

UNIVERSIDADE FEDERAL DO PARANÁ – UFPR

JEOVANE HONÓRIO ALVES

A LUNG CANCER DETECTION APPROACH BASED ON SHAPE INDEX AND  
CURVEDNESS SUPERPIXEL CANDIDATE SELECTION

CURITIBA

2016

JEOVANE HONÓRIO ALVES

A LUNG CANCER DETECTION APPROACH BASED ON SHAPE INDEX AND  
CURVEDNESS SUPERPIXEL CANDIDATE SELECTION

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, Área de Concentração Sistemas Eletrônicos, Departamento de Engenharia Elétrica, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Engenharia Elétrica

Supervisor: Lucas Ferrari de Oliveira

CURITIBA

2016

A474 Alves, Jeovane Honório  
A lung cancer detection approach based on shape index and  
curvedness superpixel candidate selection / Jeovane Honório Alves.  
Curitiba, 2016.  
91 f.: il., tabs, grafs.

Orientador: Lucas Ferrari de Oliveira  
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de  
Tecnologia, Curso de Pós-Graduação em Engenharia Elétrica.  
Inclui Bibliografia.

1. Câncer - Pulmão. 2. Nódulos - Detecção. 3. Detecção precoce de  
câncer. 4. Diagnóstico por imagem. I. Oliveira, Lucas Ferrari. II. Título. III.  
Universidade Federal do Paraná.

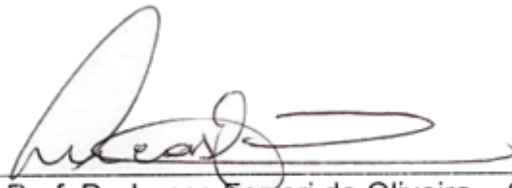
CDD 616.99424

## TERMO DE APROVAÇÃO

JEOVANE HONÓRIO ALVES

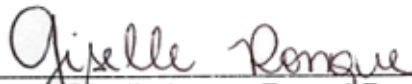
### A LUNG CANCER DETECTION APPROACH BASED ON SHAPE INDEX AND CURVEDNESS SUPERPIXEL CANDIDATE SELECTION

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Paraná.



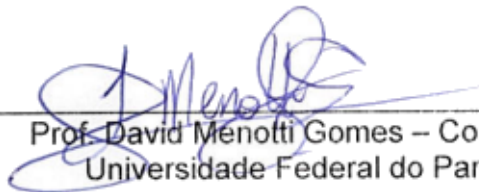
---

Prof. Dr. Lucas Ferrari de Oliveira – Orientador  
Universidade Federal do Paraná



---

Profa. Dra. Giselle Lopes Ferrari Ronque – Convidada  
Universidade Federal do Paraná



---

Prof. David Menotti Gomes – Convidado  
Universidade Federal do Paraná

Curitiba, 29 de agosto de 2016.

## RESUMO

Câncer é uma das causas com mais mortalidade mundialmente. Câncer de pulmão é o tipo de câncer mais comum (excluindo câncer de pele não-melanoma). Seus sintomas aparecem em estágios mais avançados, o que dificulta o seu tratamento. Para diagnosticar o paciente, a tomografia computadorizada é utilizada. Ela é composta de diversos cortes, que mapeiam uma região 3D de interesse. Apesar de fornecer muitos detalhes, por serem gerados vários cortes, a análise de exames de tomografia computadorizada se torna exaustiva, o que pode influenciar negativamente no diagnóstico feito pelo especialista. O objetivo deste trabalho é o desenvolvimento de métodos para a segmentação do pulmão e a detecção de nódulos em imagens de tomografia computadorizada do tórax. As imagens são segmentadas para separar o pulmão das outras estruturas e após, detecção de nódulos utilizando a técnicas de superpixels são aplicadas. A técnica de Rótulamento dos Eixos teve uma média de preservação de nódulos de 93,53% e a técnica *Monotone Chain Convex Hull* apresentou melhores resultados com uma taxa de 97,78%. Para a detecção dos nódulos, as técnicas *Felzenszwalb* e *SLIC* são empregadas para o agrupamento de regiões de nódulos em superpixels. Uma seleção de candidatos à nódulos baseada em *shape index* e *curvedness* é aplicada para redução do número de superpixels. Para a classificação desses candidatos, foi utilizada a técnica de Florestas Aleatórias. A base de imagens utilizada foi a LIDC, que foi dividida em duas sub-bases: uma de desenvolvimento, composta pelos pacientes 0001 a 0600, e uma de validação, composta pelos pacientes 0601 a 1012. Na base de validação, a técnica *Felzenszwalb* obteve uma sensibilidade de 60,61% e 7,2 FP/exame.

**Palavras-chaves:** Câncer de pulmão. Detecção de nódulos. Superpixel. *Shape index*.

## ABSTRACT

Cancer is one of the causes with more mortality worldwide. Lung cancer is the most common type (excluding non-melanoma skin cancer). Its symptoms appear mostly in advanced stages, which difficult its treatment. For patient diagnostic, computer tomography (CT) is used. CT is composed of many slices, which maps a 3D region of interest. Although it provides many details, its analysis is very exhaustive, which may has negatively influence in the specialist's diagnostic. The objective of this work is the development of lung segmentation and nodule detection methods in chest CT images. These images are segmented to separate the lung region from other parts and, after that, nodule detection using superpixel methods is applied. The Axes' Labeling had a mean of nodule preservation of 93.53% and the Monotone Chain Convex Hull method presented better results, with a mean of 97.78%. For nodule detection, the Felzenszwalb and SLIC methods are employed to group nodule regions. A nodule candidate selection based on shape index and curvedness is applied for superpixel reduction. Then, classification of these candidates is realized by the Random Forest. The LIDC database was divided into two data sets: a development data set composed of the CT scans of patients 0001 to 0600, and a untouched, validation data set, composed of patients 0601 to 1012. For the validation data set, the Felzenszwalb method had a sensitivity of 60.61% and 7.2 FP/scan.

**Key-words:** Lung cancer. Nodule detection. Superpixel. Shape index.

## LIST OF FIGURES

FIGURE 1 – Representation of the bronchus and bronchial tree in the lungs . . . . .	6
FIGURE 2 – A slice from a chest CT scan. . . . .	7
FIGURE 3 – Representation of the axes . . . . .	8
FIGURE 4 – Example of the binary morphological operation of dilation . . . . .	12
FIGURE 5 – Example of an erosion . . . . .	12
FIGURE 6 – Opening application in an object . . . . .	13
FIGURE 7 – Closing operation using a $B$ structuring element . . . . .	13
FIGURE 8 – Illustration of the process in the lower hull . . . . .	14
FIGURE 9 – Result of the Monotone Chain application . . . . .	15
FIGURE 10 – Representation of different shapes of a shape index value . . . . .	17
FIGURE 11 – Representation of intensity of curvedness in certain shapes. . . . .	18
FIGURE 12 – Example of SLIC superpixel generation with different images . . . . .	19
FIGURE 13 – Application of the Felzenswalb in a beach photography . . . . .	21
FIGURE 14 – Another example of the Felzenszwalb segmentation . . . . .	22
FIGURE 15 – Basic flow of the methodology. . . . .	39
FIGURE 16 – Representation of the lung segmentation stage. . . . .	40
FIGURE 17 – Overview of the base segmentation flow. . . . .	40
FIGURE 18 – Applying thresholding on the CT scan . . . . .	41
FIGURE 19 – Segmentation of the thorax region . . . . .	42
FIGURE 20 – The thoracic area with the lung inside. . . . .	43
FIGURE 21 – The original thresholded image show in Figure 19 a) after completely segmentation process. Parts of lung were lost. . . . .	43
FIGURE 22 – Representation of the AL approach . . . . .	44
FIGURE 23 – Illustration of the Monotone Chain Convex Hull approach . . . . .	45
FIGURE 24 – Flow of the nodule detection stage . . . . .	46
FIGURE 25 – Example of the AL approach with partial loss of a nodule region . . . . .	51
FIGURE 26 – Example of the AL approach with nodule preservation. . . . .	51
FIGURE 27 – MCCH applied in Figure 25. The nodule was completely preserved, but part of another organ remained. . . . .	52
FIGURE 28 – Bad segmentation of the MCCH approach . . . . .	53
FIGURE 29 – Boxplot of the CT scans' candidates after application of the candidate selection for each superpixel generation approach. . . . .	56
FIGURE 30 – Diameters of the nodules which did not met the requirements of both rules . . . . .	57
FIGURE 31 – Boxplot of the distribution of curvedness mean of nodules. . . . .	58

FIGURE 32 – Application in the validation data set of the candidate selection approach for each superpixel generation approach. . . . . 66



## LIST OF TABLES

TABLE 1 – Approximate values of $HU$ for diverse substances . . . . .	9
TABLE 2 – An overview of recent works about lung nodule detection . . . . .	38
TABLE 3 – Average of superpixels generated and selected per CT scan, showing the amount of superpixels selected if compared to the original amount (in %) . . . . .	55
TABLE 4 – Average of candidates selected labeled as nodules for each superpixel method using the MCCH lung segmentation . . . . .	56
TABLE 5 – Average of true nodule candidates selected. Lungs segmented using the AL approach. . . . .	57
TABLE 6 – Relation of the total of nodules selected from the development data set for $T = 1$ and union (1u) . . . . .	58
TABLE 7 – A more detailed data about nodules selected to the next phase for the MCCH lung segmentation . . . . .	59
TABLE 8 – Results of the nodule candidate selection using the AL segmentation .	59
TABLE 9 – Overall results for classification with three types of superpixel segmen- tation, using the MCCH lung segmentation approach . . . . .	61
TABLE 10 – Overall results for classification based on the AL approach . . . . .	62
TABLE 11 – Sensitivity results considering only the selected nodules. Percentages based on the data from Tables 7 and 8. . . . .	62
TABLE 12 – Overall sensitivity based on the nodule size using the AL lung seg- mentation method and FH superpixel . . . . .	63
TABLE 13 – Overall sensitivity based on the nodule size using the AL lung seg- mentation method and SLIC superpixel . . . . .	63
TABLE 14 – Overall sensitivity based on the nodule size using the AL lung seg- mentation method and SLIC supervoxel . . . . .	64
TABLE 15 – Overall sensitivity based on the nodule size using the MCCH lung segmentation method and FH superpixel . . . . .	64
TABLE 16 – Overall sensitivity based on the nodule size using the MCCH lung segmentation method and SLIC superpixel . . . . .	65
TABLE 17 – Overall sensitivity based on the nodule size using the MCCH lung segmentation method and SLIC supervoxel . . . . .	65
TABLE 18 – Amount of original superpixels generated for each approach and the amount after application of the shape index and curvedness based candidate selection. . . . .	65
TABLE 19 – Relation of the superpixels selected for each approach to the amount of superpixels labeled as nodules. . . . .	66

TABLE 20 – Classification results of the validation data set by different approaches.	67
TABLE 21 – Classification results of the validation data set by different approaches.	67
TABLE 22 – Comparison of different works with our approach. Results from validation stage are shown. . . . .	68

## LIST OF ABBREVIATIONS AND ACRONYMS

2D	2-dimensional
3D	3-dimensional
AIDS	Acquired Immune Deficiency Syndrome
AL	Axes' Labeling
ANOVA	Analysis Of Variance
ASM	Angular Second Moment
AUC	Area Under the Curve
BFGS	Broyden-Fletcher-Goldfarb-Shanno
CAD	Computer-Aided Diagnosis
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CCA	Connected Component Analysis
CLAHE	Contrast Limited Adaptive Histogram Equalization
CT	Computed Tomography
CXR	Chest X-Ray
DICOM	Digital Imaging and Communications in Medicine
DWT	Discrete Wavelet Transform
ELCAP	Early Lung Cancer Action Program
FH	Felzenszwalb-Huttenlocher
FN	False Negative
FNIH	Foundation for the National Institutes of Health
FP	False Positive
GGO	Ground-Glass Opacity
GLCM	Gray Level Co-occurrence Matrix

GTSDMs	Gray-Tone Spatial-Dependence Matrices
HRCT	High-Resolution Computed Tomography
HU	Hounsfield Unit
IDRI	Image Database Resource Initiative
INCA	Instituto Nacional de Câncer
IQR	Interquartile Range
ITK	Insight Segmentation and Registration Toolkit
kV	Kilovolts
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LDCT	Low-Dose Computed Tomography
LIDC	Lung Image Database Consortium
LVQ	Linear Vector Quantization
mA	Milliamperere
MaxA	Maximum Axis Length
MCCH	Monotone Chain Convex Hull
MGRF	Markov-Gibbs Random Field
MinA	Minimum Axis Length
mm	Millimeter
NLM	National Library of Medicine
NNE	Neural Network Ensemble
NSCLC	Non-Small-Cell Lung Cancer
PB	Probably Benign
PCA	Principal Component Analysis
PM	Probably Malignant
PSNR	Peak Signal-to-Noise Ratio

PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RBP	Radial Basis Probabilistic
RG	Region Growing
ROI	Region of Interest
RU	Random Undersampling
SCLC	Small-Cell Lung Cancer
SE	Structuring Element
SFS	Step-wise Feature Selection
SLIC	Simple Linear Iterative Clustering
SMOTE	Synthetic Minority Oversample Technique
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TN	True Negative
TP	True Positive
UC	Uncertain
VQ	Vector Quantization
VTK	Visualization Toolkit
XML	eXtensible Markup Language

# CONTENTS

<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Motivation	2
1.2 Objectives	2
1.3 Contributions	2
1.4 Challenges	3
1.5 Document Structure	3
<b>2 THEORETICAL BASIS</b>	<b>5</b>
2.1 MEDICAL BASIS	5
2.1.1 Lung Cancer	5
2.1.2 LIDC/IDRI Database	8
2.2 COMPUTATIONAL BASIS	10
2.2.1 Digital Image Processing	10
2.2.2 Pattern Recognition and Machine Learning	21
2.2.3 Applications and libraries	28
<b>3 STATE OF THE ART</b>	<b>29</b>
3.1 LUNG SEGMENTATION	29
3.2 NODULE DETECTION AND SEGMENTATION	30
3.2.1 Discussion	37
<b>4 METHODOLOGY</b>	<b>39</b>
4.1 DATA HANDLING	39
4.2 LUNG SEGMENTATION	40
4.2.1 Base segmentation	40
4.2.2 Axes' Labeling	42
4.2.3 Monotone Chain Convex Hull based approach	44
4.3 NODULE DETECTION	45
4.3.1 Superpixel segmentation	46
4.3.2 Nodule candidate selection	47
4.3.3 Feature extraction and classification	49
<b>5 RESULTS AND DISCUSSION</b>	<b>50</b>
5.1 LUNG SEGMENTATION	50
5.1.1 Axes' Labeling	50
5.1.2 Monotone Chain Convex Hull based approach	52

5.1.3 Comparison . . . . .	52
5.2 NODULE DETECTION . . . . .	53
5.2.1 Development stage . . . . .	54
5.2.2 Validation stage . . . . .	63
5.3 DISCUSSION . . . . .	67
<b>6 CONCLUSION . . . . .</b>	<b>70</b>
<b>BIBLIOGRAFY . . . . .</b>	<b>72</b>

## 1 INTRODUCTION

Cancer, also denominated malignant neoplasm or tumor, is a group of diseases known by its enormous growth of abnormal cells and propagation to other parts of the body. Internal factors, such as inherited mutations, hormones, lack of physical activity, obesity, metabolism problems; and external factors, such as radiation, tobacco and chemicals are some causes of cancer. (CENTER; SIEGEL; JEMAL, 2011).

Cancer is one of the leading causes of death worldwide, causing more than the combination of AIDS, Malaria and Tuberculosis deaths (CENTER; SIEGEL; JEMAL, 2011). The GLOBOCAN project aims to collect and estimate data about incidence and mortality of cancer (excluding non-melanoma skin) in 184 countries. Their recent estimates informed there was 14.1 million new cases of cancer worldwide, about 8.2 million deaths and, within 5 years of diagnosis, 32.6 million cases of cancer in 2012. 20% of the new cases and almost 16% of the world deaths occurred in the Americas (FERLAY et al., 2015; VINEIS; WILD, 2014; CENTER; SIEGEL; JEMAL, 2011). According to the Brazilian National Cancer Institute (INCA), the estimated incidence of 2014 was around 576 thousands new cases of cancer (total of 394 excluding non-melanoma skin), 52.44% male and 47.56% female (INCA, 2014). Estimated mortality of 2012 was around 224 thousands deaths (excluding non-melanoma skin), 53.89% male and 46.11% female (FERLAY et al., 2015).

There are many types of cancer and this study will focus on the pulmonary malignant neoplasm. According to estimates of 2012, lung cancer is the most common malignant neoplastic disease, with an incidence of 1.8 million new cases, and the leading cause of cancer deaths, with 1.59 million worldwide. Because its symptoms in early stages are not common, the detection of lung cancer is mainly done in later stages, hindering its treatment and cure, causing high rates of mortality. In Brazil, estimates for lung cancer incidence in 2014 were 27.33 thousands (16.4 and 10.93 thousands for male and female, respectively). As for mortality, 28.3 thousands deaths were the estimation for year 2012, 17.2 and 11.1 thousands for male and female, respectively (FERLAY et al., 2015; INCA, 2014).

Among lung cancer, there are two major types: the majority, non-small-cell lung cancer (NSCLC), representing 85% cases, and small-cell lung cancer (SCLC), representing 15%. NSCLC is grouped into large cell carcinoma, adenocarcinoma and squamous cell (ROY; HERBST; HEYMACH, 2008). Different reasons may cause each type of lung cancer. Smoking (including second-hand) and carcinogens exposure are the mainly causes of lung cancer, so prevention and treatment are necessities. First, reduction (or extinction) of these problems are crucial to pulmonary malignant neoplastic disease cases' diminution. Second, early diagnosis and treatment is encouraged. Usage of lung cancer screening for



early diagnosis is usually done with Chest Radiography, commonly called Chest X-Ray (CXR), and Computed Tomography (CT) (COLLINS et al., 2007; CHILES, 2014).

The CT technique is more suitable to lung cancer diagnosis, as they can be more sensitive to pulmonary nodules than the CXR technique. Chest CTs are taken in the transverse plane, generating many slices, where each slice represents a 2D plane in a specific depth, to make a chest 3D representation. Although the CT sensitivity improves its analysis, the technique produces numerous slices, being tiresome to analyze each one of them, increasing the possibility to miss a nodule, to misdiagnose or even to mark a non-nodule region. Computer-Aided Detection (CADe) systems are developed to assist the radiologist in lung cancer detection, usually by the analysis of CT scans. Another type of system is the Computer-Aided Diagnosis (CADx), which serves as a second opinion in diagnosis of nodules. These systems can be abbreviated as CAD (AWAI et al., 2004).

## 1.1 MOTIVATION

As CT examinations can be exhaustive, since they have many slices to analyze, radiologists may interpret exams incorrectly because of fatigue. Not only that, but lack experience in some situations may lead to misinterpretation. Utilization of CAD systems (CADe, for detection, and CADx, to aid in diagnosis), as a second opinion to radiologists, can improve detection's speed and accuracy, assist in the determination of tumor characteristics and patient's prognosis, reduce exams' workload and increase early detection (FIRMINO et al., 2014).

## 1.2 OBJECTIVES

Aiming to propose a different approach to aid the diagnosis of lung cancer, the objectives of this work are:

- Development of lung segmentation methods for scope reduction;
- Development of automatic nodule detection methods;
- Analysis of results obtained for guidance of further works.

## 1.3 CONTRIBUTIONS

The expected contributions of the methodology proposed are:

- segmentation of lung areas including every nodule present, excluding another unimportant areas;
- detection of nodules with different sizes and shapes;

- a classification protocol using the entire database for training and testing;
- developed code available for analysis and testing purposes, which will be available at <http://web.inf.ufpr.br/vri/alumni/jeovane-honorio-alves-msc-2016>;
- furthermore, to make methods not present in ITK (Insight Toolkit, an open-source toolkit for mostly segmentation and registration in medical images) publicly available, contributing to the ITK and medical imaging communities.

#### 1.4 CHALLENGES

Although recent studies in nodule detection and diagnosis achieved promising results, some points need to be improved for implementation of CAD system in clinical practice (FIRMINO et al., 2014; EL-BAZ et al., 2013). A few challenges are listed below and may increase CAD systems' usage:

- Efficient lung segmentation, decreasing execution time and diminishing possible false positives (FP);
- High sensibility and specificity, aiming to detect every present nodule with few (or preferentially none) FP nodules;
- Communication between scanner, CAD and hospital systems, where the CAD system would receive DICOM files from a scanner, process them and send results to the clinic's system or generate a diagnostic report;
- Develop a standard communication between CAD and hospital systems, facilitating in the process of data exchange;
- High detection speed and low cost of implementation and utilization;
- Possible to detect if a nodule is solid, semisolid or ground glass, and according its location, solitary, juxtapleural or juxtavascular;
- Diagnose a nodule (malign or benign), by its shape and appearance or by its growth, to serve as a second opinion.

#### 1.5 DOCUMENT STRUCTURE

The next chapter, Theoretical Basis, is divided in two sections: medical and computational. Information about lung cancer, diagnosis with CT and details about the public database used in this work are discussed in the medical section. In the second section, image processing, pattern recognition and feature descriptors are explained and their usage are detailed. At last, utilized software are introduced.

The state of the art lung segmentation and nodule detection approaches are discussed in Chapter 3. Their methods, problems and results are detailed, aiming to get their pros and cons for improvement of new methods.

Chapter 4 is about our methodology. Preprocessing techniques applied to chest CT images and segmentation of lung areas are explained in the first section. Nodule detection with the superpixel technique, shape index and curvedness based candidate selection and classification by Random Forest are described.

Results about the extraction of lung areas and nodule detection, then discussion about the proposed method is detailed in Chapter 5. At last, the Conclusion chapter concludes with a brief explanation of the proposed methodology, its results and a description of future works in the lung cancer CAD systems' area.

## 2 THEORETICAL BASIS

This chapter introduces medical and computational concepts that are necessary for understanding of the presented work. These two fields are separated in sections for better organization of the document. First, some concepts about radiology, including Housfield unit (*HU*) and computed tomography (CT), the DICOM format and pulmonary carcinoma are discussed in Section 2.1 where specific concepts about the problem are presented. Next, a basic concept about digital image processing and pattern recognition and machine learning are presented, followed in each subsection by explanations of methods used in this work from these fields.

### 2.1 MEDICAL BASIS

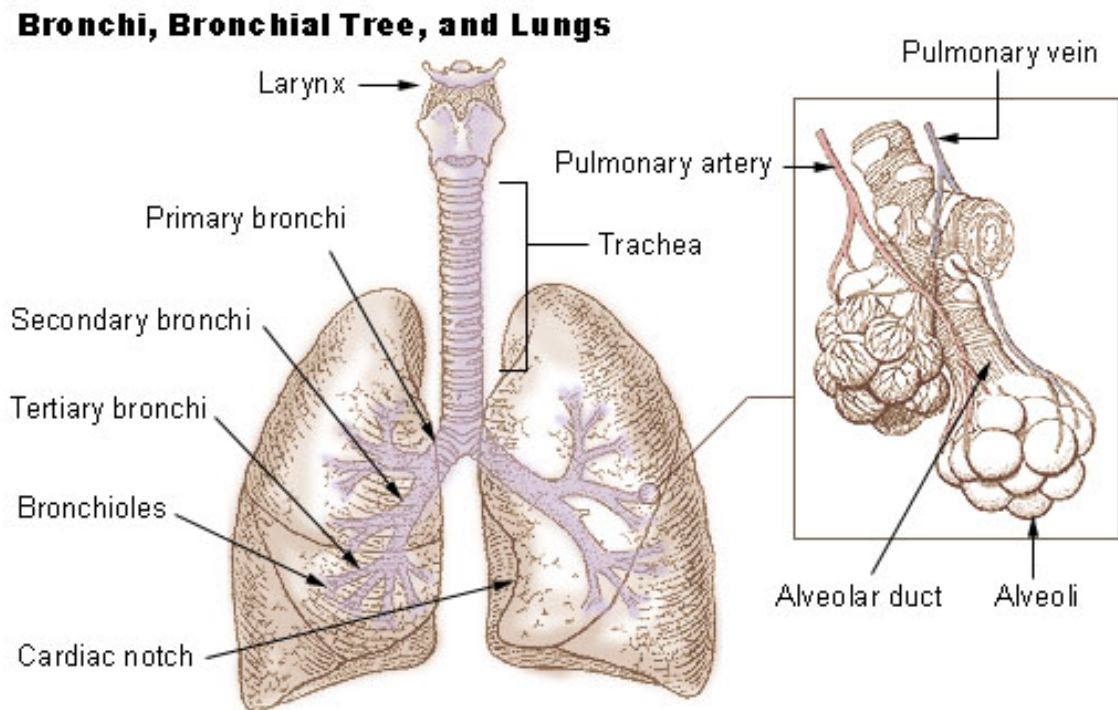
Understanding concepts around the problem is necessary for better development of a solution to it. In this section, concepts regarding lung cancer, as a brief description of the lungs, and its analysis, with computed tomography, are described. Details like anatomy of the lungs, how the DICOM format works, using CAD systems to improve diagnosis, are showed in further subsections. Also, a description of the LIDC database regarding quantity of patients, nodules and the process of annotation is depicted.

#### 2.1.1 Lung Cancer

Lung cancer are malignant neoplastic diseases from the pulmonary region. Grouped in two types, non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). Commonly NSCLC is diagnostically in advanced stages, decreasing the chances of patient survival. NSCLC is sub-grouped in three major types: squamous-cell carcinoma, adenocarcinoma and large-cell lung cancer. Pulmonary tumors from current or former smokers generally are SCLC or squamous-cell carcinoma. Adenocarcinoma is the most common lung cancer for non-smokers. The most common types of lung cancer are squamous-cell carcinoma and adenocarcinoma. Most cases of primary lung cancer and tumors from (former) smokers are developed in the central airway. Non-smoker malignant neoplasms, generally adenocarcinomas, are likely to be developed in the peripheral airways. (HERBST; HEYMACH; LIPPMAN, 2008).

