



M 2015

U. PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

AUTOMATIC LUNG NODULE CLASSIFICATION IN CHEST COMPUTERIZED TOMOGRAPHY IMAGES

LUÍS DO COUTO GONÇALVES

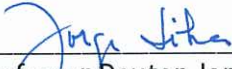
DISSERTAÇÃO DE MESTRADO APRESENTADA
À FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO EM
ENGENHARIA BIOMÉDICA

A Dissertação intitulada

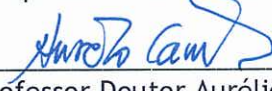
“Automatic lung nodule classification in chest computerized tomography images”

foi aprovada em provas realizadas em 06-07-2015

o júri


Presidente Professor Doutor Jorge Alves da Silva
Professor Auxiliar do Departamento de Engenharia Informática da Faculdade de Engenharia da U. Porto


Professor Doutor Jaime dos Santos Cardoso
Professor Associado do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da U. Porto


Professor Doutor Aurélio Joaquim de Castro Campilho
Professor Catedrático do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da U. Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.


Autor - Luís do Couto Gonçalves

Faculdade de Engenharia da Universidade do Porto

Resumo

O objectivo deste trabalho consiste no desenvolvimento de um sistema de diagnóstico assistido por computador para a classificação de nódulos pulmonares em benigno ou maligno. O cancro do pulmão é o cancro mais letal do mundo com uma taxa global de sobrevivência a 5 anos de apenas 10 a 15 %. Um relativamente pobre diagnóstico numa fase precoce é a principal causa de morte quando se definem as probabilidades da taxa de sucesso na sobrevivência do paciente. A razão por trás disto reside na dificuldade que existe no processo de diagnóstico, onde para haver uma detecção e caracterização precoce destas patologias, os radiologistas têm ser capazes de fazer uma busca exaustiva em todos os exames. Este procedimento é muito demorado e muitas vezes fisicamente exigente, o que pode levar a erros. Engenheiros biomédicos têm, portanto, o objectivo de proporcionar sistemas de diagnóstico assistido por computador, a fim de ajudar e assistir os radiologistas no processo de diagnóstico. Os sistemas CAD envolvem algoritmos baseados em computador desenvolvidos com o objectivo de processar imagens usando a análise de imagem e técnicas de *machine learning*.

Neste trabalho, é apresentado um sistema de diagnóstico assistido por computador para a classificação de nódulos pulmonares em imagens de tomografia computadorizada. Este sistema determina a malignidade de um nódulo usando unicamente informação recolhida a partir da região em torno dos nódulos e foi desenvolvido em dois tipos: 1) Um sistema que proporciona uma classificação de nódulos pulmonares semelhante a uma classificação feita por radiologistas. 2) Um sistema que proporciona uma classificação de nódulos pulmonares semelhante a uma biopsia, cirurgia ou acompanhamento durante alguns anos.

Para esse efeito, alguns estudos foram realizados para otimizar o sistema, incluindo uma análise do banco de dados *Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI)* no que diz respeito ao acordo nas segmentações dos nódulos feitas pelos radiologistas e na análise da gama de intensidades das imagens. Foi também desenvolvido um novo algoritmo de segmentação de nódulos pulmonares.

O estudo mostrou que existe um baixo acordo entre radiologistas no que toca às segmentações e que a gama de intensidades entre nódulos é semelhante. O novo algoritmo de segmentação de nódulos pulmonares apresentou bons resultados quando comparado com as segmentações dos radiologistas.

Finalmente, os dois sistemas foram implementadas e otimizados por meio de um conjunto ideal de características e usando diferentes classificadores. Os resultados mostraram bom desempenho dos sistemas quando utilizados na classificação de dados semelhantes.

Abstract

The aim of the present work is the development of a computer-assisted diagnosis system for lung nodule classification into benign or malignant. Lung cancer is the world's deadliest type of cancer with a 5-year overall survival rate of only 10 to 15 %. A relatively poor early stage diagnosis is the main cause of death when defining the odds for the success of the patient's survival rate and the primary reason for this lies on the difficulty in the diagnosis process, where for early detecting and characterizing these pathologies, the radiologists must be capable of performing an exhaustive search throughout the scans. This procedure is very time consuming and often physically demanding, that may lead to errors. Biomedical engineers have, therefore, the objective of providing Computer-aided Diagnosis (CAD) systems in order to aid and assist radiologists in the diagnostic process. CAD systems involve computer based algorithms designed to process images using image analysis and machine learning techniques.

