Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance*.

Bellotti, R. (2007). *A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model*.

Bengio, Y. (2012). *Practical Recommendation for Gradient-based Training of Deep Architecture.*

Bhattacharyya, K. (2016). *Godfrey Newbold Hounsfield (1919–2004): The man who revolutionized neuroimaging*.

Castellino, R. A. (2005). *Computer aided detection (CAD): an overview*.

D. Zinovev J. Feigenbaum, J. Furst D. Raicu (2011). *Probabilistic lung nodule classification with belief decision trees*.

Dehmeshki, J. (2007). *Automated detection of lung nodules in CT images using shape-based genetic algorithm*.

Farag, A. (2011). *Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose CT scans of the chest; Paper presented at the Biomedical Imaging: From Nano to Macro*.

G. Rubin J. K. Lyo, D. S. Paik A. J. Sherbondy (2005). *Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection.*

Haralick, R. (1987). *Image Analysis Using Mathematical Morphology*.

Hossain, M. (2015). *Automatic Lung Tumor Detection Based on GLCM Features*.

I. Sluimer A. Schilham, M. Prokop B. van Ginneken (2006). *Computer analysis of computed tomography scans of the lung: A survey*.

K. Hua C. Hsu, Y. Chen (2015). *Computer-aided classification of lung nodules on computed tomography images via deep learning technique*.

Mathworks (2016a). *Graycoprops*. URL: https://nl.mathworks.com/help/images/ref/graycoprops.html.

— (2016b). *Regionprops*. URL: https://nl.mathworks.com/help/images/ref/regionprops.html.

Messay, T. (2010). *A new computationally efficient CAD system for pulmonary nodule detection in CT imagery*.

Mukhopadhyay, S. (2016). *A Segmentation Framework of Pulmonary Nodules in Lung CT Images*.

N. S. Lingayat, M. R. Tarambale (2013). *A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image*.

NKR (2016). *Incidentie — Long; Landelijk; Man Vrouw; Invasief*. URL: https://www.cijfersoverkanker.nl/selecties/Incidentie_borst/img5b17cb0cd70e6.

Nunzio, G. De (2011). *Automatic lung segmentation in CT images with accurate handling of the hilar region*.

Oda, T. (2002). *Detection algorithm of lung cancer candidate nodules on multislice CT images*.

P. Lambin J. K. Lyo, D. S. Paik A. J. Sherbondy (2008). *Comparison of different feature extraction techniques in content-based image retrieval for CT brain images*.

P. Lin P. Huang, C. Lee M. Wu (2013). *Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model*.

P. Monnin N. Sfameni, A. Gianoli S. Ding (2016). *Optimal slice thickness for object detection with longitudinal partial volume effects in computed tomography*.

Sahiner, B. (2010). *Effect of CAD on Radiologists' Detection of Lung Nodules on Thoracic CT Scans: Analysis of an Observer Performance Study by Nodule Size*.

Singh, S. (2016). *An Evaluation of Features Extraction from Lung CT Images for the Classification Stage of Malignancy*.

T. Buzug A. Schilham, M. Prokop B. van Ginneken (2008). *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*.

W. Yang K. Wang, W. Zuo (2012). *Neighborhood Component Feature Selection for High-Dimensional Data*.

Ye, C. (2009). *Shape-based CT lung nodule segmentation using five-dimensional mean shift clustering and mem with shape information*.