An overall idea of the structure of the lungs can assist in the interpretation of exams and ways to resolve our problems. An illustration of the lungs, primary of the bronchi and bronchial tree belonged to the lungs is shown in Figure 1. As seen in this figure, some parts are outside the lungs, from the larynx to the trachea, then the primary bronchi are found between the inside and outside of the lungs. From beyond the primary

bronchi, the secondary and tertiary ones, then the bronchioles form the bronchial tree inside the lungs. Through the bronchioles, little sacs in their ends are the alveoli, which stores oxygen and carbon dioxide. The membrane which evolves the lungs is called pleura (NHI, 2016).

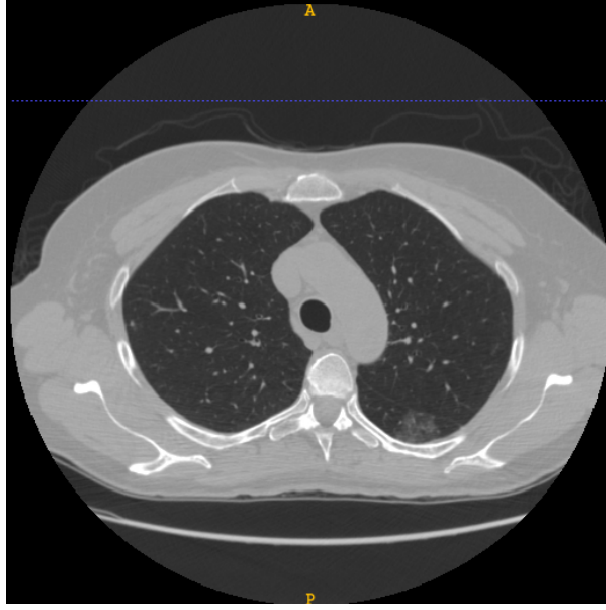


**FIGURE 1** – Representation of the bronchus and bronchial tree in the lungs (NHI, 2016).

Computed tomography (CT) is an imaging procedure for generation of slices from some part of the body via emission in different angles of x-ray beams, projecting an 3D image. CT scans are characterized by higher contrast resolution and less structure noises. CT scans are stored in a standard format. The Digital Imaging and Communications in Medicine<sup>1</sup> (DICOM) format is the standard format for communication of medical imaging, created by the National Electrical Manufacturers Association (NEMA) from the United States. The DICOM format stores data and meta-data (header) of scans from some part of the human body. These meta-data includes information (separately) about the process of image acquisition and the patient (BEUTEL; KUNDEL; METTER, 2000). Figure 2 shows an image of lung CT's exam.

The CT scanner obtains 2D images, slices, through covering the body region of interest. These images have a height and width of  $512 \times 512$ , but may have different height and width physical sizes according to the scanner process, which are stored in the DICOM header as the Pixel Spacing tag. Each slice have a slice thickness, which represents a physical depth of the slice. Through the scanning process, the CT scanner get a slice of the

<sup>1</sup> Informations about the standard is available at <<http://dicom.nema.org/standard.html>>. Last accessed in 18 Jul 2016.



**FIGURE 2** – A slice from a chest CT scan.

body with a specific thickness, averages it and obtains a plane region, which is store in a  $512 \times 512$  image. Slice thickness are typical in the range of  $[1mm, 10mm]$  (lower values can be found, such as some CT scans from the LIDC database) and generally have high values than the pixel spacing, which may interfere in the results of 3D image techniques, as they are processing anisotropic voxels. A solution for this problem is the interpolation of the image for isotropization, i.e. pixel spacing and slice thickness with same sizes (BUZUG, 2008).

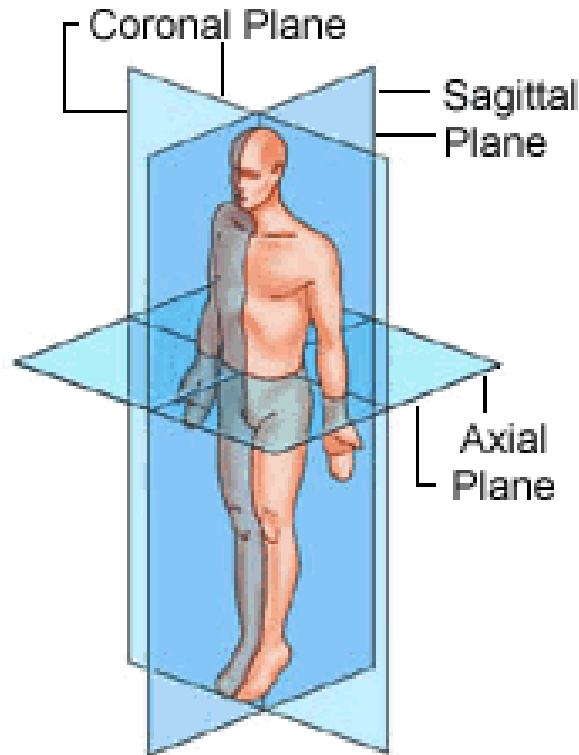
CT scans can be represented in three different axes: axial, coronal and sagittal. Although chest CT scans are gotten in the axial axis, visualizations in coronal and sagittal axes are possible, since the scanning process is 3D and data about it (such as pixel spacing) is available in the header. A representation of the axes is shown in Figure 3.

Data related to the CT scan stored in the DICOM format is represented in the Hounsfield scale (from a CT scan developer, Sir Godfrey Newbold Hounsfield), a intensity scale which represents the density of a tissue calculated based on water, air and its attenuations. Values are presented as Hounsfield Unit ( $HU$ ), also called CT values(BUZUG, 2008; RADIOPAEDIA, 2016). Given the attenuations  $\mu$  from some location,  $\mu_{water}$  and  $\mu_{air}$ , the  $HU$  value of this location can be calculated as:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \quad (2.1)$$

Since the  $\mu$  is being subtracted by  $\mu_{water}$  in Equation 2.1,

$$HU_{water} = 1000 \times \frac{\mu_{water} - \mu_{water}}{\mu_{water} - \mu_{air}} = 1000 \times \frac{0}{\mu_{water} - \mu_{air}} = 0, \quad (2.2)$$



**FIGURE 3** – Representation of the axes. From <<http://www.coloradospineinstitute.com/subject.php?pn=anatomy-anatomical-planes-18>>. Last accessed in 6 Sep 2016.

it can be established the  $HU$  value of water as  $HU_{water} = 0$ , and calculating the  $HU_{air}$ , we obtain:

$$HU_{air} = 1000 \times \frac{\mu_{air} - \mu_{water}}{\mu_{water} - \mu_{air}} = 1000 \times \frac{-1(-\mu_{air} + \mu_{water})}{-\mu_{air} + \mu_{water}} = -1000. \quad (2.3)$$

Table 1 shows values of different tissues (they can have different but close values, dependable of their attenuations).

Usage of Computer-aided Diagnosis (CAD) systems can improve the diagnosis of lung cancer (SONG et al., 2011; CHRISTE et al., 2013). These type of systems, in order to aid radiologists or other specialists, employ automatic or semi-automatic methods for detection and diagnosis of nodules.

### 2.1.2 LIDC/IDRI Database

The LIDC/IDRI lung database is a cooperative work by the Lung Image Database Consortium (LIDC), created by the American National Cancer Institute (NCI) and formed by five institutions, University of California, Los Angeles (UCLA), University of Chicago (U of C), University of Iowa (U of I), University of Michigan (UMich) and Weill Cornell Medical College, and the Image Database Resource Initiative (IDRI), created by the US

**TABLE 1** – Approximate values of  $HU$  for diverse substances. From <http://www.fpnotebook.com/Rad/CT/HnsfldUnt.htm>. Last accessed in 2 Aug 2016.

Substance	$HU$
Air	-1000
Lung	-700
Soft Tissue	-300 to -100
Fat	-50
Water	0
Cerebrospinal fluid	15
Blood	30 to 445
Muscle	40
Calculus	100 to 400
Bone	1000 to 3000 (dense bone)

Foundation for the National Institutes of Health (FNIH), with two academic centers, the University of Texas MD Anderson Cancer Center and the Memorial Sloan Kettering Cancer Center, and eight medical imaging companies, Agfa HealthCare, Carestream Health, Fuji Photo Film Co., Ltd, GE Healthcare, iCAD Inc., Philips Healthcare, Riverain Medical, and Siemens Healthcare, for the development of a public database of CT lung scans (ARMATO et al., 2011).

This database is composed of 1018 helical chest CT scans from 1010 patients, taken by distinct scanners with various configurations. Almost every patient has only one exam in the database – 8 patients have two different exams. As the scans were taken by distinct scanners, some data (like slice thickness) are different between the scans in the database. The peak kilovoltages 120 kV, 130 kV, 135 kV and 140 kV were used to take 818, 31, 69 and 100 CT scans, respectively. The current flow ranged from 40 to 627mA, with a mean of 221.1 mA, reconstruction interval from 0.461 to 0.977 mm, with a mean of 0.688 mm, and pixel spacing from 0.461 to 0.977 mm, mean of 0.688 mm. The soft, standard, slight enhancing and over-enhancing convolution kernels were utilized for 67, 560, 264 and 127 CT scans, respectively.

In the process of nodule annotation, four institutions contributed with a total of 12 radiologists. Objects annotated by four radiologists were classified in three groups:  $\text{nodule} \geq 3\text{mm}$ ,  $\text{nodule} < 3\text{mm}$  and  $\text{non-nodule} \geq 3\text{mm}$ . Nodules with diameter ranging from 3 to 30 mm were classified into the  $\text{nodule} \geq 3\text{mm}$  group. Nodules with diameter less than 3 mm which are possibly malignant ones were grouped into the  $\text{nodule} < 3\text{mm}$ . Last, other pulmonary lesions not classified into nodules with diameter  $\geq 3$  mm were grouped into the  $\text{nodule} \geq 3\text{mm}$ . For nodules  $\geq 3$  mm, radiologists utilized a computer interface to construct the border from the region they think it is part of a nodule, in each slice these nodules were present. For the other two, a central point was obtained. The analysis was realized only in trans-axial sections, as other views were not present in every scan.



A blinded read, which the radiologists were not allowed to share their annotations between them, was realized in the first phase. After, unblinded read, where radiologists may see the annotations from each other and modify or remove theirs, was done. Then, subjective characteristics like internal structure, lobulation, margin, sphericity, spiculation, solidity, subtlety and likelihood (level) of malignancy were scored by four radiologists for every nodule  $\geq 3$  mm. The likelihood of malignancy was attributed with natural numbers between 1 and 5, where 1 was scored when the nodule had a high probability of being benign, and 5 for a high likelihood of malignant. The label 3, when a radiologist was uncertain of this likelihood. These data were available in XML files (one for each series). For each CT scan, the unblinded read data was stored in an XML file, grouped by radiologist. IDs were designated by them to distinct the reads in the file.

The Nodule Size Report<sup>2</sup>, with estimates of nodule volume from every one of them (with at least one nodule  $\geq 3mm$  classification), relations between the IDs from each nodule attributed by the radiologists and some other data, was provided by the LIDC. In cases where one region was analyzed as one nodule by a radiologist and more than one nodule by another radiologist, they were grouped in the same nodule (same line in the report).

A total of 928 nodules were marked as nodule  $\geq 3mm$  by the four radiologists and 2669 by at least one radiologist.

## 2.2 COMPUTATIONAL BASIS

Through each step of lung segmentation and nodule detection, computational techniques are performed, so understanding the process behind them is necessary to a better assimilation of our methodology. In this section, concepts about image processing and pattern recognition used in our approach are presented, as some softwares to execute them.

### 2.2.1 Digital Image Processing

According to Gonzalez & Woods (2008), digital image processing is the field about processing by digital computers of digital images, which may be defined as function  $f(x, y)$ , where  $x$  and  $y$  are spatial discrete – and finite – coordinates with a discrete and finite value, called intensity.

#### 2.2.1.1 Thresholding

Thresholding technique is a method that separated an image's intensities in two or more values based on one or more thresholds (GONZALEZ; WOODS, 2008). For example,

<sup>2</sup> Available at <http://www.via.cornell.edu/lidc/>

given an image  $f(x, y)$  with intensities ranging from  $N$  to  $M$  where  $N < M$ , a threshold  $T$ , where  $N \leq T \leq M$ , is utilized to separate the image's values into two intensities. Given  $T$ , the resulted image  $g(x, y)$  is calculated by:

$$g(x, y) = \begin{cases} 1 & \text{if } f(x, y) > T \\ 0 & \text{if } f(x, y) \leq T \end{cases} \quad (2.4)$$

In the above case, a unique threshold was used to divide the image  $f(x, y)$  in two values, 0 and 1. This approach is called binary thresholding. Other values different from 0 and 1 can be used for thresholding. The equation 2.4 is changed to:

$$g(x, y) = \begin{cases} a & \text{if } f(x, y) > T \\ b & \text{if } f(x, y) \leq T \end{cases} \quad (2.5)$$

To divide the image  $f(x, y)$  intensities in  $n$  values, a multiple thresholding approach is needed. Given the thresholds  $T_1, T_2$  and  $T_3$ , where  $N < T_1 < T_2 < T_3 < M$ , the resulted image  $g(x, y)$  is given by:

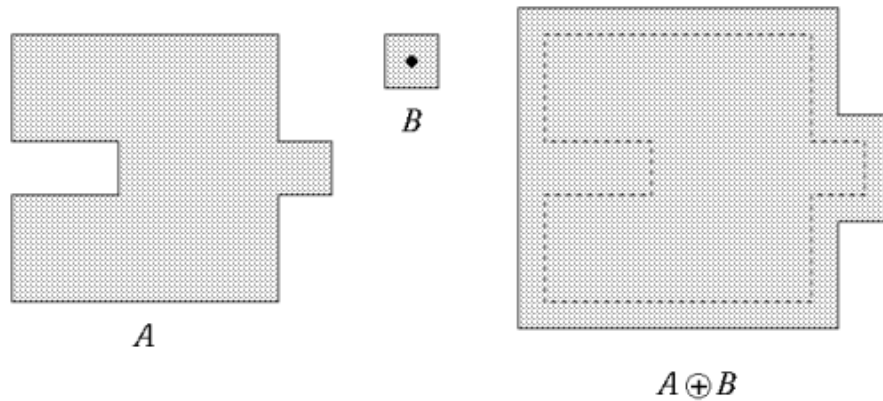
$$g(x, y) = \begin{cases} a & \text{if } f(x, y) > T_3 \\ b & \text{if } T_2 < f(x, y) \leq T_3 \\ c & \text{if } T_1 < f(x, y) \leq T_2 \\ d & \text{if } f(x, y) \leq T_1 \end{cases} \quad (2.6)$$

### 2.2.1.2 Mathematical Morphology

The field which studies processing of shapes from image objects using mathematical properties is called mathematical morphology. Algorithms from this field usually process data using a kernel-like denominated as structuring element (SE). SEs from various shapes (circular, rectangular, cross-like) are used to better attend the problem in question. The two most basic methods from mathematical morphology are erosion and dilation. In a binary image, the dilation operator expands the foreground according to the shape and size of the structuring element and number of iterations this operator is applied (GONZALEZ; WOODS, 2008). Assuming  $A$  is the set (foreground objects) and  $B$  to be the structuring element, the dilation equation is expressed below:

$$A \oplus B = \{z | [(\hat{B})_z \cap A] \subseteq A\}, \quad (2.7)$$

Illustration of this process is present in Figure 4.

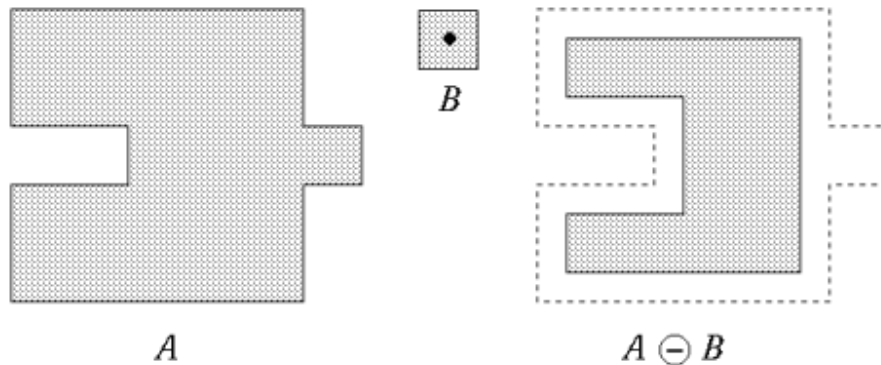


**FIGURE 4** – Example of the binary morphological operation of dilation. From <http://www.inf.u-szeged.hu/ssip/1996/morpho/morphology.html>. Last accessed in 6 Sep 2016.

Contrary to the dilation, the erosion operation reduces the foreground objects based on the  $B$  structuring element. Its equation is given below:

$$A \ominus B = \{z | (B)_z \subseteq A\}. \quad (2.8)$$

Figure 5 shows an example of the erosion operation in an object.

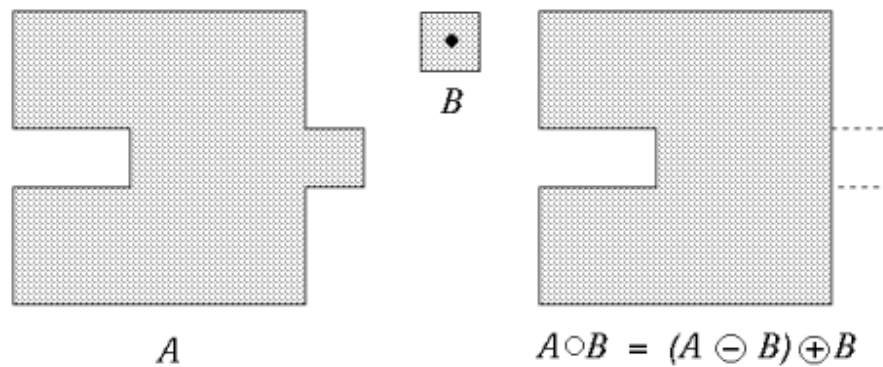


**FIGURE 5** – Example of an erosion. From <http://www.inf.u-szeged.hu/ssip/1996/morpho/morphology.html>. Last accessed in 6 Sep 2016.

Another common morphological operators are the opening and closing ones. They are combinations of the erosion and dilation ones. The opening operator can be defined by Equation 2.9, given below:

$$A \circ B = (A \ominus B) \oplus B \quad (2.9)$$

Generally, the opening operator smooths contours, removes spikes and isthmuses. An example of this operator can be seen in Figure 6.

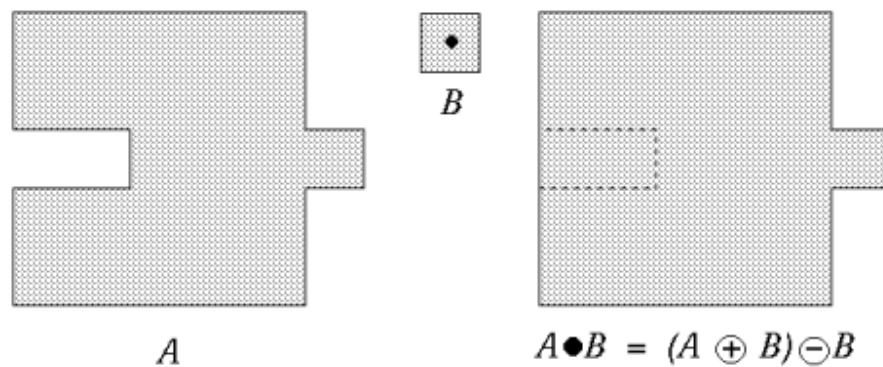


**FIGURE 6** – Opening application in an object. From <http://www.inf.u-szeged.hu/ssip/1996/morpho/morphology.html>. Last accessed in 6 Sep 2016.

The closing operator, also a combination of the dilation and erosion operators, is defined below:

$$A \bullet B = (A \oplus B) \ominus B \quad (2.10)$$

This operator can also smooth contours but, differently from the opening operator, focus in elimination of small holes and fills gaps in the object's contour. Figure 7 illustrates this operation.

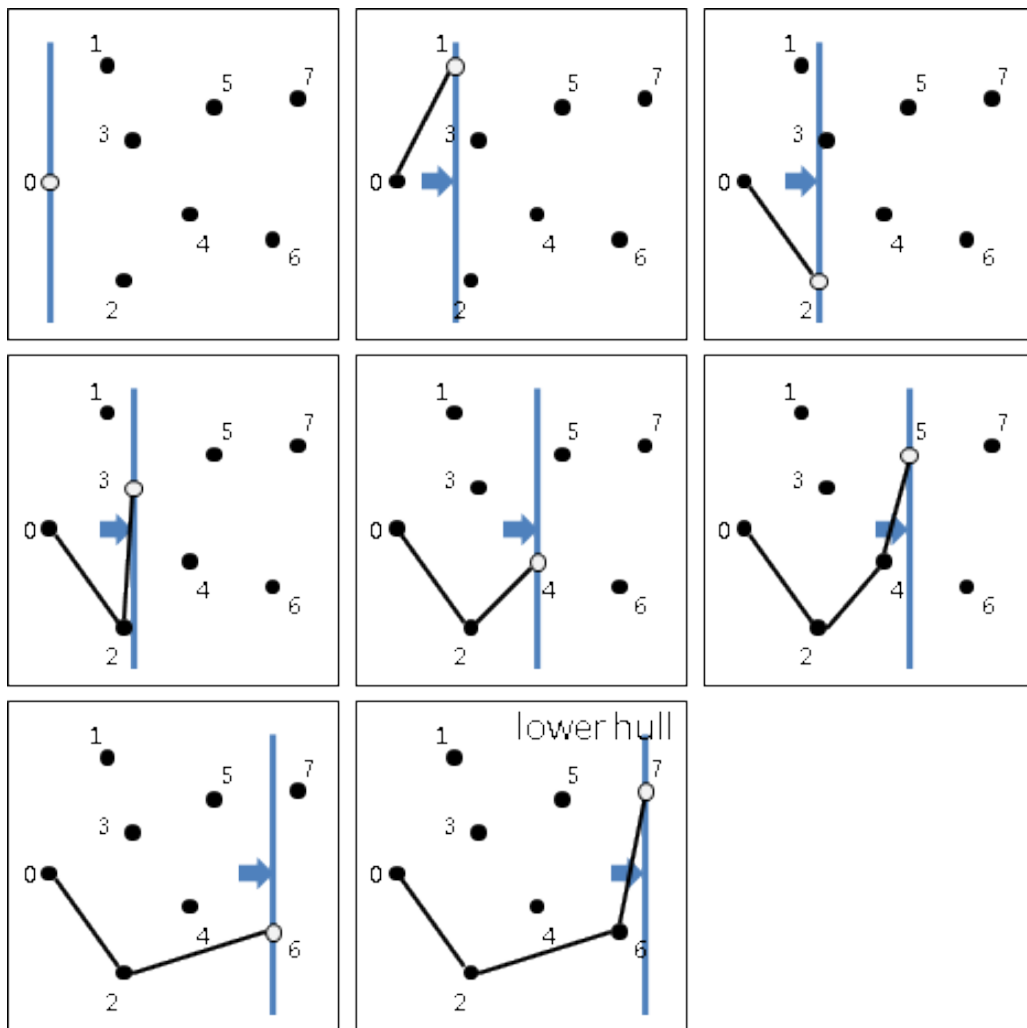


**FIGURE 7** – Closing operation using a  $B$  structuring element. From <http://www.inf.u-szeged.hu/ssip/1996/morpho/morphology.html>. Last accessed in 6 Sep 2016.

### 2.2.1.3 Monotone Chain Convex Hull

The convex hull of an object is the smallest convex region containing it. There are many techniques to obtain the convex hull, and one of them is the monotone chain, created by Andrew (1979). This technique sorts the points of an image like  $f(x, y)$ , first ordering by  $x$  and, if two points have the same value for  $x$ , the ordering is realized by the value of  $y$ . Then, the leftmost and rightmost points are found to create two chains, one for the upper hull and another for the lower hull, and the object's points are attributed

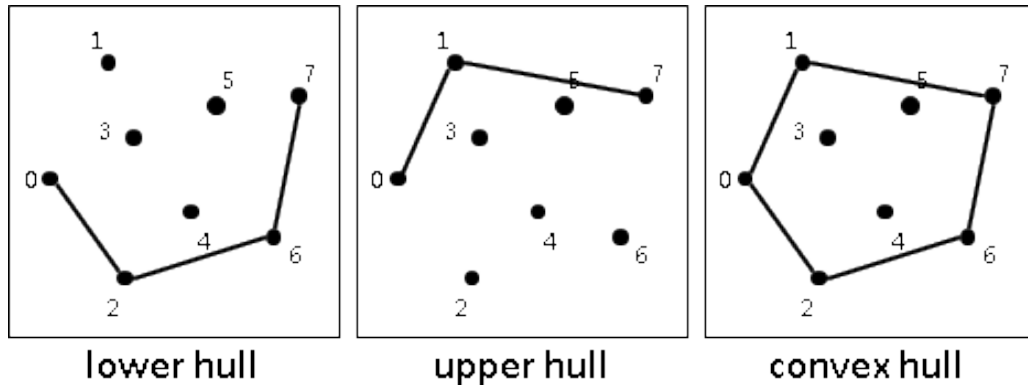
to one of them. To realize this, we iterate the sorted points ascending and descending to attribute the convex points to the upper and lower hull, respectively. For each point, a counter-clockwise verification between it (point A) and the two last points of the current hull (B and C, orderly) is done. Until this is true and the current hull has at least two points, the last point of the current hull is removed. After this, the point A is inserted in the last position of the current hull. This is realized until the entire sorted points are iterated (SUNDAY, 2010). Figure 8 illustrates the process in the lower hull. After the process is realized in both hulls, the convex hull is obtained (Figure 9).



**FIGURE 8** – Illustration of the process in the lower hull. From <http://www.csie.ntnu.edu.tw/~u91029/ConvexHull.html>. Last accessed in 6 Sep 2016.

#### 2.2.1.4 Gaussian filter

In image processing, a filter commonly used for noise reduction and smoothing is the Gaussian. For obtaining of partial derivatives of an image, which amplifies high frequencies (and noises too), the Gaussian filter can be applied before as a low-pass filter



**FIGURE 9** – Result of the Monotone Chain application. From <http://www.csie.ntnu.edu.tw/~u91029/ConvexHull.html>. Last accessed in 6 Sep 2016.