In the present work, an automatic CAD system for lung nodule classification in CT images is presented. This system determines the malignancy of a nodule using information retrieved solely from the region around the nodules and was designed in two ways: 1) A system that provides a lung nodule classification similar to the radiologists. 2) A system that provides a lung nodule classification similar to a real biopsy, surgery exam or follow up during several years.

For this purpose, some studies were performed to optimize the system, namely, an analysis of the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database, in what concerns the agreement of the radiologists' segmentations and intensity ranges of the images and the development of a novel lung nodule segmentation algorithm. The agreement study showed that the segmentations of radiologists differ greatly from each other, indicating a low agreement, and the range of intensities between nodules is similar. The novel lung nodule segmentation algorithm presented good results when compared with the radiologists' segmentations.

Finally, the two systems were implemented and optimized by using a set of optimal features and different classifiers. The results showed good performance for both when used for classification in similar data.

Agradecimentos

Gostaria de agradecer ao meu orientador, o Professor Aurélio Campilho, pela orientação, interesse e, principalmente, pela disponibilidade e apoio constantes ao longo deste ano.

Gostaria de agradecer ao meu co-orientador, o Ph. D. Jorge Novo Buján, pelo apoio, orientação e disponibilidade que sempre ofereceu, mesmo não podendo estar fisicamente presente. Um agradecimento ao Ph. D. José Rouco Maseda, que sempre se mostrou disponível para esclarecer e ajudar em diversas situações.

Gostaria de deixar também um agradecimento ao INESC TEC e FCT pelo apoio fornecido na elaboração deste trabalho de dissertação.

Aos meus amigos em Coimbra, que sempre me apoiaram e incentivaram ao longo deste desafio. Sem dúvida uma fonte de grande inspiração. Aos meus colegas e amigos, Diogo, Joana Machado, Inês, Gonçalo, Hélder, Joana Silva e tantos outros que sempre estiveram presentes e foram, sem dúvida, de grande importância ao longo destes dois anos e na escrita desta dissertação.

Por fim, um agradecimento aos meus pais e irmão. Agradeço-lhes pelo privilégio que é ter oportunidades, pela pessoa que sou e pela demonstração de apoio e amor incondicional. O seu exemplo é uma fonte de inspiração e uma força, que dão um sabor diferente a este caminho que é a vida.

Luís Gonçalves

*“Our virtues and our failings are inseparable, like force and matter.
When they separate, man is no more.”*

Nikola Tesla

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Objectives and contributions of the dissertation	4
1.3	Resulted publications	5
1.4	Dissertation Overview	5
2	Methods for Lung Nodule Characterization	7
2.1	Lung nodules. Classification and characteristics	7
2.1.1	The pulmonary nodule	7
2.1.2	Radiologic features of the pulmonary nodule	8
2.2	CAD systems for lung nodule classification	11
2.2.1	Pre-processing	11
2.2.2	Feature measurement	12
2.2.3	Feature selection	15
2.2.4	Classification	18
2.2.5	Validation	21
2.3	Concluding remarks	24
3	The LIDC-IRDI Database	25
3.1	Computed tomography images	25
3.2	Database. Overall description	27
3.2.1	Construction and data	28
3.3	Database analysis	31
3.3.1	Agreement of inter-observer segmentations	31
3.3.2	Intensity ranges of the images	35
3.4	Concluding remarks	36
4	Methodology for Lung Nodule Segmentation and Classification	39
4.1	Lung nodule segmentation	39
4.1.1	Nodule segmentation using Hessian matrix	40
4.2	Feature measurement	42
4.2.1	Shape features	42
4.2.2	Intensity features	44

4.2.3	Texture features	45
4.3	Feature selection and classification	49
4.3.1	Feature selection	50
4.3.2	Classification	50
4.4	Validation	50
5	Experimental Results	53
5.1	Construction of the ground truth framework	53
5.1.1	Radiologists' data. Ground truth	54
5.1.2	Diagnostic data. Ground truth	56
5.2	Lung nodule segmentation	56
5.2.1	Agreement analysis between methods' segmentations and radiologists' segmentations	58
5.3	Feature selection and classification	61
5.3.1	Radiologists' data. Procedure and results	62
5.3.2	Diagnostic data. Procedure and results	64
5.3.3	Inter-datasets validation	68
5.4	Evaluation and discussion of results	71
5.4.1	Evaluation of the radiologists' dataset classification	71
5.4.2	Evaluation of the diagnostic dataset classification	73
5.4.3	Evaluation of the inter-datasets classification	76
6	Conclusions and Future Work	79
	Appendices	83
A		85
B		89
C		91

List of Figures

1.1	General scheme of an automatic pulmonary nodule classification system using CAD.	2
2.1	a) Example of a solid nodule. b) Example of a sub-solid nodule. c) Example of a Ground-Glass-Opacity nodule [15].	8
2.2	a) Well-circumscribed nodule. b) Vascularized nodule. c) Pleural-tail nodule. d) Justa-pleural nodule.	10
2.3	General scheme of an automatic pulmonary nodule classification system using CAD.	11
2.4	A classification example taken from [47] and [45].	20
2.5	Non-linear mapping via Φ , converting a nonlinear into a linear decision boundary [47].	21
2.6	Example of a ROC curve. The ideal operating point would be in the upper left corner where sensitivity is 1 and 1-specificity is 0 [51].	23
3.1	The histogram of a 2D CT image.	26
3.2	Example of a CT image [15].	27
3.3	Outlines of the observers in red, blue, yellow and green. (a) and (d) are two small nodules, (b) and (e) are two medium sized nodules, (c) and (f) are two big nodules, (g) and (h) are two sub-solid nodules.	34
3.4	Histograms of the 3D CT images for different cases.	37
3.5	Histograms of nodules included in a 3D ROI.	38
4.1	Nodule segmentation. A - Multiscale Gaussian smoothing using σ of 0.5 to 3.5, step 0.5. B - Eigenvalues computed from the Hessian Matrix. C - Nodule enhancement for each method for every size of σ (we show only the response for $\sigma = 3.5$). D - Maximum response. E - Final mask.	42
5.1	Diagram of the construction of the GT using the radiologists assessment. The input is the classification of four radiologists and the output are three different GT, one for each degree of agreement. m is the array of labels for the degree malignancy and p a weight value.	54
5.2	All stages of the segmentation procedure for a particular σ	57

5.3	Examples of segmentations for both methods, their combination and for the radiologists, presented as contours.	59
5.4	Classification results for the Radiologists' data, presented as AUC (%) value, for 12 features selected by two model searches and six classifiers.	63
5.5	Classification results for the Diagnostic data, presented as AUC (%) value, for 5 features selected by two model searches and six classifiers.	65
5.6	Classification results for the Diagnostic dataset. The columns are the true labels of the nodules, where the left column presents the benign nodules and right column the malignant. The red contours represent nodules classified as malignant and green contours nodules classified as benign.	66
5.7	Classification results for the Diagnostic dataset. Classification results for the Diagnostic dataset. The columns are the true labels of the nodules, where the left column presents the benign nodules and right column the malignant. The red contours represent nodules classified as malignant and green contours nodules classified as benign.	67
5.8	Classification results for the Diagnostic dataset. Classification results for the Diagnostic dataset. The columns are the true labels of the nodules, where the left column presents the benign nodules and right column the malignant. The red contours represent nodules classified as malignant and green contours nodules classified as benign.	68
5.9	Block diagrams of <i>test 1</i> and <i>test 2</i>	69
5.10	Examples of correctly classified nodules from <i>test 1</i>	72
5.14	ROC curves of the classification performances using the Radiologists' data for six classifiers. a) Results presented for 12 features selected by the CFS algorithm. b) Results presented for 12 features selected by the Relief-F algorithm.	72
5.11	Examples of incorrectly classified nodules from <i>test 1</i>	73
5.12	Examples of correctly classified nodules from <i>test 2</i> , Ground Truth 1.	74
5.13	Examples of incorrectly classified nodules from <i>test 2</i> , Ground Truth 1.	75
5.15	ROC curves of the classification performances using the Diagnostic data for six classifiers. a) Results presented for 5 features selected by the CFS algorithm. b) Results presented for 5 features selected by the Relief-F algorithm.	75
A.1	Jaccard results for all small nodules.	85
A.2	Jaccard results for all medium nodules.	86
A.3	Jaccard results for all big nodules.	86
A.4	Jaccard results for all sub-solid nodules.	87