(HALE, 2006). An one-dimensional Gaussian can be calculated by:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (2.11)$$

Number of calculations will vary according to the value of  $\sigma$ . Through the steps of our methodology, a recursive implementation<sup>3</sup> of the Gaussian filter is used. This implementation has a fixed number of calculations, producing faster results with large values of  $\sigma$ . Also, partial derivatives of first and second order can be calculated (YOUNG; VLIET, 1995).

#### 2.2.1.5 Anti-geometric diffusion

The anti-geometric diffusion (or anti-geometric heat flow) was proposed by (MANAY; YEZZI, 2003) for adaptive thresholding and fast segmentation, aiming to smear the edges. Calculation (Equation 2.12) of the anti-geometric diffusion can be achieved using the first and second order derivatives obtained from the Gaussian filter.

$$\frac{\partial I}{\partial t} = \frac{I_x^2 I_{xx} + 2I_x I_y I_{xy} + I_y^2 I_{yy}}{I_x^2 + I_y^2} \quad (2.12)$$

#### 2.2.1.6 Shape index and curvedness features

Calculation of shape index and curvedness features are realized using eigenvalues extracted from a calculated Hessian matrix, which contains the partial derivatives in many directions and orders. The Hessian matrix for a voxel  $p(x, y, z)$ , in a 3D image, can be

<sup>3</sup> RecursiveGaussianImageFilter class from the Insight Toolkit. Available at [https://itk.org/Doxygen/html/classitk\\_1\\_1RecursiveGaussianImageFilter.html](https://itk.org/Doxygen/html/classitk_1_1RecursiveGaussianImageFilter.html). Last accessed in 15 Jul 2016

calculated as follows:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} \quad (2.13)$$

Then, eigenvalues are extracted from the Hessian matrix. But before calculation, it is desirable to understand what is the eigenvalues, and its related feature, the eigenvectors. Say, for example, we have a matrix  $A$  and want to power it to higher values (100, e.g.). Multiply a matrix for itself too many times can be exhaustive. The resulted matrix can be attained using the eigenvalues. Some vectors  $\mathbf{x}$  do not change direction, when multiplied by matrix  $A$  (they have the same direction as  $A\mathbf{x}$ ), are the eigenvectors (STRANG, 2016). Given a scalar  $\lambda$ , the vector  $A\mathbf{x}$  is calculated as follows:

$$A\mathbf{x} = \lambda\mathbf{x}, \quad (2.14)$$

where the  $\lambda$  value is an eigenvalue of  $A$ . This eigenvalue  $\lambda$  can express if a vector  $A\mathbf{x}$  was changed (stretch, reversed, shrunk) or not. Calculation of the eigenvectors and eigenvalues of a matrix are explained in the following. Given a  $2 \times 2$  matrix  $A$ ,

$$A = \begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix}, \quad (2.15)$$

the eigenvalues  $\lambda$  can be obtained equaling the determinant of the matrix  $\det(A - \lambda I)$  to zero.

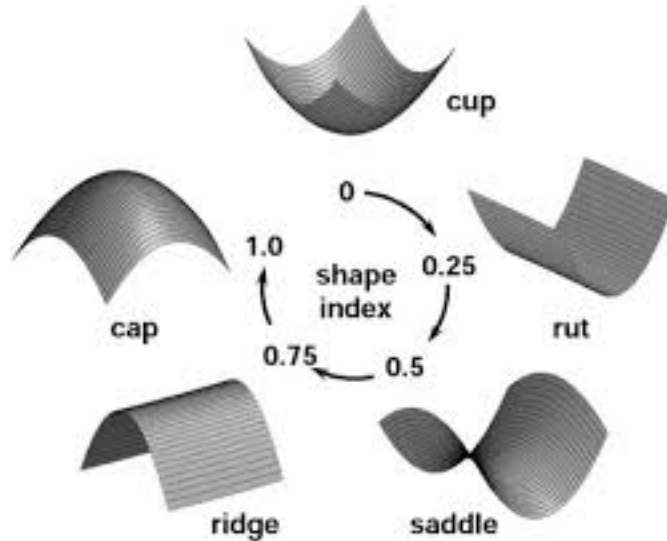
$$\det(A - \lambda I) = \det \left[ \begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right] = \det \begin{bmatrix} 5 - \lambda & 1 \\ 4 & 2 - \lambda \end{bmatrix}, \quad (2.16)$$

thus obtaining the polynomial  $\lambda^2 - 7\lambda + 6$  and its roots 1 and 6, that is, the eigenvalues of  $A$ . To find the eigenvector  $\mathbf{x}$  related to each eigenvalue, solve the equation  $(A - \lambda I)\mathbf{x} = 0$ . The maximum and minimum eigenvalues, the principal curvatures  $k_1(p)$  and  $k_2(p)$  respectively, are used to calculate the values of shape index and curvedness (SPIVAK, 1999).

Shape index describes the shape of a voxel  $p(x, y, z)$  into an interval of  $[0, 1]$ . Certain values can represent different shapes. The shapes represented are cup (0), rut (0.25), saddle (0.5), ridge (0.75) and cap (1), showed in Figure 10. A thing to consider is a voxel can have a shape index value between values of two shapes (0.85 e.g., being

something between a cap and a ridge, but more like a cup). Given the principal curvatures  $k_1(p)$  and  $k_2(p)$  ( $k_1(p) \geq k_2(p)$ ), the shape index is calculated as follows:

$$S(p) = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_2(p) + k_1(p)}{k_2(p) - k_1(p)} \quad (2.17)$$



**FIGURE 10** – Representation of different shapes of a shape index value (YOSHIDA et al., 2002).

Curvedness feature at determined voxel  $p(x, y, z)$  can be calculated as follows (NAPPI; FRIMMEL; YOSHIDA, 2005):

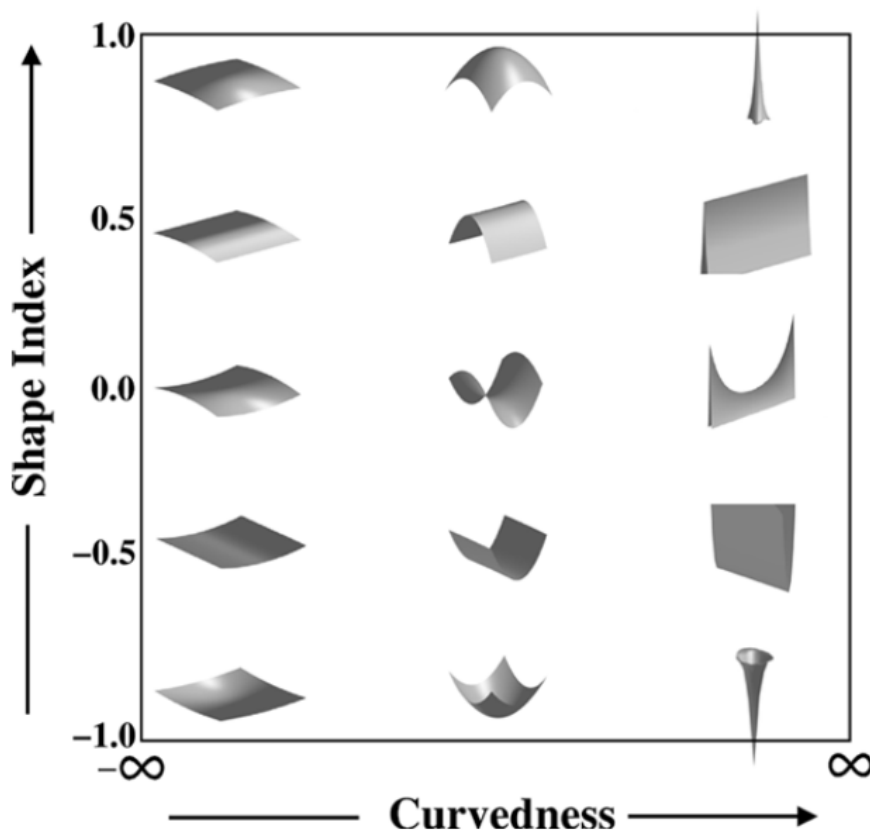
$$C(p) = \frac{2}{\pi} \ln \sqrt{\frac{k_1(p)^2 + k_2(p)^2}{2}} \quad (2.18)$$

### 2.2.1.7 SLIC Superpixel

Superpixel is a technique for grouping of pixels in an image which can be used to replace the usage of grids for segmentation or local analysis. Achanta et al. (2012) developed the Simple Linear Iterative Clustering (SLIC) superpixel algorithm, based on  $k$ -means clustering. According to the authors, SLIC has a good adherence to borders and it is an interesting choice for segmentation, outperforming others state-of-the-art superpixels methods in many aspects, such as boundary recall, segmentation speed and under-segmentation error. A simple algorithm describing the SLIC method is present below with a detailed explanation:

First, values of step  $S$  and weight  $m$  are chosen. The  $S$  value means the step the algorithm will take to initialize the cluster centers and the initial dimension of the clusters (the authors calculated this value based on the desired number of superpixels, with  $S = \sqrt{N/k}$ , where  $N$  is the number of pixels and  $k$  is the number of desired superpixels).





**FIGURE 11** – Representation of intensity of curvedness in certain shapes. Through the interval is set to  $[-1, 1]$ , it describes the previous reported shapes (NAPPI; FRIMMEL; YOSHIDA, 2005).

To each cluster center, it was found the lowest gradient pixel in an area of  $3 \times 3$  from the center and this pixel is chosen as the new cluster center (its intensity is stored too). According to the authors, this was realized to avoid cluster centers at borders. For each pixel, two values are stored: cluster label and its distance to the cluster center. These values are initialized with a non label value (for cluster label) and the maximum possible value (cluster distance). This is realized to make sure that every pixel has a cluster tied to it.

For the next part, the clusters are iterated  $t$  times to get the correct shapes of superpixels. Authors have chosen ( $t = 10$ ), as many tests executed by them demonstrated it has great results and good performance. For each cluster, the  $2S \times 2S$  window around the cluster center is analyzed for distance calculation. For each pixel from this area, the Euclidean distance between the cluster center is calculated. Distances are calculated based on the spatial and intensity values. The authors used the LAB model for color distance calculation. Its equation is given below:

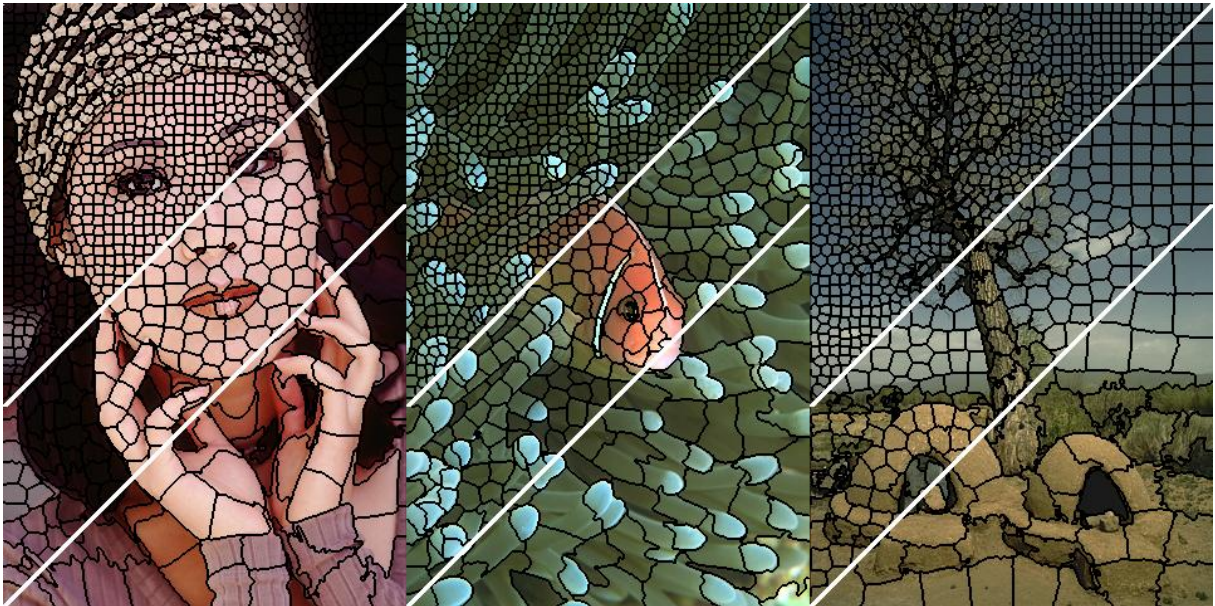
$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2}. \quad (2.19)$$

---

**Algorithm 1** SLIC superpixel algorithm
 

---

- 1: Set values of step  $S$  and weight  $m$
  - 2: Set loop threshold  $t$  to 10 and variable  $i$  to 1
  - 3: Generate clusters centers within  $S \times S$  areas
  - 4: Move them to the lowest gradient in an area of  $3 \times 3$
  - 5: Reset cluster label and distance from each image pixel
  - 6: **while**  $i \leq t$  **do**
  - 7: **for** each cluster **do**
  - 8: **for** each pixel within an area of  $2S \times 2S$  around the cluster center **do**
  - 9: Calculate distance between pixel and cluster center
  - 10: **if** distance from current cluster center  $<$  distance from pixel's nearest cluster center **then**
  - 11: Change the pixel's cluster label and distance to the current one
  - 12: **end if**
  - 13: **end for**
  - 14: **end for**
  - 15: Update new cluster centers
  - 16: Increment value of  $i$
  - 17: **end while**
  - 18: Connectivity is enforced
- 



**FIGURE 12** – Example of SLIC superpixel generation with different images. From <http://ivrl.epfl.ch/research/superpixels>. Last accessed in 6 Aug 2016.

Distance in gray-scale images can be calculated with the equation below:

$$d_c = \sqrt{(g_j - g_i)^2}. \quad (2.20)$$

For spatial distance calculation, the following equation is used:

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}. \quad (2.21)$$

Generation of superpixel from 3D images, called supervoxels, can be done by the SLIC algorithm with the following equation for calculation of the  $d_s$  value:

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2} \quad (2.22)$$

The distance  $D'$  between pixel and cluster center is the result of a combination of Equations 2.19 (2.20 for gray-scale images) and 2.21 (2.22 for 3D images), but a normalization of these values is necessary. Normalization is realized with the maximum distances within a cluster of spatial  $N_s$  and color distance  $N_c$ . According to the author, the maximum spatial distance should correspond to the  $S$  value,  $N_s = S$ , but the maximum color distance would vary too much between clusters. Then, they used a weight  $m$  parameter for color distance normalization. The equation below is used to calculate the distance  $D'$ :

$$D' = \sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2} \quad (2.23)$$

This  $m$  parameter can balance the importance of spatial and intensity distances. A high value prioritizes the spatial distance and a low value the intensity. If the distance from the current cluster center is less than the distance from the pixel's nearest cluster center, the pixel's cluster label and distance are changed. After the iteration, a new cluster center (and intensity) is calculated based on the mean from the sum of pixels belonging to the cluster for every one of them. At last, a connected component algorithm is applied to enforce connectivity between pixels from each superpixel. If a pixel or region of pixel is not directly connected to the belonged superpixel, then it is attached to the nearest superpixel.

#### 2.2.1.8 Felzenszwalb-Huttenlocher Segmentation

The Felzenszwalb-Huttenlocher (F-H) segmentation (or Felzenszwalb segmentation) is an image segmentation approach based on graphs developed by Felzenszwalb & Huttenlocher (2004). The idea of this method is a fast formation of components while adapting its segmentation process according to the variability of local intensities. Thus, regions with high variance are correctly segmented. Figures 13 and 14 show examples of the Felzenszwalb segmentation.

With a graph based approach, each pixel is treated as a vertex, and edges are generated with its neighbor pixels. Each edge has a weight, which is the absolute difference between the intensities of two vertices.

1. A preprocessing step is executed prior to the segmentation. A Gaussian filter with a low  $\sigma$  (the authors stated that the  $\sigma = 0.8$  is the value used by them) is applied to the image, for noise reduction without compromising the image information.

2. Generate a vector  $E$  with edges and their weights and sort them into a non-decreasing order
3. Initialize components, where each vertex is part of a different component. Set component's threshold with a value equal to  $k$ .
4. For each edge inside  $E$ , which its two vertices have different components, verify if the edge weight is less or equal than the minimum threshold of the these components. If true, the components are merged.
5. Repeat step 4 until the entire  $E$  has been iterated.
6. Although not stated in the method's article but in the code available in the website from one of the authors<sup>4</sup>, a minimum size is determined. If some component has a size less than this value, it is annexed to the nearest component (obtained in the edge with the lowest weight related to the component).



**FIGURE 13** – Application of the Felzenszwalb in a beach photography. Parameters:  $\sigma = 0.5$ ,  $K = 500$ ,  $min = 50$ . From <http://cs.brown.edu/~pff/segment/>. Last accessed in 6 Sep 2016.

### 2.2.2 Pattern Recognition and Machine Learning

Feature is a measure value to represent a pattern from groups of samples, aiming to improve the distinction between classes, samples' groups of interest, which may be denominated as label. These features are extracted using different methods, such as texture,

<sup>4</sup> To download the code, visit <http://cs.brown.edu/~pff/segment/index.html>



**FIGURE 14** – Another example of the Felzenszwalb segmentation. Parameters:  $\sigma = 0.5$ ,  $K = 1000$ ,  $min = 100$ . From <http://cs.brown.edu/~pff/segment/>. Last accessed in 6 Sep 2016.

shape and size feature descriptors. Then, samples with their feature vectors and their labels serves as an input for the classifier, a technique for classification, labeling a sample with unknown class with one from a determined range. The example given is part of the supervised learning. Supervised learning is a pattern recognition task where the labels for training are available, which gives the classifier options for sample classification. There may be cases where the labels are not known. This problem is known as unsupervised learning (or clustering). Data set is a group of samples with  $N$  features and labels (classifications with samples having more than one label are denominated as multi-label classification), which are divided in different ways for training and testing. The process when the data set is divided into two subsets, one for training and another for testing, is called holdout method. Another approach is the leave-one-out method. Given a data set with  $N$  samples, for each one them, a sample is taken from this data set and the remaining samples ( $N - 1$ , excluding the current one) serves as a training set to test the left sample. (THEODORIDIS; KOUTROUMBAS, 2008; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; DUDA; HART; STORK, 2012)

Diverse ways of evaluating a classification can be employed, but selection of the appropriated metric for a specific problem is necessary, as the wrong choice may lead to incorrect representation of the solution, such as describing a high accuracy with a unbalanced data set biased towards the majority class. Considering a binary classification problem, where there is samples from a positive (P) or from a negative (N) class, true positive (TP) and true negative (TN) are correctly classification of a sample from the positive and negative classes, respectively. False positive (FP) is when a sample from the negative class is predicted as from the positive class and false negative (FN) is the inverse (POWERS, 2011). According to these definitions, some popular metrics are defined below:

- Sensitivity, recall, true positive rate (TPR):

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (2.24)$$

- Specificity, inverse recall, true negative rate (TNR):

$$TNR = \frac{TN}{N} = \frac{TN}{FP + TN} \quad (2.25)$$

- Precision, confidence, true positive accuracy (TPA):

$$TPA = \frac{TP}{TP + FP} \quad (2.26)$$

- Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.27)$$

One problem in works involving classification is the class imbalance, which can be defined as the presence of a difference between the number of samples from one class if compared to another (class imbalance can be found in multi-class approaches too). This problem appears in many types of applications, such as text classification and diagnosis of diseases. Class imbalance can impact negatively in results, focusing more in the majority class. Classifications may present high accuracy because the classifier is biased to the majority class and the wrong classification of minority class sample did not presented high changes to the accuracy (some cases with class imbalance would use other evaluations, as the accuracy did not correctly represented the classification results). Class imbalance problems are commonly treated with two main approach, resampling and cost sensitive (THEODORIDIS; KOUTROUMBAS, 2008).

Resampling is the technique for amplification (oversampling) and/or reduction (undersampling) of samples of one or more classes to balance the amount of samples for each class. This resampling can be realized randomly or using different calculations based on the original samples. Some problems can be seen in this approach. First, undersampling the majority class may lead to loss of important information. Second, depending on how the class distribution is done, the classification would be impacted negatively.

Another example is the cost-sensitive approach, where classifiers are modified to be adapt to the class imbalanced by applying different weights according to the amount of samples for each class.

### 2.2.2.1 Random Forest

Random Forest is a classification and regression learning method based on decision tree and random selection of features. At each tree, a random vector of features serves as an input for class prediction. (BREIMAN, 2001)

For better classification of samples, good features are needed to be extracted. Some features are described below, which will be used for the development of our approach.

### 2.2.2.2 Shape and Statistical Features

Features like shape, size and statistical are important types for classification of samples, like for example, diagnosis of a nodule into benign and malign, as some features may have different patterns between classes (THEODORIDIS; KOUTROUMBAS, 2008). Below some features are defined:

- Perimeter: Given a boundary with  $N$  points and each point is represented as  $x_i$ , calculation of the object's perimeter  $P$  is compute by

$$P = \sum_{i=1}^{N-1} \|x_{i+1} - x_i\| + \|x_N - x_1\| \quad (2.28)$$

- Area: A way to calculate the area  $A$  of an object is simply count the pixels inside its boundary. Considering we are working with an 2D binary mask image  $M(x, y)$ , where values inside the object have an intensity  $M(x, y) = 1$ ,  $M(x, y) = 0$  otherwise, calculation of the area of an object is given by

$$A = \sum_{x,y} M(x, y) \quad (2.29)$$

- Volume: Like the area, the volume  $V$  of an object can be calculated with voxel counting. Computation is given by

$$V = \sum_{x,y,z} M(x, y, z) \quad (2.30)$$

- Physical Size: Since CT scans may have different pixel spacing ( $p_x, p_y$ ) and slice thickness ( $s$ ), a more close calculation to the real volume  $V_{real}$  is given below:

$$V_{real} = \sum_{x,y,z} M(x, y, z) p_x p_y s \quad (2.31)$$

- Roundness: Computation of the roundness  $\gamma$  of an object is realized using the following equation:

$$\gamma = \frac{P^2}{4\pi A} \quad (2.32)$$

- Sum of intensities: Given a gray-scale image  $I(x, y, z)$  and the binary mask  $M$ , the sum of intensities of an object is calculated as

$$S = \sum_{x,y,z} I(x, y, z)M(x, y, z) \quad (2.33)$$

- Maximum and minimum: The range of intensities of an object.
- Mean:

$$\mu = \frac{S}{V} \quad (2.34)$$

- Standard deviation:

$$\sigma = \sqrt{\frac{1}{V-1} \sum_{x,y,z} (I(x, y, z) - \mu)^2}, \quad (2.35)$$

- Equivalent spherical radius and perimeter: Represent the radius and perimeter of a spherical object with same size than the object of interest.
- Elongation: Result of the largest principal moment divided by the smallest one.

### 2.2.2.3 Hu moments

The seven Hu moments were created by (HU, 1962) and can be used for feature description of images. They are invariant to rotation, scaling and translation, being useful for description of images with these type of variations. Hu moments are based on central moments, thus in the geometric moments. Definition of these moments is necessary to define the seven Hu moments. Given a digital image  $I(x, y)$ , a geometric moment of order  $p + q$  is calculated as

$$m_{pq} = \sum_x \sum_y x^p y^q I(x, y) \quad (2.36)$$

These moments depends on the image coordinates, thus are variant to rotation, scaling and translation. Invariance to translations was proposed with the normalized central moments. One thing to note is the true invariance is only for analog images. Digital images have approximation error according to the sampling from analog to digital (THEODORIDIS; KOUTROUMBAS, 2008). Calculation of a central moment  $\mu_{pq}$  is achieved using the moments  $m_{00}$ ,  $m_{01}$  and  $m_{10}$  for the definition of  $\bar{x}$  and  $\bar{y}$ , being calculated as

$$\bar{x} = \frac{m_{10}}{m_{00}}, \quad (2.37)$$



$$\bar{y} = \frac{m_{01}}{m_{00}}, \quad (2.38)$$

getting the following equation:

$$\mu_{pq} = \sum_x \sum_y I(x, y)(x - \bar{x})^p (y - \bar{y})^q \quad (2.39)$$

A new value  $\gamma$ , based on the vales  $p$  and  $q$ , is necessary to calculated the normalized central moments.  $\gamma$  is computed as

$$\gamma = \frac{p + q + 2}{2} \quad (2.40)$$

The normalized central moments, invariant to scaling and translation, are calculated using the following equation:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma} \quad (2.41)$$

Now that calculation of the normalized central moments was defined, calculation of the seven Hu moments is possible. These moments, being calculated from the normalized central moments, are invariant to scaling and translation, but also to rotation and reflection (the last moment changes it sign). Hu moments are represented with the Greek letter  $\phi$ . The first two moments,  $\phi_1$  and  $\phi_2$ , are of order  $p + q = 2$ , and the remaining moments of order  $p + q = 3$  (THEODORIDIS; KOUTROUMBAS, 2008). Definition of the seven moments is given below:

$$\begin{aligned} \phi_1 &= \eta_{20} + \eta_{02} \\ \phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ \phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2 \\ \phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{03} + \eta_{21})^2 \\ \phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{03} - 3\eta_{21})(\eta_{03} + \eta_{21})[(\eta_{03} + \eta_{21})^2 - 3(\eta_{12} + \eta_{30})^2] \\ \phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21}) \\ \phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{30} - 3\eta_{12})(\eta_{03} + \eta_{21})[(\eta_{30} + \eta_{21})^2 - 3(\eta_{30} + \eta_{12})^2] \end{aligned}$$

#### 2.2.2.4 Gray-Level Run Length

Analysis of direction of the pixel values from an object may help to distinguish it, for example, an object with values in one direction may suggest a current, which can be

blood vessel. Some features of this kind are based on the gray-level run length. A gray-level run is a set of consecutive pixels with the same gray-scale value in a determined direction ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ ,  $135^\circ$ ), and the size of it is denominated run length. Given a matrix with the run length data  $Q_{RL}$  from an image  $I$  with  $N_g$  gray-levels and maximum run length  $N_r$ , the number of times a run happened with gray-level  $i$  and length  $j$  is given by  $Q_{RL}(i, j)$  (GALLOWAY, 1975; THEODORIDIS; KOUTROUMBAS, 2008). From this matrix, we can extract the following features:

- Short-run emphasis:

$$SRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} (Q_{RL}(i, j)/j^2)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} Q_{RL}(i, j)} \quad (2.42)$$

- Long-run emphasis:

$$LRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} (Q_{RL}(i, j)j^2)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} Q_{RL}(i, j)} \quad (2.43)$$

- Gray-level non-uniformity:

$$GLNU = \frac{\sum_{i=1}^{N_g} \left[ \sum_{j=1}^{N_r} Q_{RL}(i, j) \right]^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} Q_{RL}(i, j)} \quad (2.44)$$

- Run length non-uniformity:

$$RLN = \frac{\sum_{j=1}^{N_r} \left[ \sum_{i=1}^{N_g} Q_{RL}(i, j) \right]^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} Q_{RL}(i, j)} \quad (2.45)$$

- Run percentage:

$$RP = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} Q_{RL}(i, j)}{L}, \quad (2.46)$$

where  $L$  can be treated as the total of number of pixels, as this value is the total possible number of runs in  $I$  with runs length equal to one.