B.1	Number of nodules for each label of Ground Truth 1 versus the labels from radiologists.	89
B.2	Number of nodules for each label of Ground Truth 2 versus the labels from radiologists.	89
B.3	Number of nodules for each label of Ground Truth 3 versus the labels from radiologists.	89
C.1	Bland-Altman results for small nodules. a) Murphy's method. b) Krissian's method. c) Combination of both.	92
C.2	Bland-Altman results for medium sized nodules. a) Murphy's method. b) Krissian's method. c) Combination of both.	93
C.3	Bland-Altman results for large nodules. a) Murphy's method. b) Krissian's method. c) Combination of both.	94

List of Tables

2.1	Tumor stages of the lung nodules.	8
2.2	Radiologic features of the nodules.	9
2.3	Comparison of feature selection methods and classifiers of the most relevant lung nodule classification works, taking into account the number of citations.	16
2.4	Comparison of methods taking into account the number of used nodules in both benign and malignant, and system validation.	23
3.1	Size of nodules presented in LIDC-IDRI database.	28
3.2	Different characteristics and the corresponding degree of appearance used to characterize the nodules by radiologists.	30
3.3	Results for the mean inter-observer agreement in terms of Jaccard's agreement index.	33
4.1	Threshold values.	41
5.1	Weights for the radiologists classification. m is the array of labels for the degree malignancy and p the weight value.	55
5.2	Number of nodules having benign, malignant and indeterminate labels for each of the GTs.	56
5.3	Results for the Bland-Altman study. Mean of the difference between the mean Jaccard value of the radiologists and the mean Jaccard value of the methods segmentations.	61
5.4	Measured features	61
5.5	Selected features using CFS and Relief F for the Radiologists' data and Diagnostic data.	63
5.6	Classification results for the Radiologists' data, presented as the mean and standard deviation of 50 AUC (%) values, for 12 features selected by two model searches and six classifiers.	64
5.7	Selected features using CFS and Relief F for the Radiologists' data and Diagnostic data.	64
5.8	Classification results for the Diagnostic data, presented as the mean and standard deviation of 50 AUC (%) values, for 5 features selected by two model searches and six classifiers.	65

5.9	Confusion matrix of <i>test 1</i>	69
5.10	Confusion matrix of <i>test 2</i> using Ground Truth 1.	70
5.11	Confusion matrix of <i>test 2</i> using Ground Truth 2.	70
5.12	Confusion matrix of <i>test 2</i> using Ground Truth 3.	71
5.13	Performance of the CAD system in form o Sensitivity and Specificity for <i>test 1</i> and <i>test 2</i>	71
5.14	Performance of the CAD system versus the performance of the ra- diologists. Only 19 nodules nodules were used for comparison as they were the only ones to be labelled as non-Indeterminate (1 and 5 labels) by radiologists.	76

List of abbreviations

ANN	Artificial Neural Network
AUC	Area Under Curve
CAD	Computer Aided Diagnosis
CFS	Correlation-based Attribute Subset evaluator
CT	Computed Tomography
CV	Curvedness
GA	Genetic Algorithms
GE	General Electric
GT	Ground Truth
GGO	Ground-Glass Opacity
GLCM	Gray-level Co-occurrence Matrix
GLIH	Gray-level Intensity Histogram
IDRI	Image Database Resource Initiative
LIDC	Lung Image Database Consortium
MHOG	Multioriented Histogram of Oriented Gradients
HU	Hounsfield Units
HRCT	High Resolution Computed Tomography
LAWS	Lattice Aperture Waveform Sets
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
MR8	Maximum Response 8 Filter
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SI	Shape Index
SIFT	Scale Invariant Feature Transform
SPN	Solitary Pulmonary Nodule
SVM	Support Vector Machine
QP	Quadratic Programming

Chapter 1

Introduction

1.1 Motivation

Lung cancer is the world's deadliest type of cancer with a 5-year overall survival rate of only 10 to 15 % [1]. According to the International Agency for Research on Cancer [2], in 2012, approximately 1.8 million new cases (13% of all diagnosed cancers) and 1.6 million related deaths (19.4% of all cases) were accounted all over the world. This represents approximately 20% of all medical cases with lung nodules, as a relatively poor early stage diagnosis is the main cause of death when defining the odds for the success of the patient's survival rate [1]. This is a consequence of two factors:

- Poor screening programs using computerized tomography (CT) - unlike, for example, in prevention of breast cancer, where a mammography exam is performed frequently, chest CT screening is not viable (in fact 20%-30% of the detections come from X-ray exams [3]) because of the patient's radiation uptake and overall cost. Also, to date, there is no conclusive study that demonstrates the advantage in performing a screening program for any risk group regarding neoplasms ¹ of the lung [4].
- Difficulty in the diagnosis process - for early detection and characterization of these pathologies, the radiologists must be capable of performing an exhaustive search throughout the scans. This procedure is very time consuming and often physically demanding, that may lead to errors [5].