Each feature extracted can be represented within a unspecific range of values. Some may have low ranges and others higher. Depending of how these features are represented, classifiers would be biased towards them. More balanced weights for the features can be attained by data normalization. This technique analysis each feature and normalize their values according to a specific rule, such as the normalization to zero mean and unit variance, called Z-score. One of the pros of using the Z-score is reduction of effects of outliers, samples with feature values farther from the mean which contribute to training

errors, in the data set (PRIDY; KELLER, 2005; THEODORIDIS; KOUTROUMBAS, 2008). Before calculation of the Z-score, computation of the mean and standard deviation is necessary. Given a data set with  $N$  samples and  $l$  features, a feature  $k$  can be normalized, its mean  $\bar{x}$  is calculated as:

$$\bar{x}_k = \frac{1}{N} \sum_{i=1}^N x_{ik}, \quad (2.47)$$

and its standard deviation as  $\sigma_k$  computed as:

$$\sigma_k = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_{ik} - \bar{x}_{ik})^2}, \quad (2.48)$$

then calculation of the Z-score is computed by the Equation 2.49.

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k} \quad (2.49)$$

### 2.2.3 Applications and libraries

This subsection describes briefly three main applications and libraries used in this work: Insight Toolkit, MongoDB and Shark.

The Insight Segmentation and Registration Toolkit (ITK) is an open-source toolkit mainly for segmentation and registration of medical images, developed with the C++ language initially by GE Corporate R&D, Kitware, Insightful, University of North California, University of Utah and University of Pennsylvania with funding of the American National Library of Medicine of the National Institutes of Health. ITK is currently at version 4.10. (JOHNSON et al., 2015).

MongoDB is an open-source document-based database. Data in MongoDB is stored in collections, with a format called BSON (or GridFS, for large data), with a JSON-like structure. MongoDB is currently at version 3.2 and it is available on many platforms, such as Linux and Microsoft Windows<sup>5</sup>.

Shark is a C++ open source machine learning library developed by Igel, Heidrich-Meisner & Glasmachers (2008), focusing in speed and flexibility. This library is mainly composed of supervised, unsupervised and evolutionary algorithms. Popular methods like SVM, LDA, RF, KNN, PCA and k-means are implemented, as many other novel methods. Currently, Shark is at version 3.1<sup>6</sup>.

<sup>5</sup> Information available at <https://www.mongodb.org/> and <https://docs.mongodb.org>. Last accessed in: 30 Jul. 2016

<sup>6</sup> More information available at <http://image.diku.dk/shark/>. Last accessed in 30 Jul 2016.

### 3 STATE OF THE ART

In the development of lung segmentation and nodule detection techniques, a study about previous developed techniques is necessary to understand some negative points and as improve results. A brief study about lung segmentation and nodule detection has been done for a better understanding of the problem. Some recent works about each one of these topics are described in separated sections below.

#### 3.1 LUNG SEGMENTATION

Before processing the CT scans for detection of nodules, a lung segmentation is realized to reduce the area to process. Most works about lung segmentation are developed with signal thresholding, deformable boundaries, shape models and edge techniques (ELBAZ et al., 2013). Bounding Box and Threshold were used to segment the lungs from CT scans by Liu et al. (2009). The Rolling Ball algorithm was applied to recover lost nodules candidates. For lung segmentation within juxta-pleural nodules, Ye et al. (2009) developed a segmentation approach that utilizes adaptive fuzzy thresholding, to create a initial segmentation, and chain code, to refine its contours and include nodules not initially segmented.

A new approach involving 3D region growing with wavefront simulation to segmentation of lungs' area was made by Nunzio et al. (2011). Analyzing the CT scan image histogram, a threshold is found and an application of a Simple-threshold 3D Region Growing (RG) is done for lungs, bronchi and trachea segmentation, followed by a wavefront simulation model to remove the external airways, resulting in the first mask. In cases the left and right lungs are connected, a separation surface is used in the first mask. In the next step, the simple-threshold 3D RG is applied in the left and right sides to create two masks (one for each side). For inclusion of pleural and internal nodules, a 3-dimensional morphological closing operator is applied.

For the preprocessing step, Ashwin et al. (2012) applied the Adaptive Median Filter and Contrast Limited Adaptive Histogram Equalization (CLAHE) techniques to enhance image contrast. Multilevel Thresholding was used for lung segmentation. Tariq, Akram & Javed (2013) used mathematical morphology techniques and euclidean distance to segment lung areas. Peak Signal-to-Noise Ratio (PSNR) calculation and median filter are used in the preprocessing step for the method created by Parveen & Kavitha (2013). For the lung segmentation, border extraction and flood fill algorithms are applied to the CT images.

An active contour model (ACM) and adaptive fuzzy thresholding based lung

segmentation algorithm was proposed by Keshani et al. (2013). First, an adaptive fuzzy thresholding is applied in the CT scan to obtain a binary representation. After, two windows with different sizes ( $5 \times 5$  and  $23 \times 23$ ) are applied to the scan to get lung areas assigned as non-lung ones (or to remove non-lung areas assigned as part of the lungs). This is done by verifying if the two opposite sides of the window have pixels with the same binary value, filling the windows with this value if true. Then, the same process is realized but with  $45^\circ$  rotated windows ( $25 \times 25$  and  $50 \times 50$ ), and the filling is realized only when three sides have the same value. At last, the resulted image is utilized as a mask for the ACM in combination with the Yezzi energy code.

In the method developed by Han et al. (2015), a simple thresholding is applied to segment the chest area from the rest and high level Vector Quantization (VQ) to segment the lungs. For juxta-pleural nodules segmentation, a 3D morphological closing operator is utilized. A K-Means clustering lung segmentation method, with trachea and bronchi removal using Euclidean distance, developed by Cid et al. (2015) took part in the VISCERAL Anatomy3 Challenge, obtaining 0.972 and 0.052 minimum Dice coefficient and maximum Hausdorff distance respectively, presenting satisfactory results if compared to other challenge's candidates.

### 3.2 NODULE DETECTION AND SEGMENTATION

After the segmentation of lung regions, the correct detection of nodules is desired. In this section, some recent works (2009-2015) about nodule detection are described in ascending order of date of publication. Some authors developed CAD systems to only detect nodules; others aimed correctly segment nodules after their detection; and works with only specific type(s) of nodules are presented below. Table 2 shows an overview of these works, and a discussion about them concludes this section.

In the CAD framework proposed by Ye et al. (2009), several steps were realized to detect and properly segment nodule regions, beginning by the application of the antigeometric diffusion, extraction of geometric features for nodule candidate selection, adaptive thresholding and a modified expectation-maximization (MEM) technique for segmentation, and finally 3D geometric features extraction and rule-based and weighted SVM classification. Ye et al. (2009) use antigeometric diffusion in the preprocessing step, to smooth edges, helping the calculation of the geometric features. Shape index feature is used to detect the potential nodules candidates because, as stated by the authors, GGO nodules do not have high sphericity but are more probable to be malignant than solid nodules. For these potential nodules, some method are applied on them depending if they are or not pleural nodules and based on their intensity mean. For pleural nodules, the chain code-based critical point is used. For non-pleural nodules, a adaptive threshold is applied and if its mean is greater than -500 HU, then a multiscale dot enhancement

filtering is applied in sub image containing the potential candidate. Some features (3D maximum distance based on distance transform, 3D object filtering based on motion tracking, sphericity, effective diameter and parameter selection) were extracted and used in a rule-based classifier to reduce the quantity of candidates which are non-nodule. Finally, a weighted SVM with RBF kernel classifies the remaining candidates, using the following features: intensity (maximum, minimum, mean and standard deviation), compactness, shape index mean, skewness, kurtosis, correlation, elongation, volume, location, sphericity, effective diameter and 3D maximum distance to the boundary. A total of 108 CT scans were utilized for validation of the method. Two data sets of 56 CT scans each were utilized for training and testing, separately. A detection rate of 90.2% was obtained in the testing data set classification, with 8.2 FP/scan.

An ensemble method using random forest (RF) and classification aided by clustering (CAC) was proposed by Lee, Kouzani & Hu (2010) for lung nodule detection. Experiments were realized with CT scans from the LIDC database. A total of 32 CT scans were utilized. The authors compared the RF with SVM and decision tree (DT), with and without CAC to validated the study. Nodules were extracted based on the annotation available in the LIDC database. A  $30 \times 30$  region size was utilized to fit nodule and non-nodule patterns. Nodule patterns with sizes greater than  $30 \times 30$  were resized to fit this size. One collection for nodule patterns, containing 1203 samples, and two non-nodule collections, one containing 1156 patterns from regions marked by radiologists as non-nodules and another with 1203 randomly regions (with sizes  $30 \times 30$ ,  $56 \times 56$  and  $82 \times 82$ , being resized to  $30 \times 30$ ), were generated. Overall, results with RF were better than with SVM or DT. Also, the RF CAC EM approach was slightly better than the non-CAC RF, but it had a higher execution times. The best RF CAC EM result achieved a sensitivity of 98.33%, specificity of 97.11% (80% of samples for training and 20% for testing, with number of trees grown equal to 100 and number of variables at each split to 25) and execution time with an average of 210.81 seconds. As for the non-CAC RF's best result, the sensibility, specificity and execution time were 95%, 96.28% and 45.42 seconds, respectively (same proportion of samples and parameters as RF CAC EM case). Experiments were realized on a Dell Precision 490 Workstation with Intel Xeon CPU 5130 @ 2 GHz.

An artificial neural network approach based on the BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm was presented by Ashwin et al. (2012) for nodule detection. Preprocessing with median filter and Contrast Limited Adaptive Histogram Equalization (CLAHE) was performed to reduce noise and improve image contrast. A multilevel thresholding approach for detection of nodule candidates and a two-layer neural network for candidate classification were employed. The BFGS quasi-Newton back propagation algorithm played an important role as the training input, generating an approximate Hessian matrix at each iteration. This approach resulted in a sensitivity of 92% for the 40

CT scans tested, with a 0.2 FP rate per scan (specificity of 94.3%).

A lung nodule detection based on SVM classification of 2D and 3D features with active contour-based nodule extraction was developed by Keshani et al. (2013). The segmented lung was divided in  $3 \times 3$  windows and some features were extracted from them. Extracted features were mean, variance, 3D averaging, 2D and 3D cross correlation. A SVM classifier with gaussian RBF kernel was utilized for classification of these windows into nodule and non-nodule, and verifying if the nodule was attached to the lung wall, to the bronchioles or solitary. Since the obtained nodules may have deformable contours, the ACM was applied to get a better shape of these nodules. Experiments were realized with four groups of data: a group with 13 nodules with high slice thickness (45 slices/scan); a second group with six nodules and slice thickness of 0.625 mm; a third group using the ANODE09 data set with 39 nodules; and the fourth group using CT scans from the ELCAP data set, containing 397 nodules from 50 patients (early scans from the LIDC database). To training 15 CT scans from each group were utilized. Reported results with 19 nodules from the first two groups showed a detection rate of 90% and 5.63 FP/scan, with a Dice coefficient of 82%. A detection rate of 66.2% and 8 FP/scan was achieved with the ANODE09 data set using the value for parameter  $c = 3$ , from the 3D feature descriptors, which influences the quantity of analyzed slices (the default value is  $c = 5$ ). At last, for the ELCAP data set, a detection rate of 89% and 7.3 FP/scan was obtained.

A hybrid classifier called neuro fuzzy, based on neural networks and fuzzy logic, was utilized by Tariq, Akram & Javed (2013) to classify nodule candidates. For selection of these candidates, some steps were realized on the image. From an optimal thresholded image, the morphological operations of opening and closing were applied, for reduction of noise and improvement of borders, respectively. Then, a Sobel in horizontal and vertical was performed to enhance boundaries. The two objects with largest boundaries were considered as pulmonary lobes, and the boundaries from other objects were reconstructed and a filling step was realized to get the area from each object, serving as samples for the neuro fuzzy. In the classification step, the following features were extracted: area, energy, entropy, eccentricity, mean and standard deviation. Experiments achieved an accuracy of 95%, but they were done with only 100 slices of different patients, thus compromising its reliability. Sensitivity and specificity were not reported.

A intra-type lung nodule classification approach with patch division, superpixel labeling and a new feature description (context curve) for nodule characterization, was proposed by Zhang et al. (2013). They developed an adaptive patch division approach with quick shift clustering method for superpixels generation, applying on quad-amplified images (then return them to their original size). The Otsu thresholding method was used to label the superpixels into foreground or background. For the extraction of the context curve feature, circular sections involving the superpixel and its around are done, a foreground

ratio is extracted for each section and used to generate the feature vector. Finally, the four nodule types (juxta-pleural, pleural-tail, vascularized and well-circumscribed) are classified by a SVM classifier with polynomial kernel. Experiments were realized with the ELCAP database, containing 50 CT scans with 379 nodules. The four types compose the database with 30% juxta-pleural, 39% pleural-tail, 16% vascularized and 15% well-circumscribed. The percentages of CT scans for training for each type were chosen randomly between 10% and 90%, with the remaining scans compounding the testing set. The average classification rate was around 90%, achieving the best result using 70% of training percentage.

Aiming to standardize the evaluation of CADe systems, some rules and a new method were proposed by Brown et al. (2014). Thresholds of 4 and 8 mm were used to set the minimum size of a nodule to be evaluated. A minimum of four times the slice thickness was also determined. Their method is composed of intensity thresholding, Euclidean Distance Transformation (EDT) and watershed segmentation for candidate selection. Then, a 3D connected component analysis is performed in each candidate to cover neighbor voxels around it that satisfies a local EDT maxima. To exclude non-nodules, volume and shape rules are applied. The first 200 scans from the LIDC data set were utilized for the evaluation of their method (cases 0001-0080 for development and 0081-0200 for testing stage). Since the LIDC data set presents annotations of four radiologists, a majority rule was applied, where only nodules marked by at least three radiologists with size greater or equal than the threshold were considered. From the testing stage, 68 and 58 nodules were evaluated (for thresholds of 4 and 8mm, respectively). Sensitivity of 79.8% and 2.05 FP/scan were achieved with size  $\geq 4$ mm. For size  $\geq 8$ mm, 82.2% sensibility and 1.01 FP/scan.

Another work based on shape feature descriptor is the CADe system developed by Choi & Choi (2014). Using the eigenvalues of a generated Hessian matrix, nodule candidates are detected based on the values of the multi-scale dot enhancement filter. First, a Gaussian filtering is used for noise reduction and generation of the Hessian matrix. Since the objects of interest (nodules) have different sizes, a five smoothing scales filtering approach is utilized, where the values of  $\sigma$  are calculated according to the nodule size range of the data set (this work also utilizes the LIDC database). For the nodule candidate detection, the dot-enhanced image for each scale is thresholded, based on the local maximum dot values average and the regions are extracted. From these nodule candidates, a novel feature descriptor based on shape description of small 3D objects, Angular Histogram of Surface Normals (AHSN), is proposed. This feature descriptor can be calculated with the eigenvalues extracted from the Hessian matrix. Before the calculation of AHSN, angular histograms of surfaceness (representation of the angular direction of the surface normal vector on the surface saliency), are extracted. Elevation  $\theta$  and azimuth  $\varphi$  are calculated and represented in a range of  $[0, 180]$  and  $[0, 360]$  degrees, respectively, then two angular histograms are generated based on these values.



Since some nodules may be attached to other objects (pleura, vessels, etc), they would interfere in the AHSN calculation, so a wall elimination approach was developed to remove these objects and improve nodule detection. This approach consists in an analysis of peaks in the AHSN. These peaks can be obtained with local maxima. Then, a connected component technique is applied to voxels with similar normal vector orientations. Large reconstructed areas are removed. Finally, the AHSN feature is recalculated for the nodules candidates. In this step the number of non-nodules is far greater than of nodules. To correctly select the nodules, a SVM classifier with three different kernels (polynomial, RBF and Minkowski) is applied with a  $k$ -fold cross-validation ( $k = 10$ ). Data set consists of 148 nodules from 84 chest CT scans, available from the LIDC database. A candidate is labeled as nodule if the distance between it and the nearest nodules is smaller than 1.5 times the nodule's radius, or if the candidate's radius is between 0.8 to 1.5 times the nodule's radius. Candidates not following this rule are labeled as non-nodules. Best results were achieved with the RBF kernel, with a sensitivity of 97.5% and 6.76 FP/scan (CHOI; CHOI, 2014).

The sub-solid nodule have higher malignancy rate and detect this nodule kind Jacobs et al. (2014) developed a method involving extraction of context features. For an initial candidate selection, an interval thresholding with values of -750 and -300 HU was applied to the CT scan, followed by a 3D erosion operation with spherical structuring element ( $r = 1$ ). Connected component analysis was applied to the image to cluster the remaining objects. Clustered regions with volume smaller than  $34 \text{ mm}^3$  were removed, as they present a diameter inferior to 5 mm (if treated as a spherical volume), which according the authors, would not be further analyzed. For a proper segmentation, they utilized the nodule segmentation approach described by Kuhnigk et al. (2006), which treat nodules attached to other objects (but to proper adapt to the sub-solid problem, they changed a global lower threshold value utilized from -450 to -750 HU). Four groups of features were extracted from these nodule candidates to describe them: intensity, texture, shape and context features. These features were extracted not only from the segmented nodule candidate, but from the surrounding region. Particularly, the context features represented characteristics of the nodule candidate with relation to the lung segmentation and airway tree, additionally the relationship between other candidates from the same CT scan. Some features of the other three groups were LBP, Hu moments, maximum vesselness, entropy, mean, 2D Haar wavelets, sphericity and compactness. Classification of these candidates was executed with two different approaches, an one-stage and a two-stage classification. In the first stage of the two-stage classification, a Linear Discriminant Classifier (LDC) was executed with five selected features with Sequential Forward Floating Selection (SFFS) or Fisher's linear discriminant ratio, for the purpose to speed up and simplify the process. In the second stage (same for the one-stage classification), various classifiers were separately tested, attempting to found out the best classifier for this sub-solid nodule detection approach. The classifiers tested were kNN, RF, GentleBoost (GB), Nearest Mean (NM)

and SVM with RBF kernel. Two independent data sets from the NELSON trial were employed in the development and testing of the method. In the data set for training and optimization of the technique, 122 sub-solid nodules  $\geq 5$  mm from 103 patients could be found. In the second data set, 60 sub-solid nodules  $\geq 5$  mm from 56 patients were applied in the testing stage for validation. A 10-fold cross validation was applied in the training step. A sensitivity of 88% was obtained in the candidate detection step. Best results were achieved by the two-stage GB classifier, with a sensitivity of 80% and 1.0 FP/scan. Finally, a hybrid approach combining a solid nodule detection developed by Murphy et al. (2009) and this sub-solid nodule detection improved the sensitivity from 80% to 88%.

A solitary nodule detection based on best model selection with genetic algorithm was developed by Filho et al. (2014). Foremost, a quadratic enhancement was applied to the image for improvement of its contrast. A side effect of this approach was noise growth, treated by usage of the Gaussian and median filters. In the next step, detection and segmentation of nodule candidates was realized using the quality threshold (QT) algorithm, followed by the application of a region growing. Shape and texture features were extracted to represent the extracted nodule candidates. The shape features calculated were spherical disproportion, spherical density, sphericity, weighted radial distance, elongation and Boyce-Clark radial shape index; in the texture group, contrast, energy, entropy, homogeneity and momentum from the co-occurrence matrix and mean, standard deviation, obliquity, kurtosis, energy and entropy from the histogram were extracted. Also, features based on diversity analysis, the Simpson's and Shannon's indexes, were calculated. Stepwise discriminant analysis was applied to select the best discriminant features. For the classification part, the micro-genetic algorithm (MGA) was executed to generate a training set for improvement of the testing results, which would serve as input to the SVM classifier. To test 800 exams from the LIDC database, 640 for training and 160 were utilized. From the 27 extracted features, 17 were selected with the stepwise discriminant analysis. Then, the genetic algorithm was applied in the training set for best representation of the data, reducing the number of nodules and non-nodules from 458 and 47067 to 370 and 1110, respectively, greatly reducing the quantity of non-nodules, thus balancing the amount of samples for each class. In the testing stage, 149 of 182 nodules were classified, as the remaining 33 were lost in previous steps. Experiments demonstrated that usage of both diversity features and MGA achieved best results, with a sensitivity of 85.91% and 1.82 FPs/exam (140 exams with nodules in the classification stage).

A CADe system based on nodule and outer surface's features was developed by Demir & Çamurcu (2015). Before feature extraction and classification, multiple thresholding and a 2D connected component analysis (CCA) followed by a 3D CCA was executed, obtaining candidates to be classified. In the feature extraction phase, four groups of features were created: Morphological, Statistical and Histogram, Outer Surface Statistical and Histogram, Outer Surface Texture features. First group's features consists in volume,

Minimum Axis Length (MinA), Maximum Axis Length (MaxA), division between MinA and MaxA ( $MinA/MaxA$ ), equivalent radius, sphericity and compactness. The statistical and histogram groups consists in mean, maximum pixel value, minimum pixel value, most frequent pixel value, variance, standard deviation, skewness and kurtosis features. As for the Outer Surface Texture Surface group, the features present were: contrast, energy, entropy, homogeneity and moment. Haralick's Gray Level Co-Occurrence Matrix (GLCM) was utilized to generate the Outer Surface Texture features. After the feature extraction step, the support vector machine (SVM) was used to classify nodule candidates. An improvement was applied in the SVM classifier using an algorithm called Particle Swarm Optimization (PSO). With all groups combined, the method achieved a sensitivity of 93.6% and 2.45 FP/scan in the 200 CT scans (different patients) from the LIDC/IDRI data set, with a total of 609 nodules. Outer surface features may help improve nodule detection and removal of false positives.

Orozco et al. (2015) developed a CAD based on the Wavelet Transform. ELCAP data set was used to run the experiments. Nodules were classified by the SVM learning model with features from the Gray-Level Co-occurrence Matrix (GLCM) in the wavelet domain, grouped into the malignant and benign classes. A circular ROI involving most of the thorax area was extracted. Discrete Wavelet Transform (DWT) was applied to this ROI to get the Daubechies wavelets db1, db2 and db4, and their sub-bands LL<sub>1</sub>, LH<sub>1</sub>, HL<sub>1</sub> and HH<sub>1</sub> (L from low and H from high). The DWT were applied recursively to the LL sub-bands. So, the LL<sub>2</sub>, LH<sub>2</sub>, HL<sub>2</sub> and HH<sub>2</sub> from the LL<sub>1</sub> were also utilized. From each Daubechies filter, 19 second order statistical features were extracted of the sub-bands (total of 7, including the LL sub-bands) in four angles (0°, 45°, 90° and 135°). An attribute evaluated was used to reduce the dimensionality of the feature vector, decreasing it from 19 to 11 features and, to reduce even more, a sub-band selection was done. Combinations of two from 11 features were trained using a SVM with a RBF kernel. A total of 45 CT scans (23 with and 22 without lung nodules) were tested where the sub-band LH and the pair of features X and Y proved to be the best choices, with a sensitivity of 90.90% and a specificity of 73.91%.

According to Sui, Wei & Zhao (2015), although SVM is commonly used for classification because of its good results, the classifier has problems regarding unbalanced data. To address them, they proposed a SVM with the random undersampling (RU) algorithm and the Synthetic Minority Oversample Technique (SMOTE). The features used in this work were grouped in 2D and 3D. The 2D features were composed of circularity, elongation, compactness and moment. 3D features were surface-area, volume, sphericity and centroid-offset. Experiments were done with 75 nodules and 454 non-nodules from the LIDC database and a data set provided by the ShengJing Hospital. The RU-SMOTE-SVM classifier achieved an accuracy of 93.76% and a sensitivity of 77.33%. 16 non-nodules were classified as nodules (FP).

Table 2 shows an overview of the works studied.

### 3.2.1 Discussion

Various methods for nodules detection were described previously. An overview about the studied articles is shown below, detailing some general characteristics of them, so their positive and negative points. These CADe systems had mainly the objective to detect any type of lung nodules, but some specific nodule detection systems were described too. Segmentation of detected nodules is presented in some works.

In the preprocessing step, smoothing and noise reduction filters were applied, such as anti-geometric diffusion and Gaussian filtering. For nodule candidate selection, thresholding, region growing, shape and other techniques were employed. Some techniques required a preprocessing step to achieved better results. The majority of these works decreased number of false positives through classification with previous extracted features. These features are mainly grouped into texture, intensity or shape-based categories.

Some patterns can be found in the studied methods. First, number of CT scans tested are distinct between them, and even if the same data set was used, generally it was not shown the specific CT scans used for development and testing of the method. Though some of them utilized techniques for nodule segmentation, a detailed analysis of the results achieved by them are not described, probably due to the complexity of such thing. Proportion of nodule and non-nodule candidates is unbalanced in a few works, thus methods to reduce or adapt to this irregularity are encouraged. Finally, though usage of sensitivity and FP/scan evaluations were employed by most works, others used different evaluations (accuracy, specificity by %, etc), hindering evaluation of their results.

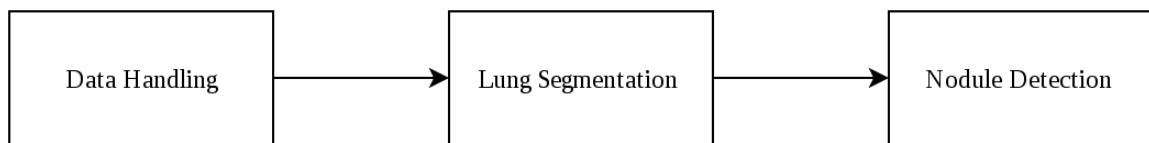
In the next chapter, our methodology for development of the automatic nodule detection is shown. Through the chapter, some choices are made based on these patterns.