Currently, chest CT imaging has shown to be a more sensitive exam for detecting and characterizing lung nodules when compared to projectional radiography, even more so with the advancements in medical imaging technology, namely, in an ever growing number of slices per scan for interpretation and the increase of image quality without further radiation uptake. It has the potential to detect and

¹Abnormal growth of tissue in the lung

evaluate malignant structures in an earlier and potentially more treatable stage. However, with the continuous increase of spatial resolution, more information is available for analysis by the radiologists, increasing fatigue and poor analysis. This includes nodule detection, malignancy assessment and nodule follow-up and management [5] [6].

Biomedical image analysis in chest CT imaging

Biomedical image analysis can prove to be very important in chest CT image diagnosis, namely, in neoplasms of the lung. In this field, biomedical engineers have the objective of providing Computer-aided Diagnosis (CAD) systems in order to aid and assist radiologists in the diagnostic process. CAD systems involve computer based algorithms designed to process images using image analysis and machine learning techniques. A typical CAD system has the block diagram in figure 1.1.

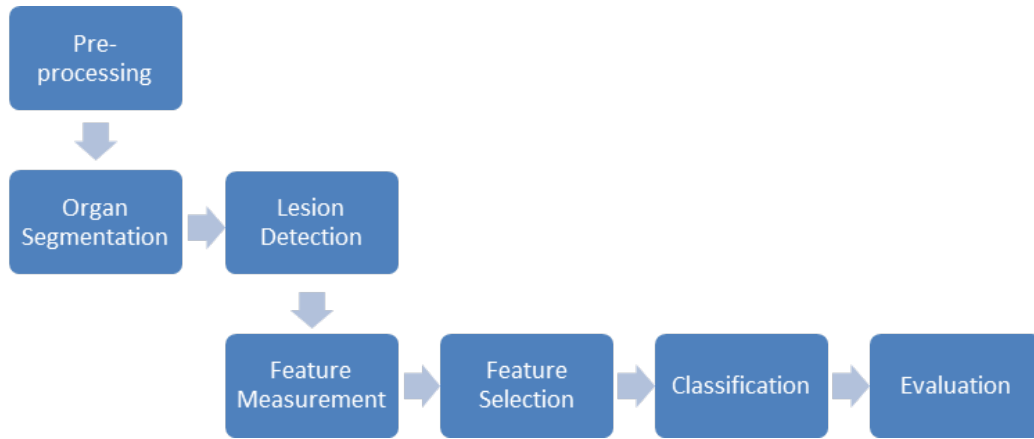


Figure 1.1: General scheme of an automatic pulmonary nodule classification system using CAD.

In the pre-processing stage the images are generally processed for a more efficient computation, noise removal, contrast enhancement for better visualization or normalization.

Though being separated from pre-processing, organ segmentation is also a processing stage and is done to increase computation efficiency, particularly when using a big amount of data. In this stage, the image is reduced to a region of interest (ROI) where all main computation is performed.

The region of interest gives a volume or window around the objects that will be in fact computed and analysed, but the CAD system does not know where they are or which of them are important. It is therefore important to identify the structures to be analysed, process performed in the lesion detection stage.

After detecting the objects the next stage is feature measurement. It is responsible for measuring different features whose objective is to differentiate two or more objects, for example, between nodule regions from non-nodule regions. The

number of possibilities is wide, although all features can be grouped into gray-level properties, gradient information, texture and shape.

Feature selection is an important part of a CAD system when the feature space is large and there are redundant and irrelevant features. It aims to reduce that space without compromising class separability and saving computation time [7], selecting the ones that best represent two or more different objects.

In the next stage, a classifier uses the selected features to build a model that can compute the probability of a certain object of belonging to a certain class. The objective is to correspond, as best as possible, those objects to their correct class with low error, or a low number of incorrect classifications. If the objective is to distinguish between nodule regions from non-nodule regions, for example, than the classifier must be able to save the nodule regions as best as possible without including many non-nodule regions.