**TABLE 2** – An overview of recent works about lung nodule detection

Authors	Data set	Samples	Candidate detection	FP reduction	Results
Ye et al. (2009)	Inhouse	54/54 CT scans for training and testing	Shape index and multi-scale z-dot filtering	Rule-based and weighted SVM classification	90.2% TPR & 8.2 FP/scan
Lee, Kouzani & Hu (2010)	LIDC	32 CT scans (80%/20%)	Radiologists' annotations	RF-CAC	98.33% TPR & 97.11% specificity
Ashwin et al. (2012)	LIDC	40 CT scans for testing	Multilevel thresholding	BFGS-based ANN	92% TPR & 0.2 FP/scan
Keshani et al. (2013)	Inhouse, ANODE09 & ELCAP	19, 39 & 397 nodules	Active contour	SVM classification	89% TPR & 7.3 FP/scan (ELCAP)
Tariq, Akram & Javed (2013)	Not specified	100 slices	Morphological and boundary operators	Neuro fuzzy classifier	95% accuracy
Zhang et al. (2013)	ELCAP	50 CT scans	Adaptive patch division	Polynomial SVM	Average of 90% TPR
Brown et al. (2014)	LIDC	200 CT scans (120 for testing)	Thresholding, EDT and watershed	3D CCA with rule-based classification	79.2% TPR & 2.05 FP/scan
Choi & Choi (2014)	LIDC	84 CT scans	Multi-scale dot enhancement	SVM with AHSN features	97.5% TPR & 6.76 FP/scan
Filho et al. (2014)	LIDC	800 CT scans (160 for testing)	Quality threshold and region growing	Genetic algorithm	85.91% TPR & 1.82 FP/scan (total of 140 scans)
Jacobs et al. (2014)	NELSON	122 and 60 nodules	Interval thresholding and connected component analysis	One and two-stage classifications	80% TPR & 1 FP/scan
Demir & Çamurcu (2015)	LIDC	200 CT scans	Multiple thresholding, 2D and 3D CCA and morphology	SVM with PSO	93.6% TPR & 2.45 FP/scan
Orozco et al. (2015)	ELCAP & LIDC	61 (training) & 45 (testing) CT scans	Hough transform and Wavelet Transform	SVM classification with pair of features	90.90% TPR & 73.91% specificity
Sui, Wei & Zhao (2015)	ShengJing & LIDC	75 nodules & 454 non-nodules	Radiologists' annotations	RU-SMOTE-SVM	77.33% TPR & accuracy of 93.76%

## 4 METHODOLOGY

Different ways of automatic nodule detection were present in the previous chapter, as some discussions about their characteristics. Aiming to improve and develop other ways for this task, a superpixel nodule candidate selection approach using the shape index and curvedness features is proposed. Our methodology is covered in this chapter, beginning with some processing steps in the utilized database, following by two approaches for lung segmentation, the proposed nodule detection refined by feature extraction and classification. Figure 15 presents the overall flow of our methodology.



**FIGURE 15** – Basic flow of the methodology.

### 4.1 DATA HANDLING

Analysis of the problem and developed of methods to resolve this problem demand some data. In our case, a database with chest CT scans containing nodules is necessary for development and validation of the proposed method. For this, the LIDC/IDRI database adopted. This database has annotation by four radiologists for CT scans from 1010 patients. Information about nodules' size, likelihood of malignancy and other characteristics has been provided. Although this database has many interest data to be processed, there are some particularities that are in interest to be addressed:

- There is no data about the relation between radiologists' reads from determined nodule in the XML files, but they are present in the LIDC Nodule Size Report.
- Diagnostic data is stored in a different place.
- Only coordinates from the nodules' borders were stored.

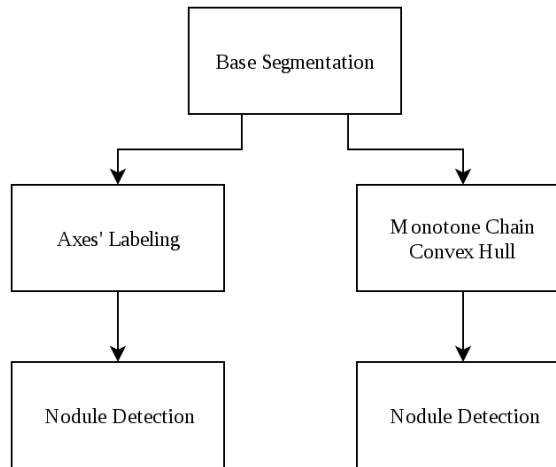
For better usage of this database, we decided to store the needed data into a Mongo DB database. An optimal choice was not realized as it is not the objective of this work. Basically, the below items were done:

- Data was organized into different collections (patient, series, nodule, diagnosis, etc).
- Every coordinate from inside the annotated nodule was stored into the diagnosis collection, minus when this coordinate is from inside an annotated non-nodule region.

- When a diagnostic data was provided, it was stored into the nodule collection.
- Resulted data from subsequent phases (lung segmentation, nodule detection, segmentation and diagnosis) were stored into various collections.

## 4.2 LUNG SEGMENTATION

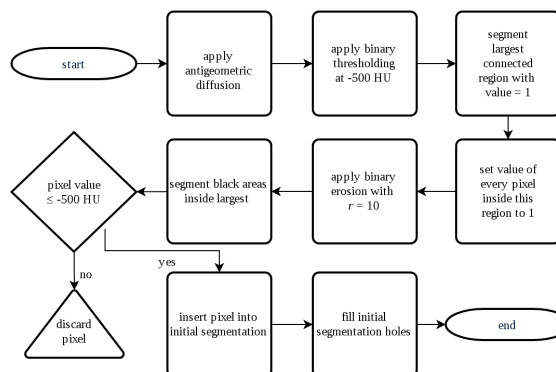
Besides pulmonary regions, chest CT scans have parts of others regions, blank areas (around the thorax) and some noises from the scanner. Removal of these areas from the image are encouraged to reduce processing time in following steps. We developed two approaches for segmentation of pulmonary areas from CT scans: the Axes' Labeling (AL) and the Monotone Chain Convex Hull (MCCH). Figure 16) shown the flow of this stage.



**FIGURE 16** – Representation of the lung segmentation stage.

### 4.2.1 Base segmentation

Before the application of these approaches, some steps are realized to reduce noise and isolate the thorax and lung regions, referenced as base segmentation. Figure 17 displays the basic flow of this step.

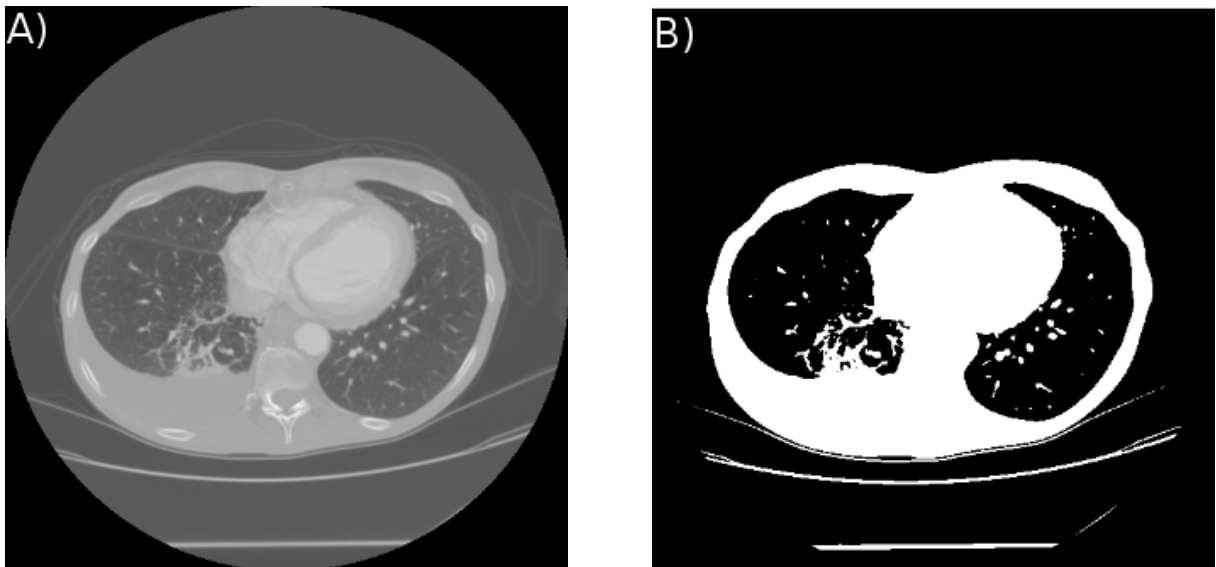


**FIGURE 17** – Overview of the base segmentation flow.

Foremost, an anti-geometric diffusion filtering (see Section 2.2.1.5) is applied to the CT scan, as it reduces image noising and improve the shape index values for nodules (YE et al., 2009). After, a thresholding using the value of -500 HU is applied to the CT scan, with the objective to obtain an initial separation of the lungs from the thorax. This specific value was used in other works for an initial lung segmentation, since it would separate most part of the lungs (as seen in Table 1 from Section 2.1.1, lungs have CT values around -700 HU) (PU et al., 2008; MESSAY; HARDIE; ROGERS, 2010). Given a CT image  $f(x, y, z)$ , the resulted image  $g(x, y, z)$  from thresholding is given by the following equation:

$$g(x, y, z) = \begin{cases} 1 & \text{if } f(x, y, z) > -500 \\ 0 & \text{if } f(x, y, z) \leq -500 \end{cases} \quad (4.1)$$

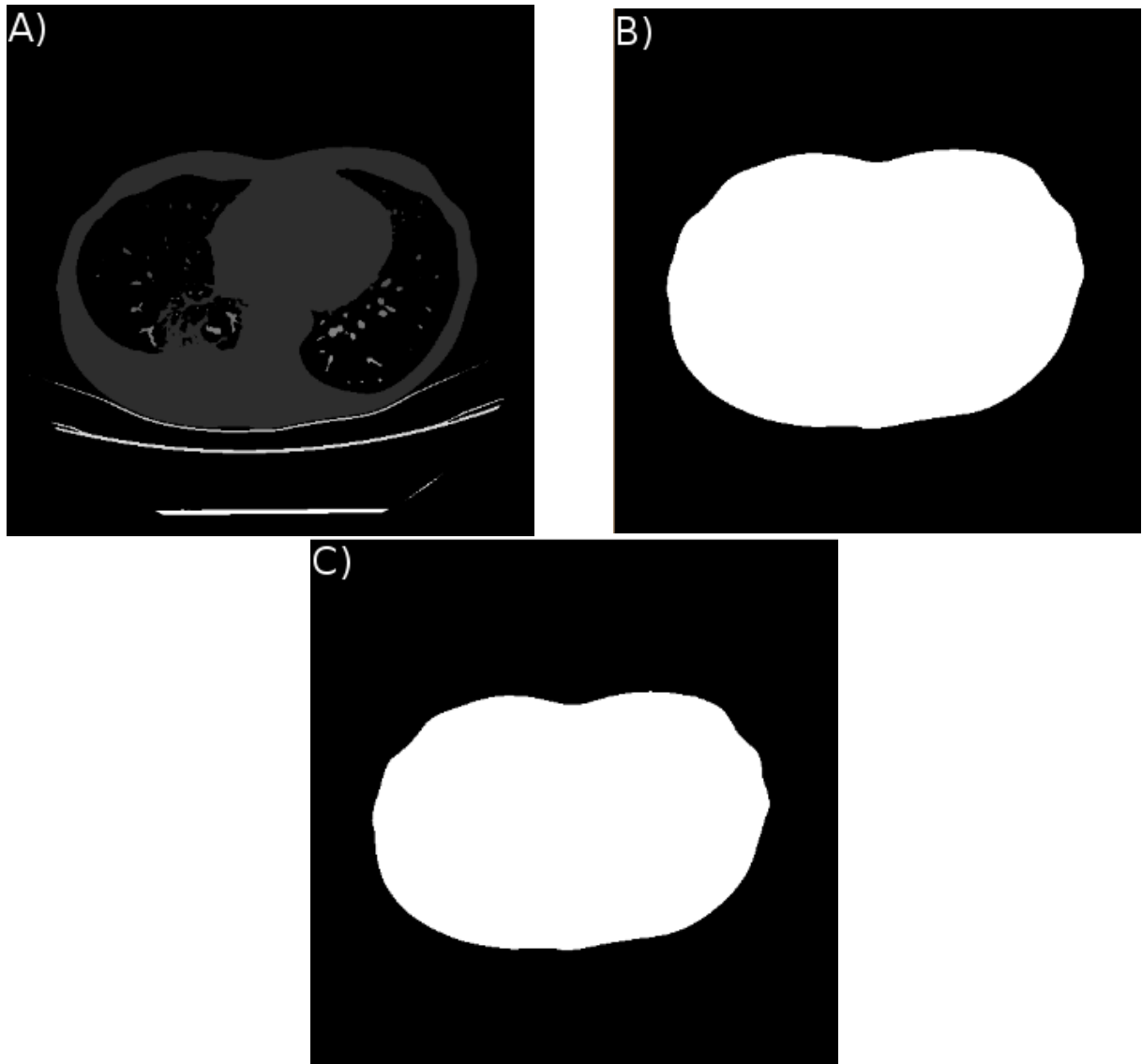
An example of thresholding using this value is shown in Figure 18. Although the majority of the lungs had CT values less or equal than -500 HU, some other regions had CT values higher than this threshold, which would be included in the thorax region. Inclusion of these regions will be address in latter steps.



**FIGURE 18** – In A) the original image and in B) Thresholded image. The method use a constant value -500 in Hounsfield Unit (HU) on threshold.

After thresholding the image, we segmented the largest connected region with value equal to 1, the thorax region. To do this, a labeling is applied in any pixel with value equal to 1. Then, every hole (lung region candidate) from the thorax region is filled with a hole filling technique. After the hole filling, we apply an erosion with a cross-shaped structuring element with a radius  $r = 10$  to remove some noises around the thorax region, which may create unintentional holes. This radius was found empirically and removed noise without compromising the lungs.





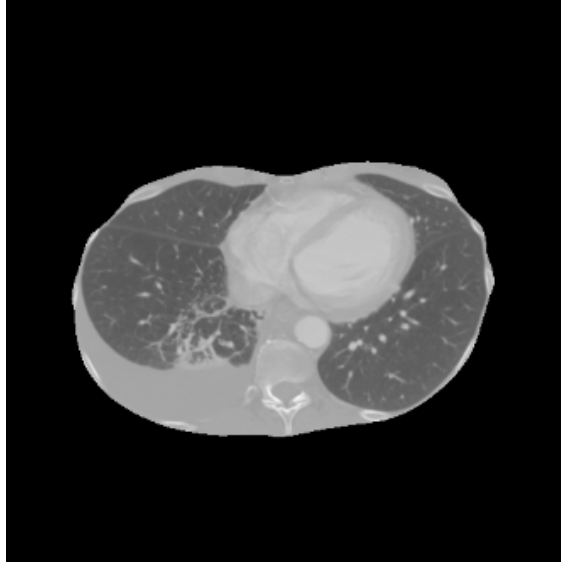
**FIGURE 19** – In A), the labeled image. In B), largest thorax region on image and in C) the image B) after applied an erosion morphological operator.

The next step is an initial segmentation of the lung region. First, we segment pixels with HU values  $\leq -500$  from the original image  $f(x, y, z)$  which are inside the thorax area (largest white area from  $g(x, y, z)$ ) and, after that, we apply a hole filling to get almost every part of the lung region (Figure 20).

Some parts of the lung region may not be segmented (Figure 21 ) as they are white regions outside the darker segmented area (juxtapleural nodules, for example). To solve this, two approaches, one based on the convex hull technique called monotone chain and another based on labeling regions in different axes are applied.

#### 4.2.2 Axes' Labeling

In this approach, we want, from the initial segmented lung region, the largest connected region from the entire trans-axial volume's left and right sides, as they are



**FIGURE 20** – The thoracic area with the lung inside.



**FIGURE 21** – The original thresholded image show in Figure 19 a) after completely segmentation process. Parts of lung were lost.

intended to be the left and right lungs. First, we divide the volume in two parts, left and right. Then, for each separately region, we get the centroid from largest connected white volume and apply a connected component on the base segmentation using these two centroids, removing regions not connected to them.

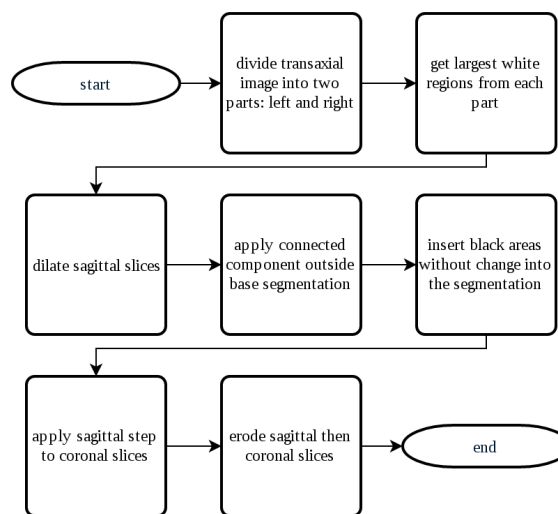
We applied a connected component in the two axes orderly: sagittal and coronal. Below, we describe the next steps:

1. For each sagittal slice  $s(i, j)$ , where  $i \rightarrow y$  and  $j \rightarrow z$ , a dilation operator with cross-shaped SE and radius  $r = 3$  is applied to close almost closed regions (holes).
2. Apply a connected component in the region around the lungs, where non-filled

regions would be from lung (similar to a hole filling). The filled region from this sagittal volume is treated as background and other regions as foreground.

3. After, transform this sagittal volume into a coronal volume. Execute the same process, dilating coronal slices  $s(i, j)$ , where  $i \rightarrow x$  and  $j \rightarrow z$ , and filling the region around the lungs.
4. Apply an erosion, with same SE than the dilation, to the sagittal and then to the coronal plane, to return the lungs to the (approximate) original volume.

This approach is summarized in the Figure 22.



**FIGURE 22** – Representation of the AL approach

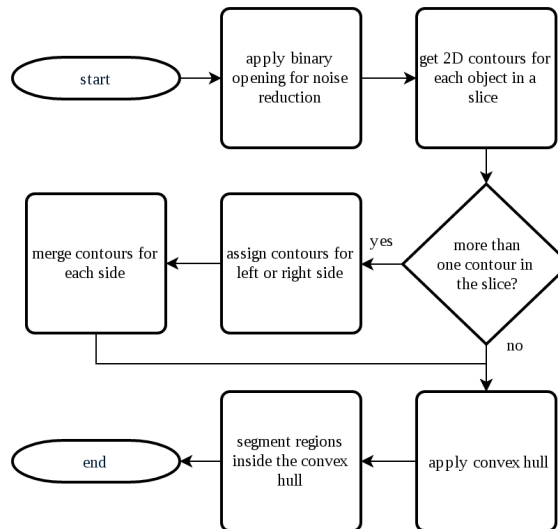
#### 4.2.3 Monotone Chain Convex Hull based approach

To reduce processing time and remove noise, an opening operator with cross-shaped structuring element and radius equal to one is applied to the CT scan. For each slice  $(s(x, y))$ , we get the contours from every object and attribute them to one from two sides: left and right lung. Choosing of which side a determined contour will be assigned is made by how near they are from the points already assigned to the sides. For each point  $p$  from some contour, the spatial Euclidean distance is calculated from every point of the two sides and the lowest distance is stored. After getting all distances, a mean is calculated to both sides and the contour is assigned to the nearest side. For this part, the following steps are executed:

1. In the first step, when no point is assigned to both sides, we choose the two nearest contours from the first and last columns (one for the left and another for the right side).

2. If only one contour is found, the method proceeds to the convex hull execution. If two contours (one for each side) is found, they serve as input to the convex hull method separately. If we have more than two contours, the process of assignment is started.
3. Since assignment of a contour may alter following assignments, two parallel iterations, one beginning in the first column and other in the last column, are executed. For example, the system checks if exists a point from a new contour in the current column from the first iteration. If true, the assignment is realized. If not, this iteration is incremented to the next column ( $y \rightarrow y + 1$ ) and the process changes to the other iteration. In this second iteration, the system realizes the same thing from the first iteration but, if no point is found, the iteration is decremented to the previous column ( $y \rightarrow y - 1$ ). This is done until every column is checked (no matter if it was from the first or second iteration).

Every point inside the convex areas is segmented, thus finalizing this lung segmentation approach. This approach is illustrated in Figure 23.

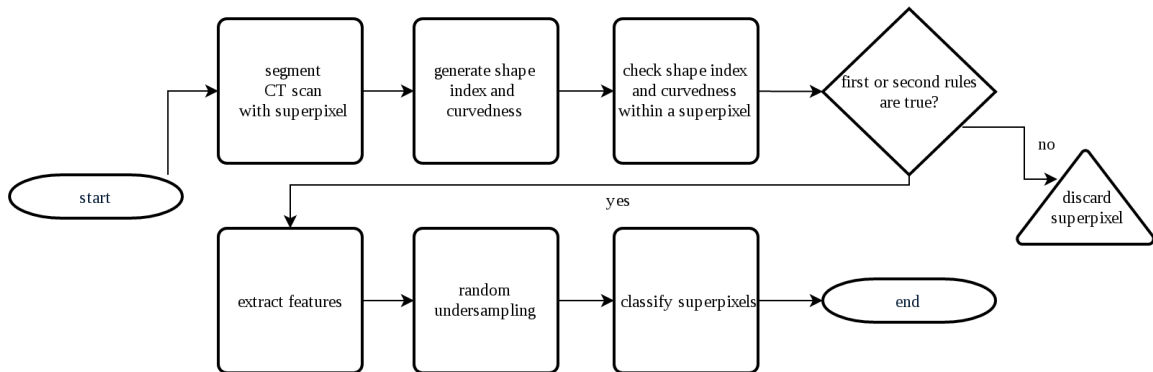


**FIGURE 23** – Illustration of the Monotone Chain Convex Hull approach

### 4.3 NODULE DETECTION

Nodule detection is a crucial step for diagnosis of cancer presence in a patient, and automatic detection may help radiologists through reduction of workload and second opinion in uncertain objects. Using the LIDC database, only reported nodules with the majority of radiologists (3 or 4) labeling them as  $\geq 3mm$  are considered nodules, in other cases there is not a consensus if it is really a nodule. For automatic nodule detection, we developed a method based on superpixel segmentation with different techniques, nodule candidate selection using shape index and false positive reduction with SVM and RF

classifiers. In Figure 24, we present our nodule detection method and its flow. For each step, a subsection describing it is present.



**FIGURE 24** – Flow of the nodule detection stage

#### 4.3.1 Superpixel segmentation

Usage of superpixel segmentation is proposed to group the lung regions in subregions according to their intensities and localization. For comparison, the SLIC (2D and 3D) and FH (only 2D) segmentations are utilized, as they are greatly used in superpixel segmentation, achieving good results in the literature.

Before execution of the superpixel and supervoxel methods, an image conversion from 16-bit to 8-bit is realized, as weight values  $m$  similar to the values from the SLIC article resulted in strange results (authors probably developed with 8 bit images in mind) and the FH technique had a similar problem. Since only a range of values are necessary, only HU values between -999 and 1000 are used in the 8-bit conversion as the intensities of objects of interest are in this interval (other values are discarded).

SLIC approach can generate 2D (superpixels) and 3D (supervoxels) regions. As nodules may be contained in different but neighbor slices, a supervoxel generation approach may result in a more faithful segmentation.

Since we are working with CT scans with different pixel spacing and slice thickness, it has been decided to modify the step value  $S$  according to these values. Instead of using a step value for pixels, it is chosen with millimeters in mind. As we are working with nodules with interval of [3mm, 30mm] of diameter, we decided to use different values of  $S$  based on this interval, but as mm. Moreover, slice thickness probably will be different than pixel spacing. So, a different step value is used when working with  $z$  dimension (only applied to the SLIC supervoxel generation). Besides, as these values may result in non-integer numbers, a ceiling is used.

Given the initial step value  $S$ , pixels spacing  $p_x$  and  $p_y$ , and slice thickness  $t$ , the step values for each dimension ( $S_x$ ,  $S_y$  and  $S_z$ ) can be calculated with Equations 4.2, 4.3

and 4.4:

$$S_x = \lceil p_x * S \rceil, \quad (4.2)$$

$$S_y = \lceil p_y * S \rceil, \quad (4.3)$$

$$S_z = \lceil t * S \rceil. \quad (4.4)$$

Superpixels with only zero intensities in the 8-bits image are discarded. As the step value is small if compared to the image size, a great number of superpixels are generated. A step value with a big value can help this, but segmentation of small nodules may contain too many non-nodule regions.

Generation of superpixels in the FH approach (described as components in its article) is only realized in 2D images so, for each separately CT slice, this technique is applied. Gaussian filtering is not applied to the image, as the resulted image from the anti-geometric diffusion is the input image. Different values of  $K$  are chosen for the experiments. Depending on the parameters of the method, too small superpixels may be generated in some areas. A minimum size based on the minimum physical area (with diameter of 3mm) is used. Superpixels smaller than the minimum size are merged into the nearest superpixel.

Superpixels in both approaches are generated based on their intensities, which would help in the segmentation of nodule areas from the lung region (brighter and darker areas, respectively). Two mainly problems can be detected in this approach: a great number of superpixels will be generated; and even if a thresholding is applied to separate nodule from non-nodule superpixels, some may contain vessels (brighter areas), which will not be removed. To address these problems, a technique to reduce the number of superpixels for further analysis (nodule candidates) and vessel or other brighter objects is presented in the next subsection.

#### 4.3.2 Nodule candidate selection

In many nodule detection methods, a candidate selection step is executed to select the candidates most likely to be nodules. For this step, the shape index and curvedness values are generated for each voxel from a CT scan. These values represent how circularity a specific region is. Some voxels from nodules generally present values of shape index and curvedness in similar intervals from other CT scans. For each superpixel, an analysis of these values is realized in windows, to verify if this window is circular enough to be part of a nodule.

First thing to do is the generation of shape index and curvedness features using the Hessian matrix. For generation of the Hessian matrix, the `itk::HessianRecursiveGaussianImageFilter`<sup>1</sup> class is utilized. A  $\sigma$  parameter from the Recursive Gaussian Filter (used to generate the partial derivatives) is necessary. Since we are working with CT scans with different pixel spacings, objects from different scans with same physical volume may have different pixels' volume, so smoothing and calculation of the derivatives independent from the size. For this, the "NormalizeAcrossScale" flag is set ON in the filter. As this normalization is enabled, a balanced candidate selection was observed if compared with a selection without normalization.

Depending on the  $\sigma$  value, smaller or larger objects (nodules) may obtain higher shape index, so it is worth to generate the Hessian at different  $\sigma$  values to cover many sizes of nodules. Some approaches used multi-scale according to the objects of interest, when working with the Hessian matrix generation (in these cases, for the dot enhancement filtering) (YE et al., 2009; CHOI; CHOI, 2014). These  $N$  scales are calculated, with a range  $[d_0, d_1]$  and  $r = (d_1/d_0)^{1/(N-1)}$ , as follows:

$$\sigma_1 = \frac{d_0}{4}, \sigma_2 = r\sigma_1, \dots, \sigma_N = r^{N-1}\sigma_1 = \frac{d_1}{4} \quad (4.5)$$

The Hessian matrix is calculated for each scale, and their eigenvalues are obtained to calculate the shape index and curvedness. Then the maximum shape index and curvedness for each voxel between the  $N$  scales are stored, being utilized in the remaining processes.

Foremost, for each superpixel, a window of  $3 \times 3$  (a  $3D$  window did not produce satisfactory results, as too many false positives were included in CT scans with small pixel spacing) iterates it, and if the shape index of the central pixel is greater or equal than 0.85, this window is checked. One of two rules needs to be true to label a window as possible part of a nodule. First rule is true if the central point and more four neighbor points have a shape index greater or equal than 0.9 and their curvedness is between  $[1.5, 3.5]$ . The second rule is similar, but the values of shape index needs to be greater or equal than 0.85, less than 0.9 and curvedness inside the interval of  $[2, 4]$ . If at least three windows fit into the first rule or at least 10 into the second rule, the superpixel is labeled as a nodule candidate. These values were obtained through several preliminary experiments.

Finally, a small group of superpixels are selected as nodule candidates. Although the number of candidates is smaller than before the nodule candidate selection step, a great number of non-nodules is present in the nodule candidates. Thus, in the next subsection, a feature extraction and classification step is explained, to reduce the number of false positives.

---

<sup>1</sup> Available at [https://itk.org/Doxygen/html/classitk\\_1\\_1HessianRecursiveGaussianImageFilter.html](https://itk.org/Doxygen/html/classitk_1_1HessianRecursiveGaussianImageFilter.html)

### 4.3.3 Feature extraction and classification

In the final step of our nodule detection approach, features are extracted from the selected nodule candidates to create a feature vector, serving as input for a binary classifier (nodule or non-nodule labels). Since nodules are more discriminant in their texture and shape and similar works employed classification with these features and achieved good results, the following features are extracted in this step: gray-level run length, binary and gray-scale Hu moments, volume (pixels and physical), perimeter, equivalent spherical radius, equivalent spherical perimeter, centroid, roundness, elongation, mean, standard deviation, minimum and maximum intensity, sum of the intensities, shape index mean and standard deviation.

Usage of gray-level run length features may help to distinguish nodules and airway objects, as they have similar intensities, but the later have more consecutive pixels/voxels (high length of the run). To label a superpixels as nodule, we check if at least 1% of it is part of the annotation done by the radiologists. Since there is annotation from four radiologists, we can select nodule the majority of radiologist marked as  $\geq 3\text{mm}$ , thus being more certain in the classification. Superpixels that does not satisfy this criteria are labeled as non-nodules.

One problem in the superpixel approach is the high number of non-nodule superpixels, thus a nodule candidate selection step prior to classification is applied. Although the number of non-nodule superpixels decreases, the proportion of non-nodule is higher than nodule superpixels. Although nodule candidate selection step reduced the number of non-nodule superpixels, this quantity still is higher than of nodule superpixels, leading to a unbalanced data set for the classifier. An unbalanced data set may significantly decreased the classification rate. Application of the Random Under-sampling technique to reduce sample from the non-nodule class is performed. A feature scaling is recommended, as features with high values can overlap ones with small values, also increases calculation complexity. Finally, RF classifier predict the generated superpixels. According to Oshiro, Perez & Baranauskas (2012), a number of trees between 64 and 128 would produce fast and satisfactory results, so a quantity of 128 was chosen for classification.

Chapter 5 details the flow of experiments and their characteristics for lung segmentation and nodule detection, followed by a discussion of the results obtained.



## 5 RESULTS AND DISCUSSION

In this chapter, a detailed description of the experiments and their results is present, followed by a discussion of their characteristics, pros and cons. Each major step is detailed in a unique section.

### 5.1 LUNG SEGMENTATION

The purpose of this step is detect the lung volume without losing important objects (*e.g.*, nodules), reducing processing time of the nodule detection step. Besides, LIDC/IDRI database do not have lung masks for evaluation of lung segmentation methods, needing another database to validate the process. So, for the preliminary tests, only analysis of volume reduction and loss of nodules is realized. For the loss of nodules (or parts from them), we used a union of the radiologists' annotation for each nodule, as even if a marked region is not a consensus, it may has important features.

For these approaches, we submitted the experiments in 563 CT scans from the LIDC/IDRI database. Given an original volume (CT scan or nodule)  $V_o$  and the segmented region  $V_s$ , the percentage  $p$  of the segmented region size if compared with the original image is calculated by

$$p = \frac{V_o}{V_s} \quad (5.1)$$

Below, the experiments and results for the two developed approaches.

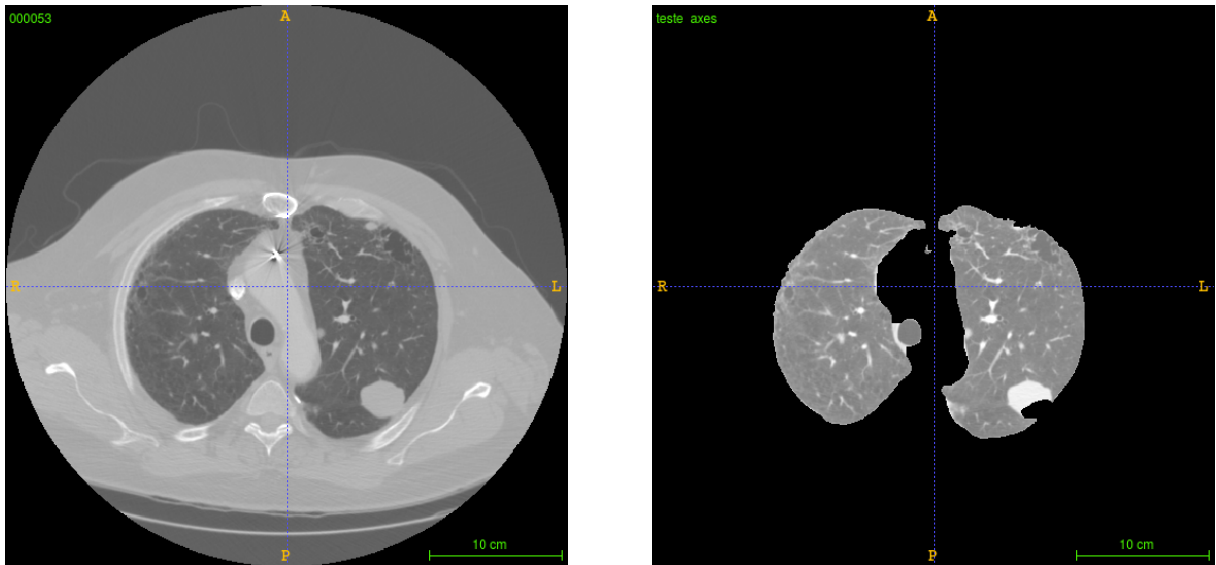
#### 5.1.1 Axes' Labeling

We realized two analysis: volume reduction and loss of nodules. The first analysis is done based on the reduction of volume, segmenting the lung to reduce the area of processing in the next step (nodule detection). None of the tested CT scans were left unsegmented. The smallest reduction, meaning the bigger value of  $p$ , for the tested CT scans was to a volume of 43.20% of the original size ( $p = 0.432017$ ). The bigger reduction was to a volume of 0.53% ( $p = 0.005339$ ). The average reduction of volume was to 12.15% ( $m = 0.121489$ ) with a standard deviation  $\sigma = 0.042640$ , demonstrating that processing volume of the next step is reduced to a little more than 1/10.

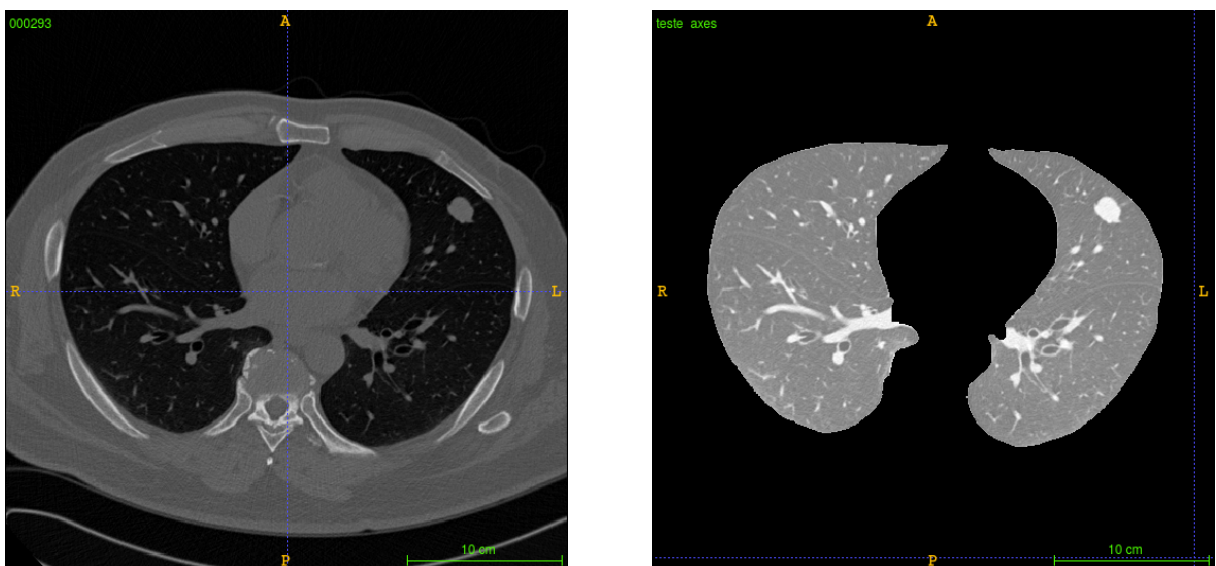
The second analysis is the most important one, as we evaluate how the segmentation process influenced the nodules present in each CT scan. From the entire CT scans tested, there was 1420 nodules with at least one radiologist marking it as  $\geq 3$  mm. From these nodules, 1010 (71.12%) were intact, not being influenced by the segmentation ( $p = 1$ ).

406 (28.60%) of the remaining nodules lost some parts of their regions and four (0.28%) were not included in the segmentation ( $p = 0$ ). From the 406 nodules, 154 preserved at least 90% of their regions ( $p \geq 0.90$ ), 85 between 80% and 90% ( $0.80 \leq p < 0.90$ ), 58 nodules between 70% and 80% ( $0.70 \leq p < 0.80$ ), 42 ranging in 60% and 70% ( $0.60 \leq p < 0.70$ ), 49 between 25% and 60% ( $0.25 \leq p < 0.60$ ) and 18 with less than 25% ( $p < 0.25$ ). Achieved a mean of  $0.93537 \pm 0.16075$ .

Figure 25(a) shows an image with a nodule in posterior right lung and Figure 25(b) shows the segmentation result using Axes' Labeling based approach. Part of the nodule was lost in the process. Figure 26 shows a good segmentation using this technique.



**FIGURE 25** – Example of the AL approach with partial loss of a nodule region. In A), a lung CT scan with a nodule in right lung. The nodule is together with parenchyma. In B), part of the nodule was lost in the process.

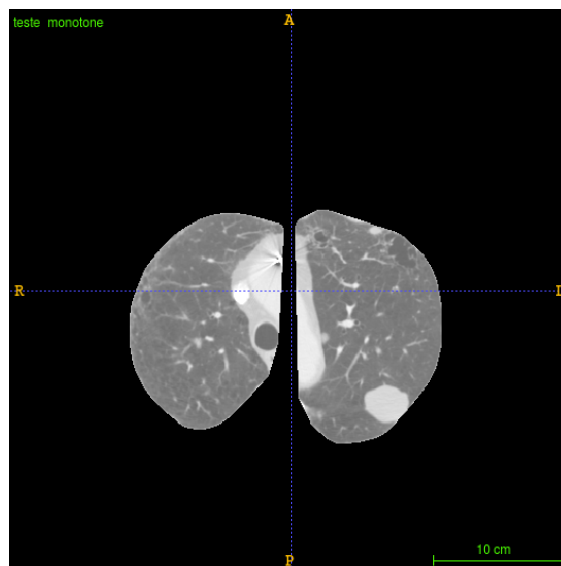


**FIGURE 26** – Example of the AL approach with nodule preservation.

### 5.1.2 Monotone Chain Convex Hull based approach

Volume reduction of the monotone chain convex hull based approach had the biggest reduction to 4.74% ( $p = 0.047370$ ) of the CT scan volume and the smallest to 36.15% ( $p = 0.361500$ ), with mean of  $p$  of 15.38% ( $m = 0.153737$ ) and standard deviation  $\sigma = 0.0473986$ . As well as the axes flood filling, this approach left no CT scan unsegmented.

For the loss of nodules analysis, we verified that 1142 (80.42%) nodules were entire segmented with the lungs (Figure 27) and one (0.07%) was not included in the segmentation ( $p = 0$ ). As for the other 277 (19.51%), 178 nodules had at least 90% of their regions preserved ( $p \geq 0.90$ ), 59 between 80% and 90% ( $0.80 \leq p < 0.90$ ), 20 nodules with preserved region between 70% and 80% ( $0.70 \leq p < 0.80$ ), 6 in 60% and 70% ( $0.60 \leq p < 0.70$ ), 11 between 25% and 60% ( $0.25 \leq p < 0.60$ ) and 3 with less than 25% ( $p < 0.25$ ). The mean of nodule preservation was  $0.9778 \pm 0.0830864$ . Figure 28 shows an bad segmentation of the lungs, where part of a nodule was removed.

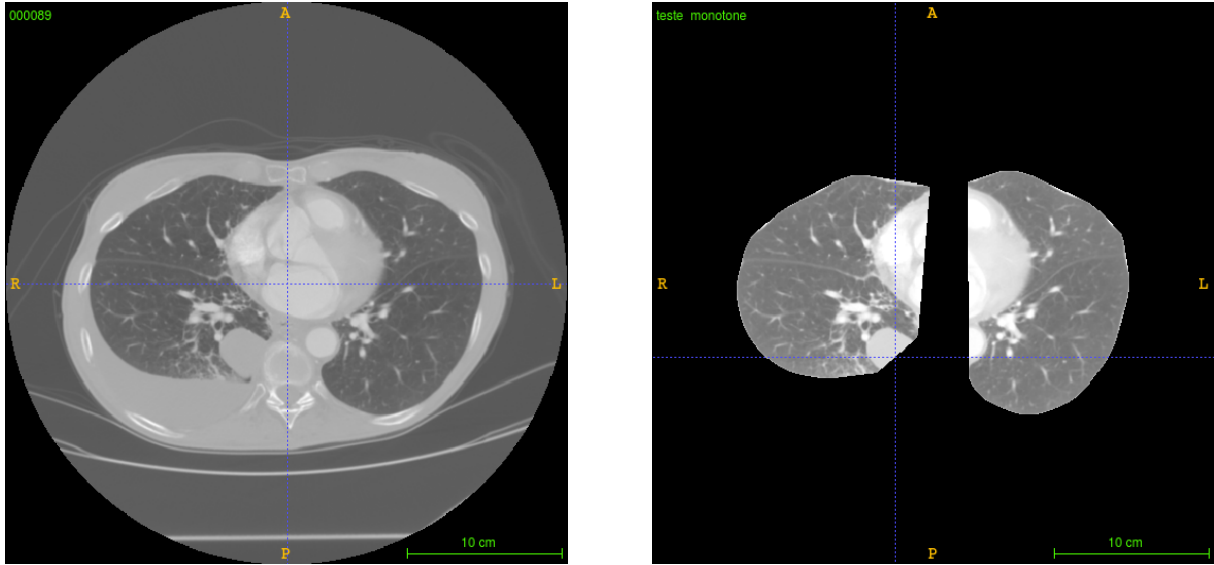


**FIGURE 27** – MCCH applied in Figure 25. The nodule was completely preserved, but part of another organ remained.

### 5.1.3 Comparison

Analyzing the results from both approaches, it was observed the AL segmentation reduced the CT scans (mean of  $p = 0.121489$  from the original size) more than the monotone chain convex hull (mean of  $p = 0.153737$ ), but it is necessary to verify if this greater reduction affected the nodules negatively.

Regarding loss of nodules, the MCCH segmentation had better results. First, left 1142 nodules intact, against 1011 from the AL approach. Then, one nodule was out of the segmentation, against four from the AL approach. With region loss at maximum of 30% ( $0.70 \leq p < 1$ ), the MCCH approach had 257, against 297 nodules from the AL approach.



**FIGURE 28** – Bad segmentation of the MCCH approach. CT scan with nodule positioned in posterior left lung in A). In B), part of the nodule was cut by the segmentation.

As for the remaining nodules ( $0 < p < 0.70$ ), the AL approach had way more nodules than the MCCH, with 109 against 20 nodules. With the above data, we can determine that the AL approach reduces more the CT volume, but with a high cost. The MCCH approach left unchanged more nodules and, even when nodule region was lost, it was not so aggressively as the AL, losing few parts for most part of these nodules. Furthermore, the MCCH method achieved a higher mean of nodule preservation ( $0.9778 > 0.93537$ ) and had a lower standard deviation ( $0.0830864$  against  $0.16075$  from flood fill), proving to be more stable, achieving similar results between CT scans.

## 5.2 NODULE DETECTION

Since we desire to verify the influence of nodules' parts' loss and lung segmentation with other parts not owned by lungs, we applied the superpixel techniques to both lung segmentation approaches, which may gives a better idea how our nodule detection approach will work if used with different lung segmentation techniques. A methodology of how to label a superpixel as nodule or non-nodule for classification is necessary. Since is desired to see the impact of the segmentation, where some superpixel may contain both nodule and non-nodule regions, a relative threshold  $T$  is utilized. A superpixel is labeled as nodule if  $T\%$  of its area/volume is contained in the radiologists' annotation; labeled as non-nodule otherwise. A total of four values were selected: 1, 25, 50 and 75%. Not only that, but the nodule annotations may differ from each radiologist, so superpixels are checked by two forms: if it is inside the union of these annotations; and inside a region with the majority of radiologists labeling as nodule (with at least three radiologists). This is analyzed to verify the impact of the radiologists' annotations in automatic nodule detection.

The LIDC database was divided into two data sets: for development (training) and validation (testing) of our methodology. Development data set was composed of the CT scans from patients with IDs from 0001 to 0600 and validation data set from IDs from 0601 to 1010. This division (almost 60%/40%) provided a great number for development and validation of our approach, and clearly separating the CT scans helps in the validation of future works with ours as more reliability in the work with usage of the entire database. A few scans presented errors to load its entirely using the ITK's DICOM series read image class (this error was verified in the ITK-SNAP<sup>1</sup>, a medical image segmentation and visualization tool). Since only a few CT scans presented this problem, they were removed from the data sets. A further verification of the DICOM series reader class is encouraged.

For classification of the development data set, the leave-one-out approach was applied. Classification was realized with every superpixel from before the down-sampling, as the results would be more reliable if every generated superpixel is checked. Training set continued to have only samples from after the down-sampling (minus the current sample in testing). Adjustments were realized for improvement of our approach. Then, a one-time classification was realized using the development data set for training and the validation data set for testing, using the parameters from the leave-one-out classification.

Experiments were divided into two main steps: superpixel generation and selection; and feature extraction and classification. FH (2D) and SLIC (2D and 3D) superpixels were analyzed to a better understanding of the problem. First, our approach was analyzed in the development data set and then, a untouched data set (validation data set) was utilized to validate the results obtained.

### 5.2.1 Development stage

In this stage, experiments were realized with CT scans from patients 0001 to 0600, containing 838 nodules which were marked by at least three radiologists as nodules  $\geq 3mm$ , for development and analysis of our approach. Several experiments were realized to adjust the aspects of the employed techniques in our approach and then, final experiments were realized with this data set for analysis of the obtained results.

#### 5.2.1.1 Superpixel generation and selection

Before superpixel generation by the FH and SLIC methods, some parametrization of both is necessary. The  $K$  threshold for FH determines how aggressive a segmentation (merging of points) is. A value of  $K = 500$  was obtained empirically, based on the values utilized in the method's article (FELZENSZWALB; HUTTENLOCHER, 2004). For the SLIC method, the values of  $m = 10$  and  $S = 12mm$  for 2D and  $S = 9mm$  for 3D were used.

---

<sup>1</sup> Tool available at <<http://www.itksnap.org/>>. Last accessed in 8 Aug 2016.

As stated in the chapter 4, the nodule candidate selection approach employed is based on the shape index and curvedness features. A window is iterated through the superpixel and for each time, two rules were checked. First rule is if the central point plus more four neighbor points had a shape index within  $[0.9, 1]$  and a curvedness between  $[1.5, 3.5]$ . As for the second rule, the shape index is necessary to be within the interval of  $[0.85, 0.]$  and curvedness  $[2, 4]$ . If the first rule is true three times or the second 10 times, the superpixel is selected.

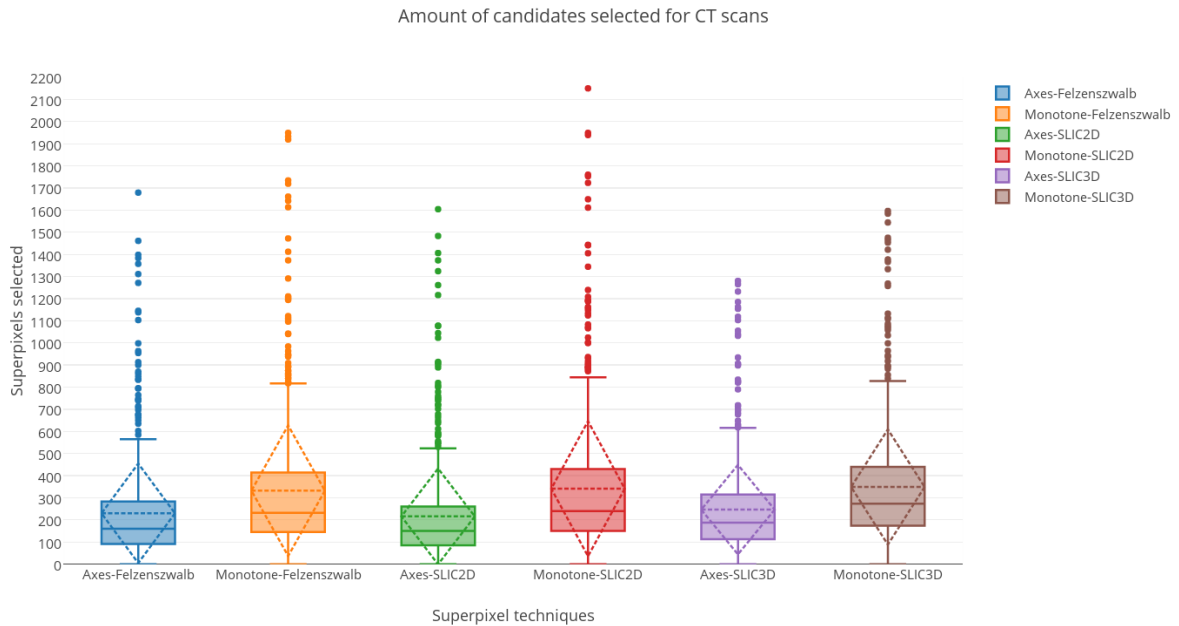
Since superpixel selection's main objective is to select only a few candidates for classification, an analysis of the results of this reduction is necessary, to show the necessity of the usage and improvement of this approach. Table 3 presents the original amount of superpixels generated for each technique and the amount of superpixels selected using the shape index and curvedness approach. A percentage of the final amount in relation to the original amount is shown.

**TABLE 3** – Average of superpixels generated and selected per CT scan, showing the amount of superpixels selected if compared to the original amount (in %)

	AL			MCCH		
	$\mu_{original}$	$\mu_{selected}$	Reduced to	$\mu_{original}$	$\mu_{selected}$	Reduced to
FH ( $K = 500$ )	20821.45	229.82	1.10%	26430.80	332.11	1.26%
SLIC ( $S = 12$ )	66073.68	215.78	0.33%	86913	341.05	0.39%
SLIC 3D ( $S = 9$ )	16295.81	246.52	1.51%	22091.54	348.86	1.58%

Analyzing the mean of nodule candidates selected, the amount of generated superpixels of the SLIC 2D technique is more than three times the amount of FH superpixels and near four times its 3D approach. A box-plot of the candidates selected for each CT scan is shown in Figure 29. This type of diagram can display graphically the distribution of data (BENJAMINI, 1988). Explaining the box-plot, the rectangular box is the interquartile range (IQR), which comprises of data from the first to the third quartile of the distribution, with the continuous line inside it being the median. Horizontal lines outside the IQR are the whiskers, which separates the outliers (points) from the remaining data. Additionally, a dashed diamond and line are displayed in further box-plots. They are the standard deviation and mean of the distribution, respectively.