Validation is the stage where an evaluation of the system is performed. There are several aspects that can be examined, though the most important is the performance of the classification. This performance is characterized by different measures that include sensitivity and specificity, which give the correct classification rate of true-positives and true-negatives, or accuracy, that presents the overall performance of the algorithm in what concerns correct decisions.

CAD systems proved to benefit the radiologists' performance when used as second readers. In lung nodule classification, for example, one study, [8], compared the radiologists' performance with and without the assistance of a CAD system to assess nodule's malignancy. The results showed that there was an increase of 2% in classification accuracy when using a CAD system. Although the results do not seem to be very significant, they show that these systems can already make some difference in cancer diagnosis.

As the reliability of the CAD systems increases, their demand for clinical application also increases. That is because radiologists seem to be increasingly comfortable using these second readers, which in turn will improve diagnostic performance. Companies like Philips, Siemens or General Electric (GE) introduced some image software tools in the market to perform several functionalities including image analysis and segmentation. Philips [9] developed a visualization software which has the capability of performing lung emphysema quantification, lung nodule detection to improve sensitivity and lung nodule characterization with respect to growth rates and shape. It also performs nodule segmentation. In Siemens [10], the Syngo Lung CAD can perform lung nodule detection, volume and diameter calculation, can measure the average and standard deviation of the nodule's density in Hounsfield (HU) values, the nodule's histogram also in HU and can perform nodule segmentation. GE Healthcare [11] developed a software called Lung VCAR that makes an automatic nodule segmentation and analysis, acquiring the nodule's volume doubling time and the relative growth between two consecutive chest CT images.

Providing these tools to health professionals is important, however, these par-

ticular software packages cannot perform any conclusive analysis for the malignancy of the nodules. The first two perform a CAD analysis, but only for nodule detection, and the third is only a visualization tool with few options for nodule characterization and only provides information for the radiologists' assessment and not an assisted diagnosis. Also, there is no performance specification, like sensitivity versus number of false positives or any type of validation.

It is concluded that CAD systems can be very important to increase the radiologists' accuracy and consistency in nodule detection and characterization. Identifying malignant nodules by image analysis and machine learning can bring new advantages to lung cancer medicine, as the number of invasive procedures performed on benign lesions and the overall health costs are minimized.

The construction of a CAD system for lung nodule classification follows the block diagram seen in figure 1.1, however, it does not have necessarily all those stages. The main problem resides mainly on the segmentation, feature measurement and classification stages, as will be explained further.

1.2 Objectives and contributions of the dissertation

Objectives

The main objective for this dissertation is to develop a robust and accurate automatic CAD system for nodule classification into benign or malignant. This purpose requires a methodology that:

- Identifies all possible nodules in the scene.
- Analyses and performs different strategies for nodule segmentation.
- Extracts and selects the features that best discriminate the nodules.
- Performs classification to distinguish between malignant and benign nodules and analyse the results concerning nodule characteristics.

Contributions

The main contributions of this dissertation are the following:

- Extensive analysis of the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database to reliably build the CAD system.
- Novel lung nodule segmentation approach using an adaptation of Murphy *et al.* [12] and the work from Krissian *et al.* [13].
- Measure and select a complete and extensive group of features that can accurately describe lung nodules.
- A system that performs two different classifications, one similar to the radiologists assessment and another, more reliable, similar to a biopsy, surgery or follow up exam.

1.3 Resulted publications

Conference:

37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy

- Title: Computer Aided Diagnostic for the Classification of Lung Nodules

MVA 2015 IAPR International Conference on Machine Vision Applications, Tokyo, Japan

- Title: 3D lung nodule candidate detection in multiple scales

1.4 Dissertation Overview

The report is divided in six chapters. The motivation and objectives have already been described in this first chapter.

Chapter 2 describes the anatomy of lung nodules, the radiologic features used by radiologists for nodule characterization and the state of the art in CAD systems for nodule classification.

Chapter 3 describes the LIDC-IDRI which is used in this work. It includes the definition of a typical Computed Tomography (CT) image, the properties of the database, an analysis on the agreement between radiologists and an analysis on the intensity ranges of the images.

Chapter 4 presents the methodology for lung nodule classification. It includes the details of all main stages including: the method used for nodule segmentation, the measured features, the method used for feature selection, the proposed classifiers and the evaluation strategy.

Chapter 5 presents the experimental results for nodule segmentation and an analysis on the agreement between that segmentation and the radiologists' segmentations, the analysis on the nodule classification in benign or malignant using different classifiers and sets of features and an evaluation of the system.