Number of outliers for each technique is noticeable, where the minimum amount of CT scans as outliers is found in the SLIC 3D ( $n = 30$ ) and maximum amount with the MCCH-FH approach ( $n = 41$ , almost 7% of the data set). Analyses shown the correlation between outliers and low values of pixel spacing and slice thickness, which the shape index–curvedness based candidate selection found too many superpixels with high shape index windows (high presence of noise). Tables 4 and 5 show the average of candidates selected for each superpixel method with different thresholds and annotation type, which



**FIGURE 29** – Boxplot of the CT scans’ candidates after application of the candidate selection for each superpixel generation approach.

demonstrates how high the candidate samples are unbalanced (calculation includes CT scans without nodules too).

**TABLE 4** – Average of candidates selected labeled as nodules for each superpixel method using the MCHH lung segmentation

	FH ( $K = 500$ )		SLIC 2D ( $S = 12$ )		SLIC 3D ( $S = 9$ )	
$T$	Union	Majority	Union	Majority	Union	Majority
1%	4.94	4.78	6.43	6.22	3.81	3.53
25%	4.23	3.75	5.27	4.95	2.33	2.15
50%	3.17	2.62	4.65	4.41	1.94	1.81
75%	2.29	1.88	4.13	3.92	1.63	1.47

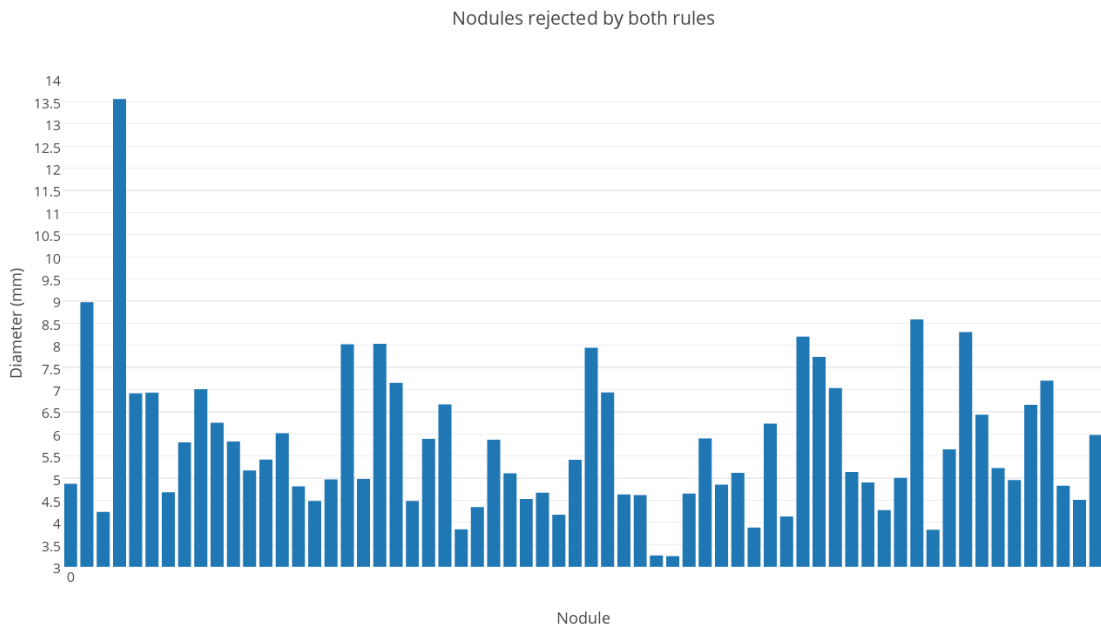
Wrong choice of shape index minimum values and curvedness intervals combined with superpixel segmentation interfere in a properly candidate selection. Verification of the shape index and curvedness impact with a ideal segmentation (radiologists’ annotation) may answer the results’ weight of both selection and segmentation steps.

From the 838 nodules, 179 did not had the minimum windows with high shape index ( $n \geq 3$ ), which 148 had none first rule’s windows. 64 of this 179 nodules did not met the requirements of the second rule too ( $\geq 10$  windows), which 19 had none windows for both rules. One thing to note is the quantity of small nodules present in these numbers. From the 64 nodules rejected by both rules, only one nodule had a diameter  $\geq 9mm$  ( $13.56mm$ ). Average of their sizes was  $5.77mm \pm 1.68$ . An overall of their sizes is presented

**TABLE 5** – Average of true nodule candidates selected. Lungs segmented using the AL approach.

	FH ( $K = 500$ )		SLIC 2D ( $S = 12$ )		SLIC 3D ( $S = 9$ )	
$T$	Union	Majority	Union	Majority	Union	Majority
1%	4.76	4.62	6.36	6.19	3.80	3.48
25%	4.17	3.71	5.33	5.01	2.42	2.25
50%	3.13	2.58	4.76	4.59	2.09	1.96
75%	2.24	1.85	4.35	4.14	1.78	1.62

in Figure 30. As for the second rule, 101 nodules were not selected, mostly small nodules, therefore the main group of non-selected nodules.

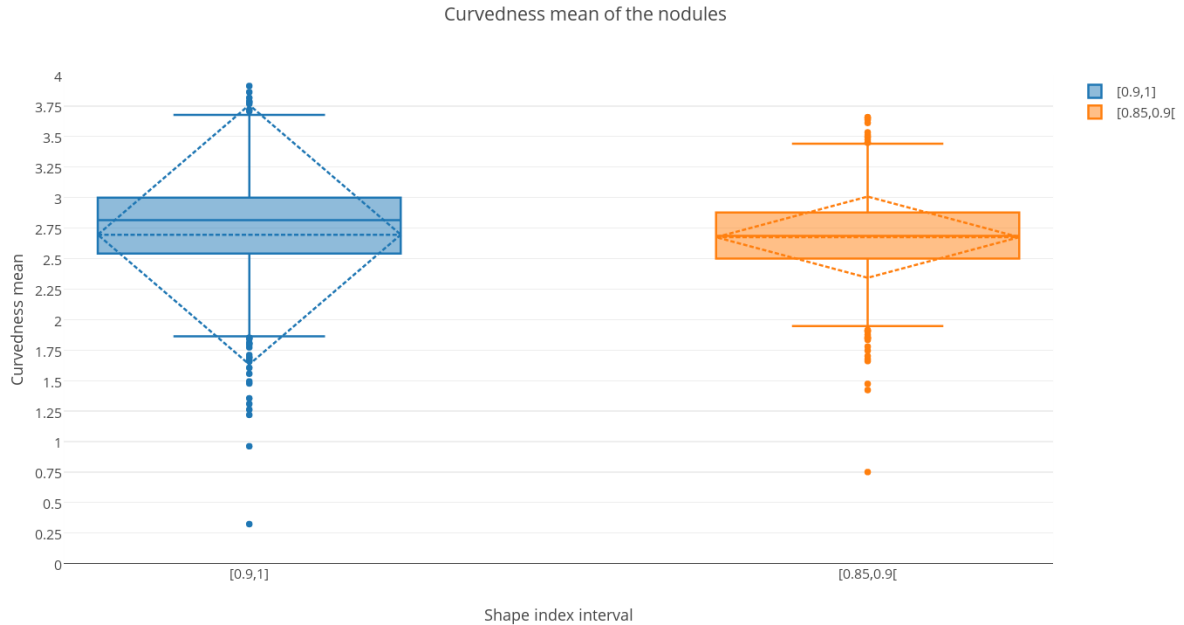
**FIGURE 30** – Diameters of the nodules which did not meet the requirements of both rules

For each nodule, a curvedness mean of its voxels was calculated for a better understand of the overall curvedness of data set. For the 838 nodules present in the data set, nine did not had any voxel with a shape index in the interval  $[0.9, 1]$  and one nodule without shape index in the interval  $[0.85, 0.9[$ . For the remaining nodules for each rule, the mean and standard deviation were calculated based on the curvedness mean previous extracted. For the nodules with shape index within  $[0.9, 1]$ , the mean and standard deviation of the curvedness were  $2.70 \pm 1.06$ . As for the second rule, the mean and standard deviation  $2.68 \pm 0.33$  were found. Distribution of the nodules' curvedness mean is shown in Figure 31 (the outliers  $[-1.64, -5.54, -8.04, -8.40, -19.02]$  are not present as they are far away from the other values, thus messing up the boxplot).

The curvedness intervals defined in our methodology were obtained empirically.



Comparison of the empirically obtained intervals with the box-plot shows that the defined interval of the first rule have close values to the whiskers (the two horizontal lines outside the box) and the standard deviation, although outliers might had increased greatly the standard deviation. One the second rule case, the whiskers had close values to the defined interval, although not so close with the maximum value. A balanced distribution is seen the second rule, as the mean and median have close values.



**FIGURE 31** – Boxplot of the distribution of curvedness mean of nodules.

It is necessary to analyze how the selection of superpixels affected the annotated nodules from the data set. Verification of the nodules selected for each technique are presented (Table 6). From a total of 838 nodules (with at least three radiologists labeling as nodule  $\geq 3mm$ ) using the MCCH approach, 743 nodules passed to the next phase (88.66%) with the FH approach. SLIC 2D achieved better results, with 760 nodules (90.69%) being part of the nodule candidates. Best candidate selection was achieved by the SLIC 3D, with a total of 779 nodules (92.96%) passing to the next phase. With the AL segmentation, lower results were obtained. 730 (87.11%), 737 (87.95%) and 769 (91.77%) were selected for the FH, SLIC and SLIC 3D methods, respectively.

**TABLE 6** – Relation of the total of nodules selected from the development data set for  $T = 1$  and union (1u)

	AL	MCCH
FH ( $K = 500$ )	730 (87.11%)	743 (88.66%)
SLIC ( $S = 12$ )	737 (87.95%)	760 (90.69%)
SLIC 3D ( $S = 9$ )	769 (91.77%)	779 (92.95%)

Tables 7 and 8 present a more detailed data about selection of true nodules in the candidate selection step. Although SLIC 3D had initially a high rate of nodule selection, its results decreased greatly with the increase of the threshold  $T$ , which may describe high over-segmentation, which not only can lead to a worse classification, as more irrelevant data is include in the sample, but the non-selection of the nodules. FH and SLIC 2D had more stable results, but many nodules were lost with high thresholds (only 54% nodules were preserved with a threshold 75m using the FH technique).

**TABLE 7** – A more detailed data about nodules selected to the next phase for the MCCH lung segmentation

	FH ( $K = 500$ )		SLIC 2D ( $S = 12$ )		SLIC 3D ( $S = 9$ )	
$T$	Nodules	%	Nodules	%	Nodules	%
Union						
1	743	88.66	760	90.69	779	92.96
25	696	83.05	651	77.68	465	55.49
50	613	73.15	546	65.16	351	41.89
75	537	64.08	514	61.34	295	35.20
Majority						
1	737	87.95	753	89.86	773	92.24
25	660	78.76	600	71.60	421	50.24
50	561	66.95	520	62.05	331	39.50
75	478	57.04	494	58.95	270	32.22

**TABLE 8** – Results of the nodule candidate selection using the AL segmentation

	FH ( $K = 500$ )		SLIC 2D ( $S = 12$ )		SLIC 3D ( $S = 9$ )	
$T$	Nodules	%	Nodules	%	Nodules	%
Union						
1	730	87.11	737	87.95	769	91.77
25	686	81.86	657	78.40	483	57.64
50	614	73.27	571	68.14	385	45.94
75	529	63.13	535	63.64	320	38.17
Majority						
1	725	86.52	734	87.59	761	90.81
25	653	77.92	550	65.63	445	53.10
50	554	66.11	550	65.63	365	43.56
75	478	57.04	526	62.77	296	35.32

For this part, some characteristics of the superpixel generation and selection can be noted. First, the quantity of nodules for each technique was close to each other. Although the intention of the candidate selection was to reduce the different between the

number of nodules and non-nodules, a great difference is seen even after the candidate selection.

Percentage of nodules passed to the next phase for the 2D techniques were worse than the SLIC 3D, especially in the FH approach. One of the motives is the scattering of high shape index of a nodule across different slices and quantity of it, not being included in the nodule candidate rules described in the previous chapter (subsection 4.3.2). Then, it is preferable to utilize 3D segmentation approaches combined with this candidate selection. Loss of these nodules will impact the overall result, even if the classification achieves good results, so an analysis not only overall, but for selected nodules is encouraged.

One thing to consider is the superpixel's size would influence in candidate selection, since a large superpixel may have many scattered high shape index windows, but overall has few high shape index proportionally, and small superpixels may have not sufficient area to have a acceptable number of high shape index windows. Therefore, a candidate selection based not only in the quantity of high shape index windows, but proportionally to the superpixel size would be more suitable.

Analyzing more the superpixel techniques, the loss of nodule regions in this step or even nodules divided in many different superpixels could influence further steps and show unique patterns of each technique.

#### 5.2.1.2 Classification

For the final step, a classification was performed using the Random Forest (RF) classifier. The objective of this classification was reduction of false positives from the nodule candidates while minimizing the number of false negatives (i.e. rejection of real nodules). A leave-one-patient classification approach was employed for a better usage and validation of the data. Before classification of the patient's samples, a random under-sampling was applied to the training data set for reduction of the majority (non-nodule) class, thus generating a balanced data set (1:1 samples). A feature vector with the features informed in Chapter 4 (as Hu moments were only extracted in 2D images, classification of SLIC supervoxels did not include them) was generated for each sample. They were normalized to a space with zero mean and variance equal to one.

Default RF classification obtained to many FPs, therefore a probability classification as used, where samples with probability of at least 75% of being a nodule were classified as it. As the superpixel segmentation can divide a nodule in more than one superpixel, analysis of the results was realized based on the nodules, not superpixels. Verification of superpixels containing nodules was realized, and for each nodule not contained in the classified superpixels, it was labeled as a false negative. Classified superpixels which does not contain nodules were labeled as false positives. This rule was applied to represent the detection of nodules, as prediction of many superpixels from one nodule may produce

a biased result.

Sensitivity (TPR) and FP/scan were extracted from the classification process and are shown in Tables 9 and 10. FH superpixel classification achieved the best results. For this technique, classification with 1% and 25% for  $T$  resulted in similar results, for both union and majority annotations, with the AL-1u having the best sensitivity (69.33%).  $T = 50$  and  $T = 75$  had worse sensitivities, especially 75m (55.97% for MCCH and 56.68% for AL segmentation). Overall, union annotation had better sensitivity than majority. Increase of  $T$  reduced the number of FP/scan and majority annotation had less FPs than union. SLIC 2D had similar results in sensitivity with FH using the AL approach, but usage of MCCH reduced drastically its sensitivity and increased the number of FP/scan. SLIC 3D had similar results than its 2D counterpart with the MCCH usage. Over-segmentation and low rate of true nodule selected contributed to its low results.

**TABLE 9** – Overall results for classification with three types of superpixel segmentation, using the MCCH lung segmentation approach

Type	FH (K=500)		SLIC 2D (S=12)		SLIC 3D (S=9)	
	TPR	FP/scan	TPR	FP/scan	TPR	FP/scan
Union						
1	66.11%	8.45	43.79%	8.68	47.97%	8.26
25	67.30%	8.73	47.97%	11.74	44.03%	7.24
50	63.96%	8.39	48.69%	13.14	41.05%	6.30
75	60.38%	7.73	48.09%	13.12	39.02%	5.57
Majority						
1	66.35%	8.26	43.44%	8.76	49.28%	8.29
25	65.99%	8.58	46.66%	12.55	43.20%	6.91
50	61.10%	7.71	49.28%	13.43	39.98%	5.61
75	55.97%	7.06	44.63%	8.59	37.35%	5.04

Overall, lung segmentation based on AL had best results than MCCH. Although the latter preserved more nodule regions, it included many non-nodule regions which interfered in further steps.

Table 11 shows the sensitivity of the classification with rejected nodules, that were not selected in the candidate selection step, not being counted as false negatives. Higher thresholds achieved better results, which may concludes the weight of the segmentation in the final results.

The LIDC database consists of nodules with diameter in the wide range of [3mm, 30mm]. Five groups of nodules based on their sizes were created to verify how our approach behaves with different sizes. Nodules were grouped according to five ranges:  $\leq 5mm$ ;  $> 5mm$  and  $\leq 10mm$ ;  $> 10mm$  and  $\leq 15mm$ ;  $> 15mm$  and  $\leq 20mm$ ;  $> 20mm$  and  $\leq 30mm$ . Results are presented in Tables 12, 13, 14, 15, 16 and 17. Analyzing

**TABLE 10** – Overall results for classification based on the AL approach

Type	FH (K=500)		SLIC 2D (S=12)		SLIC 3D (S=9)	
	TPR	FP/scan	TPR	FP/scan	TPR	FP/scan
Union						
1	69.33%	6.91	62.89%	6.59	49.28%	5.57
25	69.09%	6.88	62.29%	7.34	48.09%	5.84
50	65.63%	6.07	60.38%	8.76	44.27%	5.03
75	60.38%	5.28	59.55%	9.27	42.24%	4.52
Majority						
1	68.62%	6.87	61.22%	8.85	48.09%	6.05
25	67.18%	6.67	60.02%	10.78	46.18%	5.62
50	60.86%	5.38	60.38%	7.24	43.56%	4.80
75	56.68%	5.05	59.19%	7.07	40.45%	4.07

**TABLE 11** – Sensitivity results considering only the selected nodules. Percentages based on the data from Tables 7 and 8.

Type	FH ( $K = 500$ )		SLIC 2D ( $S = 12$ )		SLIC 3D ( $S = 9$ )	
	MCCH	AL	MCCH	AL	MCCH	AL
Union						
1	74.56%	79.59%	48.29%	71.51%	52.28%	53.71%
25	81.03%	84.40%	61.75%	79.45%	79.35%	83.44%
50	87.44%	89.58%	74.73%	88.62%	98.01%	96.36%
75	94.23%	95.65%	78.40%	93.27%	100%	100%
Majority						
1	75.44%	79.31%	48.34%	69.89%	53.43%	52.96%
25	83.79%	86.22%	65.17%	91.45%	85.99%	86.97%
50	91.27%	92.06%	79.42%	92%	100%	100%
75	98.12%	99.37%	75.71%	94.30%	100%	100%

the results obtained, some statements can be made. Foremost, the first two groups had sensitivity results much lower than the remaining groups, mostly using the SLIC method. FH superpixel achieved better results with small nodules, but these results were far behind than those with large nodules. Improvement of the sensitivity can be seen as the nodule size increases. Modifications to the approach for detection of small nodules are necessary.

Another analysis realized is the ranking of importance of each feature according to the RF classifier, using the mean decrease impurity is utilized to rank the features by their scores. Scores from extracted features were analyzed for the three superpixel generation techniques, since they may have different relevant features. For the three techniques, the most important features were the shape index based (mean and standard deviation) and the first Hu moment (binary and gray-scale). Other features such as mean, standard

**TABLE 12** – Overall sensitivity based on the nodule size using the AL lung segmentation method and FH superpixel

$T$	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 30]
Union					
1	44.12%	66.23%	86.03%	90.32%	97.62%
25	41.91%	66.23%	86.76%	90.32%	97.62%
50	40.44%	62.31%	80.15%	90.32%	97.62%
75	37.50%	55.12%	75.74%	90.32%	95.24%
Majority					
1	39.71%	66.88%	84.56%	88.71%	97.62%
25	30.11%	64.08%	88.50%	90.20%	97.30%
50	34.56%	57.30%	75%	88.71%	95.24%
75	36.76%	52.51%	66.18%	85.48%	90.48%

**TABLE 13** – Overall sensitivity based on the nodule size using the AL lung segmentation method and SLIC superpixel

$T$	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 30]
Union					
1	25%	59.91%	89.71%	90.32%	88.10%
25	22.79%	58.82%	91.18%	88.71%	92.86%
50	18.38%	57.08%	89.71%	88.71%	92.86%
75	22.06%	54.47%	88.97%	91.94%	90.48%
Majority					
1	22.79%	57.30%	90.44%	87.10%	92.86%
25	21.32%	55.34%	89.71%	87.10%	97.62%
50	20%	53.85%	94.92%	87.10%	89.47%
75	16.67%	52.75%	91.53%	87.10%	89.47%

deviation, roundness and run length features’ means scored high values. Third to seventh Hu moments achieved the lowest scores.

### 5.2.2 Validation stage

The purpose of this stage is not only to apply our work in a different, untouched data set, but to verify the similarities between the results from both development and validation stages. This data set is composed of the CT scans from patients 0601 to 1012 from LIDC database. A total of 495 nodules marked as  $\geq 3mm$  by at least three radiologists can be found in this data set.

A total of five approaches were utilized in the validation stage: using the AL segmentation, FH 1u and 25u, SLIC 1u and SLIC 3D 1u; FH 25u for MCCH segmentation. These approaches were selected based on their results in the previous stage and with the

**TABLE 14** – Overall sensitivity based on the nodule size using the AL lung segmentation method and SLIC supervoxel

$T$	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 30]
Union					
1	10.29%	39.65%	86.03%	91.94%	95.24%
25	9.56%	37.91%	86.03%	90.32%	95.24%
50	7.35%	31.59%	86.76%	90.32%	92.86%
75	5.88%	29.19%	83.82%	91.94%	90.48%
Majority					
1	11.03%	38.34%	84.56%	88.71%	92.86%
25	8.82%	34.86%	86.76%	88.71%	92.86%
50	7.35%	30.94%	86.03%	88.71%	90.48%
75	5.15%	27.89%	80.88%	88.71%	85.71%

**TABLE 15** – Overall sensitivity based on the nodule size using the MCCH lung segmentation method and FH superpixel

$T$	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 30]
Union					
1	38.97%	61.44%	86.76%	90.32%	100%
25	41.91%	62.75%	86.76%	91.94%	97.62%
50	39.71%	58.39%	85.29%	88.71%	95.24%
75	36.76%	54.90%	78.68%	87.10%	95.24%
Majority					
1	40.44%	62.53%	86.03%	87.10%	95.24%
25	42.65%	61.22%	86.03%	87.10%	95.24%
50	38.24%	55.12%	80.88%	87.10%	95.24%
75	35.29%	50.76%	70.59%	80.65%	92.86%

objective to include every lung segmentation and superpixel generation approach. Analysis of each step of the nodule detection is present below.

First, the superpixel generation and selection using the shape index and curvedness are checked. Table 18 shows the relation between the superpixels generated and then the average of superpixels selected per scan.

A complementary data is present in Table 19, which shows the average not only of superpixels selected per scan, but the average of superpixels labeled as nodules. The same problem present in the development data set occurs in this data set too. A high unbalance of sample is found, where the number of non-nodule superpixel surpass  $100\times$  the quantity of nodule superpixels.

The boxplot in the Figure 32 shows the distribution of candidates selected across the CT scans of the validation data set. If compared with the boxplot of the development

**TABLE 16** – Overall sensitivity based on the nodule size using the MCCH lung segmentation method and SLIC superpixel

$T$	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 30]
Union					
1	10.29%	34.86%	76.47%	85.48%	78.57%
25	12.50%	39.65%	80.88%	90.32%	80.95%
50	11.76%	40.52%	83.09%	90.32%	80.95%
75	10.29%	40.30%	80.88%	88.71%	85.71%
Majority					
1	8.09%	35.29%	76.47%	82.26%	78.57%
25	11.76%	38.34%	79.41%	82.26%	88.10%
50	13.24%	40.96%	80.88%	90.32%	90.48%
75	9.56%	37.03%	70.59%	82.26%	83.33%

**TABLE 17** – Overall sensitivity based on the nodule size using the MCCH lung segmentation method and SLIC supervoxel

$T$	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 30]
Union					
1	16.18%	39%	80.15%	87.10%	83.33%
25	11.03%	33.55%	80.88%	85.48%	80.95%
50	8.82%	28.10%	81.62%	87.10%	83.33%
75	8.09%	25.93%	77.21%	85.48%	85.71%
Majority					
1	17.65%	40.09%	83.82%	85.48%	83.33%
25	11.76%	31.59%	82.35%	83.87%	80.95%
50	8.82%	26.80%	81.62%	85.48%	78.57%
75	7.35%	23.97%	77.94%	83.87%	76.19%

**TABLE 18** – Amount of original superpixels generated for each approach and the amount after application of the shape index and curvedness based candidate selection.

	$\mu_{original}$	$\mu_{selected}$	Reduced to
AL-FH	29037.22	334.22	1.15%
AL-SLIC	86355.29	313.16	0.36%
AL-SLIC3D	17525.53	348.62	1.99%
MCCH-FH	38769.80	475.95	1.23%

data set previous shown (Figure 29), a higher number of selected superpixels can be seen.

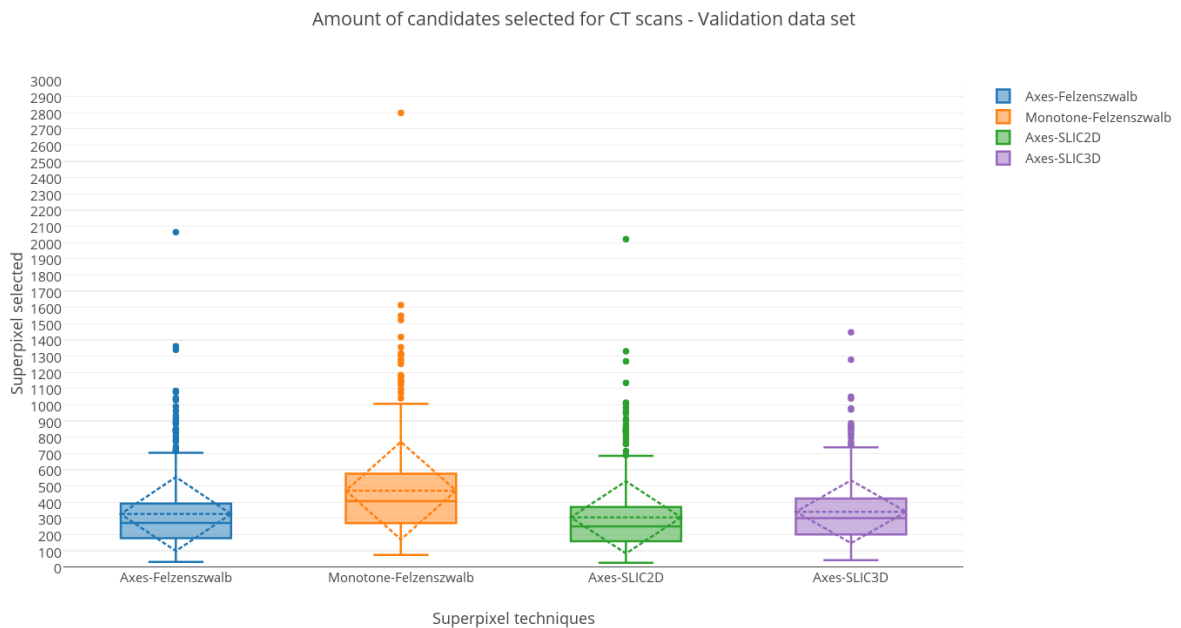
Two rules were determined for candidate selection. For both, a shape index interval was set,  $[0.9, 1]$  and  $[0.85, 0.9]$ , combined with a curvedness interval of  $[1.5, 3.5]$  and  $[2, 4]$ , respectively. A windows is counted if its central points plus four neighbors follow one of the rules. If for the first rule exists three windows or 10 windows for the second, this



**TABLE 19** – Relation of the superpixels selected for each approach to the amount of superpixels labeled as nodules.

	$\mu_{selected}$	$\mu_{true\_nodules}$
AL-FH-1u	334.22	4.25
AL-FH-25u	334.22	3.64
AL-SLIC-1u	313.16	5.61
AL-SLIC3D-1u	348.62	2.75
MCCH-FH-25u	475.95	3.77

superpixel is selected. Analyzing the windows of shape index and curvedness using as basis the radiologists' annotation, an amount of 114 (23.03%) nodules did not met the requirement of the first rule ( $n \geq 3$ ), where 94 did not had any windows ( $n = 0$ ). For the second rule, 56 (13.31%) nodules had less than 10 windows, which 6 had none. 37 nodules (7.47%) did not had sufficient windows for both rules.