Chapter 6 summarizes the conclusions of the research reported in this dissertation and discusses future work.

Chapter 2

Methods for Lung Nodule Characterization

This chapter presents the methods for lung nodule characterization in what concerns the visual analysis of the images to perform a classification based on morphological properties, internal characteristics and the location of the nodules in section 2.1. The current methodologies using CAD systems are also presented in section 2.2.

2.1 Lung nodules. Classification and characteristics

This section presents the description of the pulmonary nodule and the visual properties of the nodules, defined as *Radiologic Features*, used by radiologists to classify the nodules in benign and malignant.

2.1.1 The pulmonary nodule

The pulmonary nodule is a radiologic anomaly, commonly detected incidentally. It is defined as focal, with a round or oval shape, with increased opacity in the lung and a size less than 3 centimetres. Although many of these structures are of benign causes (60%-70%), the ones which represent stage I lung cancers (see table 2.1) must be distinguished by an inexpensive and effective manner [3], using non-invasive imaging techniques and image analysis and pattern recognition methods.

SPNs can be classified into three different groups: solid nodules, which are structures with high-contrast; partially solid or mixed, which are nodules that are both solid and sub-solid; and sub-solid or ground glass opacity (GGO), which have faint contrast and fuzzy margins [14]. Figure 2.1 shows three nodules for each group.

Table 2.1: Tumor stages of the lung nodules.

Stage	Description
I	Tumor ≤ 3 cm diameter without invasion, more proximal than lobar bronchus.
II	Tumor > 3 cm diameter or tumor of any size with any of the following: Visceral pleural invasion, Atelectasis of less than entire lung, Proximal extent at least 2 cm from carina.
III	Tumor of any size that invades any of the following: chest wall, diaphragm, mediastinal pleura, parietal pericardium. Tumor < 2 cm distal to carina.
IV	Tumor of any size that invades any of the following: mediastinum, heart or great vessels, trachea, esophagus, vertebral body, carina. Tumor with malignant pleural or pericardial effusion. Separate tumor nodules in same lobe.

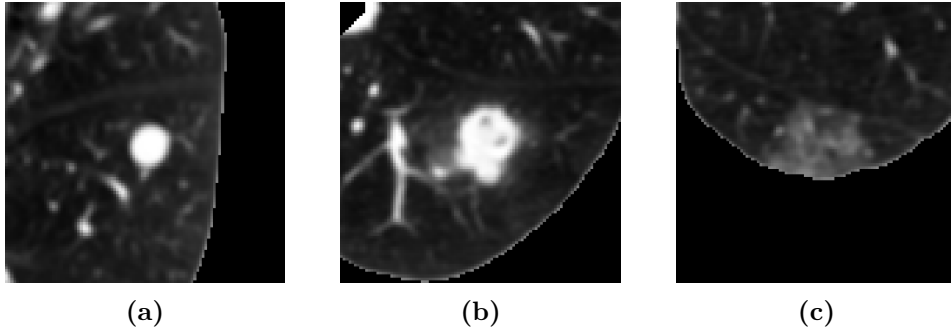


Figure 2.1: a) Example of a solid nodule. b) Example of a sub-solid nodule. c) Example of a Ground-Glass-Opacity nodule [15].

Furthermore, the nodules have different characteristics which help radiologists perform an early diagnostic when assessing if a nodule is benign or malignant. These characteristics will be discussed in the following section.

2.1.2 Radiologic features of the pulmonary nodule

The most common and used properties in lung nodule classification are the shape of the nodule, the volume, its density and calcification. Others may include the central opacity, which happens when the nodule has high intensity in the center when compared to the borders, air component, that is associated with black

Table 2.2: Radiologic features of the nodules.

Feature	Sub-type
Shape [16]	Round Oval Polygonal Complex
Volume [3] [17]	Baseline Volume Baseline Diameter
Density [3]	Homogeneous Heterogeneous
Calcification [3]	Present (Central, Diffuse Solid, Laminated, Popcorn like, Amorphous) Absent
Central Opacity [16]	Present Absent
Air Component [3] [16]	Present Absent
Margin [3] [16]	Smooth Somewhat Smooth Lobulated [17] Slightly Irregular with spiculation Irregular with spiculation
Location [17]	Intraparenchymal Juxtavascular Fissure Attached Pleural Based
Intra-nodular Fat [3]	Present Absent
Cavitation [3]	Present Absent

regions inside the nodule, the margin of the nodule, location, intra-nodular fat, represented in the image by low intensities, and cavitation, which are small, focus, low-attenuation regions within or surrounding the periphery of the nodule [3] [16].