**FIGURE 32** – Application in the validation data set of the candidate selection approach for each superpixel generation approach.

Finally, the nodule candidates are classified by the Random Forest method, using the development data set as the training set. As well as in the previous stage, the FH method with threshold  $1u$ , applied to the lungs segmented by the AL, had the best results. Although the FP/scan had similar values between the two stages (6.91 for the development and 7.2 for the validation), the sensitivity in the validation stage was far worse (60.61%) than in the development stage (69.33%).

Table 21 shows the sensitivity of groups based on nodule size. Results presented

**TABLE 20** – Classification results of the validation data set by different approaches.

	TPR	FP/scan
AL-FH-1u	60.61%	7.2
AL-FH-25u	59.60%	7.89
AL-SLIC-1u	53.94%	7.94
AL-SLIC3D-1u	33.54%	4.18
MCCH-FH-25u	60%	10.25

similar results for the first group and the two last. For the second ( $]5, 10]$ ) and third ( $]10, 15]$ ) groups, their sensitivities are more close than in the development stage, which the second group had improvements and the third group lower results.

**TABLE 21** – Classification results of the validation data set by different approaches.

$T$	$[0, 5]$	$]5, 10]$	$]10, 15]$	$]15, 20]$	$]20, 30]$
AL-FH-1u	37.50%	64.73%	69.35%	85.71%	94.44%
AL-FH-25u	37.50%	62.79%	69.35%	85.71%	94.44%
AL-SLIC-1u	28.91%	55.43%	72.58%	89.29%	88.89%
AL-SLIC3D-1u	3.91%	32.56%	59.68%	78.54%	94.44%
MCCH-FH-25u	40.62%	60.85%	69.84%	85.71%	100%

### 5.3 DISCUSSION

In this section, a overall discussion about the techniques employed in the lung segmentation and nodule detection stages, and results obtained in experiments previously commented are presented, concluding with a comparison of other nodule detection works.

Over-segmentation of the lungs resulted in lower results for nodule detection (MCCH method), which resulted in a increase of samples and many other objects had similar features to nodules. For example, usage of the AL method reduced the amount of candidates to 70% of the amount generated using the MCCH method. In the superpixel selection step, FH method had worse results for  $T = 1$ . One thing to consider is some parts of small nodules may be included in large FH superpixels, not being labeled as nodules. A possibility is the threshold value  $K$  not being sufficient good for some regions. SLIC 3D had the best candidate selection with the lowest threshold ( $1u$ ) because high shape index windows were scattered across different slice in certain nodules, but the grid-like segmentation of the SLIC approach increases the similarity between nodule and non-nodule superpixels. Not only that, but the close size of each superpixel (controlled by the step value  $S$ ) produces over and under-segmentation, since the variance of nodule sizes is high (interval of  $[3mm, 30mm]$ ), thus a superpixel with a small nodule may contains a

high proportion of non-nodule pixels and big nodules can be included in more than one superpixel.

Candidate selection based on shape index and curvedness can reduce greatly the number of samples generated by a superpixel approach. One of the problems of the approach is to determine how the superpixels will be selected for the next phase. Rules were determined empirically (described in chapter 4) based on the quantity of superpixels and % of nodules selected. Then, formation of the superpixels, as their initial quantity, would impact how strict the rules are made. An improvement on the previous step would improve this step and further, as more nodule superpixels would be selected (loosing or not the rules) and the total of selected superpixels would decrease (although only if a group of selected adjacent regions are made into one).

Final step was the feature extraction and classification of nodule candidates generated in previous phases. Shape, texture and statistical features were extract to represent the superpixels. FH method was more discriminant in relation to the SLIC superpixel (for both 2D and 3D). An overview of the results of different works and our own is presented in Table 22. Only works with similar measures (TPR and FP/scan) are included in this table, since it was the measures used in our work. Overall, the proposed approach had similar FP/scan results than other works, but had a sensitivity (TPR) lower than them. Unbalanced data and segmented superpixels not correctly representing the objects of interest contributed negatively to the classification results. High number of non-nodules combined with low discriminant objects reduced the sensitivity and increased the number of FPs.

**TABLE 22** – Comparison of different works with our approach. Results from validation stage are shown.

Authors	Data set	Samples	TPR	FP/scan
Ye et al. (2009)	Inhouse	54 CT scans	90.2%	8.2
Ashwin et al. (2012)	LIDC	40 CT scans	92%	0.2
Keshani et al. (2013)	ELCAP	397 nodules	89%	7.3
Brown et al. (2014)	LIDC	120 CT scans	79.2%	2.05
Choi & Choi (2014)	LIDC	84 CT scans	97.5%	6.76
Filho et al. (2014)	LIDC	140 CT scans	85.91%	1.82
Jacobs et al. (2014)	NELSON	122 and 60 nodules	80%	1
Demir & Çamurcu (2015)	LIDC	200 CT scans	93.6%	2.45
Our approach	LIDC	495 nodules	60.61%	7.2

A high difference in sensitivity between small and large nodules was found. First, for the AL lung segmentation method, the first two groups ( $[0, 5]$  and  $]5, 10]$ ) had worse results, mainly with the SLIC supervoxel. Sensitivities of the FH method for the first group were between 30.11% and 44.12%, much higher than for the SLIC superpixel (between

16.67% and 25%) and supervoxel (between 5.15% and 11.03%) methods. In the second group, differences between the FH and SLIC superpixel methods were lower (between 52.51% and 66.88% for FH and, for SLIC superpixel, between 52.75% and 59.91%). SLIC superpixel had better results for this group (if compared with the first one), but were lower than the FH and SLIC superpixel results (between 27.89% and 39.65%). Results for the last three groups improved significantly. For the third group ([10, 15]), SLIC superpixel had the best sensitivities, ranging from 88.97% to 94.92%, followed by the supervoxel approach (between 80.88% and 86.76%), with the FH method having more variant sensitivities (between 66.18% and 88.50%). In the fourth group, the three methods had similar sensitivities, varying from 87.10% to 91.94%). Sensitivities in the group with the largest nodules ([20, 30]) were better overall, highlighting the FH method, where the eight experiments had sensitivities higher than 90%, where seven were higher than 95% and five were higher than 97%.

The results obtained using the lung segmentation generated by the MCCH method had similar results than the AL method, but a major difference was found. Sensitivities with the SLIC superpixel were slightly lower using the MCCH segmentation, ranging from 8.09% to 13.24% in the first group, between 35.29% and 40.96% for the second group, 70.59% to 83.09% for the third group and in the fifth group, sensitivities varying from 78.57% to 90.48%. As previously stated, many superpixels from the central area of the CT scan were included in the MCCH segmentation and they had similar features than of nodule superpixels generated by this SLIC superpixel method, which would impact negatively in the classification step, lowering the sensitivity and even increasing the number of false positives per scan. Overall, the nodules with sizes equal or lower than 10mm need to be more studied using the superpixel approach, as they had significantly lower results if compared with the other groups. Analyzing only the last three groups, sensitivities were similar to the state-of-the-art nodule detection approaches previously shown in Table 22.

## 6 CONCLUSION

The lung cancer is the most common type of cancer. As its symptoms are not expressive initially, the cancer is generally diagnosed in later stages. To help improve the diagnostic of lung cancer, CAD systems may be used to aid the radiologist. This work presents two method for lung CT segmentation, Axes' Labeling and Monotone Chain Convex Hull, aiming to reduce the area of processing. The MCCH method achieved an average of 97.78% of nodule preservation and on the average, reduced the CT scan to 15.37% of the original size. The AL method achieved lower results in nodule preservation (93.53%) but on the average, reduced the CT scan to 12.14% of its original size. The two lung segmentation methods achieved interesting results, but it is necessary to improve their performances to not interfere in nodule regions. A data set with lung masks for direct evaluation of the lung segmentation is needed, if a comparison of these methods with other lung segmentation methods is desirable.

For nodule detection, different methods of superpixels were employed to group nodule regions. Since the amount of generated superpixels was too large, a nodule candidate selection approach based on shape index and curvedness was applied to reduce the amount while preserving the superpixels labeled as nodules. Then, a Random Forest classifier is employed for classification of the selected candidates. The best results were found using the FH method with the lungs segmented by the Axes' Labeling method, with a sensitivity of 69.33% and 6.91 FP/scan for the development data set, and a sensitivity of 60.61% and 7.2 FP/scan for the validation data set. Although, the FH technique had better results if compared with the SLIC method, it is clearly necessary the evaluation of the superpixel technique, as many nodules were not correctly segmented, being segmented in many superpixels or included in a large non-superpixel, which would influenced negatively in the candidate selection step. A better segmentation may lead to a more flexible candidate selection based on shape index and curvedness, where only 19 nodules of 838 (2.27%) had no presence of the high shape index windows using the radiologists' annotations. Improvement of our approach in many parts is encouraged, which may contribute positively in the lung segmentation and nodule detection results. Below, some ideas are discussed.

First, in the lung segmentation stage, a combination of both techniques may lead to improvements in the segmentation and further stages. The motive behind this idea is to segment the outer parts of the lungs correctly using the MCCH method and the inner parts (in the central area of the CT scan) with the AL method, since both methods had better segmentation results in their respectively part and worse in the other part.

Analysing the obtained results of the superpixel generation with  $T = 1$ , a 3D segmentation without pre-determination of the superpixel size would better suffice this step.

Some alterations to the FH method for 3D segmentation were made to verify its changes and how future approaches can be directed. In the original work, for a specific pixel, weight is calculated for its 8 neighbors. In a 3D approach, weight can be calculated for its 26 neighbors or less, as to not aggressively merge superpixels. After initial experiments, some problems were seen in this approach. First, superpixels are merged according to the weight between them and a threshold  $K$ . Depending on how the intensity of pixels in a region is, pixels in the same  $(x, y)$  coordinate, but in different slices, are merged, and through the course of the segmentation, superpixels had forms more widely in the Z dimension and, as it was already in the threshold for merging, did not segment correctly some specific region (a nodule, e.g.). Usage of the physical sizes (pixel spacing and slice thickness) may partially fix this, since pixels physically more close (in the same slice) are more likely to be included in the same superpixel first. But the half-done segmentation problem will still be present, cause no change was made to the limit of the segmentation (related to the threshold  $K$ ). Then, the other problem is the threshold value  $K$  for 3D segmentation. Increase of this value may correctly segment nodules with the problem prior stated, but it will increase too the aggressiveness of the merging, thus over-segmenting other parts. A study to fix this problem is necessary. Last, an optimal selection of the  $K$  based on the current CT scan analyzed is encouraged to prevent worse results in virtue of manual selection.

Another possible rule is the analysis of shape index and curvedness relative to the superpixel size, so large superpixels with high shape index windows across its area/volume would not be selected (random lung region), and small ones with concentrated but few would be selected (small nodules). Others related features such as the shape index orientation and its magnitude, which determines the orientation of the shape index and its strength respectively (LARSEN; DAHL; LARSEN, 2015), may present relevant differences between nodule and non-nodule superpixels.

Since the shape index mean and standard variation were in the highest important feature ranked by the mean decrease impurity, a more complex usage of the shape index features may boost classification. One related feature descriptor to consider is the shape index histograms, proposed by Larsen, Vestergaard & Larsen (2014) for cell classification, achieving higher results if compared with other texture descriptors for this type of classification. Application of other discriminant features may improve the results, but a focus in a better segmentation and candidate selection is necessary.

## BIBLIOGRAFY

ACHANTA, R. et al. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 34, n. 11, p. 2274–2282, 2012. Cited in page 17.

ANDREW, A. M. Another efficient algorithm for convex hulls in two dimensions. *Information Processing Letters*, Elsevier, v. 9, n. 5, p. 216–219, 1979. Cited in page 13.

ARMATO, S. G. et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical Physics*, v. 38, n. 2, p. 915, feb 2011. ISSN 00942405. Cited in page 9.

ASHWIN, S. et al. Efficient and reliable lung nodule detection using a neural network based computer aided diagnosis system. In: IEEE. *Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM), 2012 International Conference On*. [S.l.], 2012. p. 135–142. Cited 4 times in pages 29, 31, 38, and 68.

AWAI, K. et al. Pulmonary nodules at chest CT: Effect of computer-aided diagnosis on radiologists' detection performance 1. *Radiology*, Radiological Society of North America, v. 230, n. 2, p. 347–352, 2004. Cited in page 2.

BENJAMINI, Y. Opening the box of a boxplot. *The American Statistician*, Taylor & Francis, v. 42, n. 4, p. 257–262, 1988. Cited in page 55.

BEUTEL, J.; KUNDEL, H.; METTER, R. V. *Handbook of Medical Imaging: Physics and psychophysics*. SPIE Press, 2000. (Handbook of Medical Imaging). ISBN 9780819436214. Available from Internet: <[https://books.google.com.br/books?id=YKVULpCZ\\\_iEC](https://books.google.com.br/books?id=YKVULpCZ\_iEC)>. Cited in page 6.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Cited in page 23.

BROWN, M. S. et al. Toward clinically usable CAD for lung cancer screening with computed tomography. *European radiology*, Springer, v. 24, n. 11, p. 2719–2728, 2014. Cited 3 times in pages 33, 38, and 68.

BUZUG, T. M. *Computed tomography: from photon statistics to modern cone-beam CT*. [S.l.]: Springer Science & Business Media, 2008. Cited in page 7.

CENTER, M.; SIEGEL, R.; JEMAL, A. Global cancer facts & figures. *Atlanta, Georgia: American Cancer Society*, 2011. Cited in page 1.

CHILES, C. Lung cancer screening with low-dose computed tomography. *Radiologic clinics of North America*, Elsevier, v. 52, n. 1, p. 27–46, 2014. Cited in page 2.

CHOI, W.-J.; CHOI, T.-S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Computer methods and programs in biomedicine*, Elsevier, v. 113, n. 1, p. 37–54, 2014. Cited 5 times in pages 33, 34, 38, 48, and 68.

CHRISTE, A. et al. Lung cancer screening with CT: evaluation of radiologists and different computer assisted detection software (CAD) as first and second readers for lung nodule detection at different dose levels. *European journal of radiology*, Elsevier, v. 82, n. 12, p. e873–e878, 2013. Cited in page 8.

CID, Y. D. et al. Efficient and fully automatic segmentation of the lungs in CT volumes. *VISCERAL@ ISBI 2015 VISCERAL Anatomy3 Organ Segmentation Challenge*, p. 31, 2015. Cited in page 30.

COLLINS, L. G. et al. Lung cancer: diagnosis and management. 2007. Cited in page 2.

DEMIR, Ö.; ÇAMURCU, A. Y. Computer-aided detection of lung nodules using outer surface features. *Bio-Medical Materials and Engineering*, IOS Press, v. 26, n. s1, p. 1213–1222, 2015. Cited 3 times in pages 35, 38, and 68.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012. Cited in page 22.

EL-BAZ, A. et al. Computer-aided diagnosis systems for lung cancer: challenges and methodologies. *International journal of biomedical imaging*, Hindawi Publishing Corporation, v. 2013, 2013. Cited 2 times in pages 3 and 29.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Efficient graph-based image segmentation. *International Journal of Computer Vision*, Springer, v. 59, n. 2, p. 167–181, 2004. Cited 2 times in pages 20 and 54.

FERLAY, J. et al. Globocan 2012 v1. 0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11. international agency for research on cancer, lyon, france. 2013. *globocan.iarc.fr*, 2015. Cited in page 1.

FILHO, A. O. de C. et al. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial intelligence in medicine*, Elsevier, v. 60, n. 3, p. 165–177, 2014. Cited 3 times in pages 35, 38, and 68.

FIRMINO, M. et al. Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects. *Biomed Eng Online*, v. 13, p. 1–16, 2014. Cited 2 times in pages 2 and 3.

GALLOWAY, M. M. Texture analysis using gray level run lengths. *Computer graphics and image processing*, Elsevier, v. 4, n. 2, p. 172–179, 1975. Cited in page 27.

GONZALEZ, R. C.; WOODS, R. E. *Digital image processing 3rd edition*. [S.l.]: Pearson Prentice Hall, 2008. Cited 2 times in pages 10 and 11.

HALE, D. Recursive gaussian filters. *CWP-546*, 2006. Cited in page 15.

HAN, H. et al. Fast and adaptive detection of pulmonary nodules in thoracic CT images using a hierarchical vector quantization scheme. *Biomedical and Health Informatics, IEEE Journal of, IEEE*, v. 19, n. 2, p. 648–659, 2015. Cited in page 30.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009. v. 1. 337–387 p. (Springer Series in Statistics, v. 1). ISSN 03436993. ISBN 978-0-387-84857-0. Available from Internet: <<http://www.springerlink.com/index/10.1007/b94608>>. Cited in page 22.



HERBST, R. S.; HEYMACH, J. V.; LIPPMAN, S. M. Lung cancer. *New England Journal of Medicine*, v. 359, n. 13, p. 1367–1380, 2008. PMID: 18815398. Available from Internet: <<http://dx.doi.org/10.1056/NEJMra0802714>>. Cited in page 5.

HU, M.-K. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, IEEE, v. 8, n. 2, p. 179–187, 1962. Cited in page 25.

IGEL, C.; HEIDRICH-MEISNER, V.; GLASMACHERS, T. Shark. *The Journal of Machine Learning Research*, JMLR. org, v. 9, p. 993–996, 2008. Cited in page 28.

INCA. Estimativa 2014: Incidência do câncer no brasil. v. 20, 2014. Available from Internet: <<http://www.inca.gov.br/estimativa/2014/>>. Cited in page 1.

JACOBS, C. et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical image analysis*, Elsevier, v. 18, n. 2, p. 374–384, 2014. Cited 3 times in pages 34, 38, and 68.

JOHNSON, H. J. et al. *The ITK Software Guide*. Fourth. [S.l.], 2015. Updated for ITK version 4.8. Available from Internet: <<http://www.itk.org/ItkSoftwareGuide.pdf>>. Cited in page 28.

KESHANI, M. et al. Lung nodule segmentation and recognition using svm classifier and active contour modeling: A complete intelligent system. *Computers in biology and medicine*, Elsevier, v. 43, n. 4, p. 287–300, 2013. Cited 4 times in pages 30, 32, 38, and 68.

KUHNIGK, J.-M. et al. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE transactions on medical imaging*, IEEE, v. 25, n. 4, p. 417–434, 2006. Cited in page 34.

LARSEN, A. B. L.; DAHL, A. B.; LARSEN, R. Oriented shape index histograms for cell classification. In: SPRINGER. *Scandinavian Conference on Image Analysis*. [S.l.], 2015. p. 16–25. Cited in page 71.

LARSEN, A. B. L.; VESTERGAARD, J. S.; LARSEN, R. Hep-2 cell classification using shape index histograms with donut-shaped spatial pooling. *IEEE transactions on medical imaging*, IEEE, v. 33, n. 7, p. 1573–1580, 2014. Cited in page 71.

LEE, S. L. A.; KOUZANI, A. Z.; HU, E. J. Random forest based lung nodule classification aided by clustering. *Computerized medical imaging and graphics*, Elsevier, v. 34, n. 7, p. 535–542, 2010. Cited 2 times in pages 31 and 38.

LIU, Y. et al. Computer aided detection of lung nodules based on voxel analysis utilizing support vector machines. In: IEEE. *BioMedical Information Engineering, 2009. FBIE 2009. International Conference on Future*. [S.l.], 2009. p. 90–93. Cited in page 29.

MANAY, S.; YEZZI, A. Anti-geometric diffusion for adaptive thresholding and fast segmentation. *IEEE Transactions on Image Processing*, IEEE, v. 12, n. 11, p. 1310–1323, 2003. Cited in page 15.

MESSAY, T.; HARDIE, R. C.; ROGERS, S. K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Medical Image Analysis*, Elsevier, v. 14, n. 3, p. 390–406, 2010. Cited in page 41.

- MURPHY, K. et al. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, Elsevier, v. 13, n. 5, p. 757–770, 2009. Cited in page 35.
- NAPPI, J.; FRIMMEL, H.; YOSHIDA, H. Virtual endoscopic visualization of the colon by shape-scale signatures. *IEEE Transactions on Information Technology in Biomedicine*, IEEE, v. 9, n. 1, p. 120–131, 2005. Cited 2 times in pages 17 and 18.
- NHI. *SEER Training: Bronchi, Bronchial Tree, & Lungs*. 2016. Available from Internet: <<http://training.seer.cancer.gov/anatomy/respiratory/passages/bronchi.html>>. Cited 18 Jul 2016. Cited in page 6.
- NUNZIO, G. D. et al. Automatic lung segmentation in CT images with accurate handling of the hilar region. *Journal of digital imaging*, Springer, v. 24, n. 1, p. 11–27, 2011. Cited in page 29.
- OROZCO, H. M. et al. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *Biomedical engineering online*, BioMed Central Ltd, v. 14, n. 1, p. 9, 2015. Cited 2 times in pages 36 and 38.
- OSHIRO, T. M.; PEREZ, P. S.; BARANAUSKAS, J. A. How many trees in a random forest? In: SPRINGER. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. [S.l.], 2012. p. 154–168. Cited in page 49.
- PARVEEN, S. S.; KAVITHA, C. Detection of lung cancer nodules using automatic region growing method. In: IEEE. *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*. [S.l.], 2013. p. 1–6. Cited in page 29.
- POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Bioinfo Publications, 2011. Cited in page 22.
- PRIDDY, K. L.; KELLER, P. E. *Artificial neural networks: an introduction*. [S.l.]: SPIE Press, 2005. v. 68. Cited in page 28.
- PU, J. et al. Adaptive border marching algorithm: automatic lung segmentation on chest CT images. *Computerized Medical Imaging and Graphics*, Elsevier, v. 32, n. 6, p. 452–462, 2008. Cited in page 41.
- RADIOPAEDIA. *Computed tomography | Radiology Reference Article | Radiopaedia.org*. 2016. Available from Internet: <<http://radiopaedia.org/articles/computed-tomography>>. Cited 30 Jul 2016. Cited in page 7.
- ROY, S.; HERBST, J.; HEYMACH, V. Lung cancer. *N. Engl. J. Med*, v. 359, p. 1367–1380, 2008. Cited in page 1.
- SONG, K. D. et al. Usefulness of the CAD system for detecting pulmonary nodule in real clinical practice. *Korean journal of radiology*, v. 12, n. 2, p. 163–168, 2011. Cited in page 8.
- SPIVAK, M. *A Comprehensive Introduction to Differential Geometry*. 3rd. ed. [S.l.]: Publish or Perish, Inc., University of Tokyo Press, 1999. vol 3. Cited in page 16.

- STRANG, G. Introduction to linear algebra, 5th edition. Wellesley-Cambridge Press, 2016. Cited in page 16.
- SUI, Y.; WEI, Y.; ZHAO, D. Computer-aided lung nodule recognition by svm classifier based on combination of random undersampling and smote. *Computational and mathematical methods in medicine*, Hindawi Publishing Corporation, v. 2015, 2015. Cited 2 times in pages 36 and 38.
- SUNDAY, D. *Convex Hulls*. 2010. Available from Internet: <[http://geomalgorithms.com/a10-\\_hull-1.html](http://geomalgorithms.com/a10-_hull-1.html)>. Cited 30 nov. 2015. Cited in page 14.
- TARIQ, A.; AKRAM, M. U.; JAVED, M. Y. Lung nodule detection in CT images using neuro fuzzy classifier. In: IEEE. *Computational Intelligence in Medical Imaging (CIMI), 2013 IEEE Fourth International Workshop on*. [S.l.], 2013. p. 49–53. Cited 3 times in pages 29, 32, and 38.
- THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. Elsevier Science, 2008. ISBN 9780080949123. Available from Internet: <<https://books.google.com.br/books?id=QgD-3Tcj8DkC>>. Cited 7 times in pages 22, 23, 24, 25, 26, 27, and 28.
- VINEIS, P.; WILD, C. P. Global cancer patterns: causes and prevention. *The Lancet*, Elsevier, v. 383, n. 9916, p. 549–557, 2014. Cited in page 1.
- YE, X. et al. Shape-based computer-aided detection of lung nodules in thoracic CT images. *Biomedical Engineering, IEEE Transactions on*, IEEE, v. 56, n. 7, p. 1810–1820, 2009. Cited 6 times in pages 29, 30, 38, 41, 48, and 68.
- YOSHIDA, H. et al. Computer-aided diagnosis scheme for detection of polyps at CT colonography 1. *Radiographics*, Radiological Society of North America, v. 22, n. 4, p. 963–979, 2002. Cited in page 17.
- YOUNG, I. T.; VLIET, L. J. V. Recursive implementation of the gaussian filter. *Signal processing*, Elsevier, v. 44, n. 2, p. 139–151, 1995. Cited in page 15.
- ZHANG, F. et al. Context curves for classification of lung nodule images. In: IEEE. *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on*. [S.l.], 2013. p. 1–7. Cited 2 times in pages 32 and 38.