Nevertheless, there are some literature differences between authors in respect to what visual radiologic properties should be used to characterize the nodules and some are more detailed than others. Li *et al* [16] only describes four features and does not include the presence of lobulation in the sub-types features for margin, contrary to Xu *et al* [17]. The study of Xu *et al* [17] also makes reference to location and volume and Erasmus *et al* [3] gives no description about the shape or location. Table 2.2 presents all characteristics addressed by each author for a more comprehensive overview of the nodules' radiologic properties.

Malignant and benign nodules share many of the properties seen in table 2.2 [3].

Nevertheless, some studies [3, 16–18], show that some characteristics are more prominent in malignant nodules than in benign ones and vice versa. For example, Li *et al* [16] performed some statistics based on Fisher exact tests, where, for each feature, it was determined whether there were any significant differences in the proportion of malignant lesions and benign. Xu *et al* [17] used univariate logistic regression analysis to test the relation of each feature to malignancy and benignity.

The results showed that in the GGO type, nodules with a round shape were probably malignant. In the partially solid type, round shaped nodules or with central opacity were more likely to be malignant. In solid, nodules with a complex shape or irregular margins, purely intra-parenchymal, were more likely to be malignant. The presence of small cavitation, air components, or cavitation but with thick, irregular walls, will likely be related to malignancy, such as diffuse and amorphous calcification, when present (attenuation of 200 HU [3]). A big baseline size and a short time required for a nodule to double its volume are also two significant predictors of malignancy.

On the other hand, nodules with a polygonal shape or with a smooth or somewhat smooth margin are more likely to be benign. The presence of intra-nodular fat, with an attenuation of -40 to -120 HU, cavitation, but with thin and smooth walls, and calcification is also a predictor for benignity.

Apart of assessing if a nodule is benign or malignant, it is important to find out what is the nodule’s position in the lungs. This is important since lung nodules that are intra-parenchymal are more likely to be malignant than those who have attached structures like vessels or pleura [17] [19]. Regarding the nodule’s position, the most popular classification is based on Diciotti *et al* [20], which defines the nodules into four types, presented in figure 2.2. Figure 2.2a shows a well-circumscribed nodule, which is located centrally in the lung without any connection to vasculature. An example of a vascularized nodule is presented in figure 2.2b. It is found in the center of the lung but connected to neighbouring vessels. Figure 2.2c presents a pleural-tail nodule, and the name is due to a portion of the nodule being connected to the pleura surface by a thin tail. Finally, a justa-pleural nodule is given in 2.2d. This type of nodules have a large portion of the volume connected to the pleural surface.

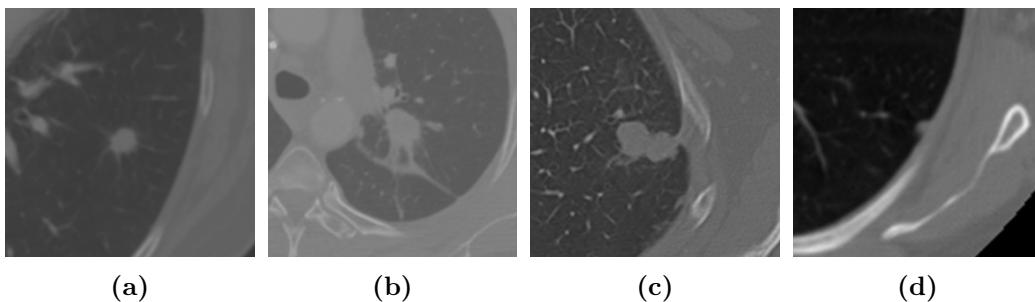


Figure 2.2: a) Well-circumscribed nodule. b) Vascularized nodule. c) Pleural-tail nodule. d) Justa-pleural nodule.

With this type of characterization we will be able to label a nodule by analysing its position and connections, which is somehow more straightforward and conclusive than to analyse multiple features that are common to both malignant and benign. Nevertheless, there are still problems to overcome in *Classification by Location*, because some nodules belong to different types and have similar characteristics. For example, some nodules present shapes that lie between the well-circumscribed and vascularized types (the nodule can have very few vessel connection) or between juxta-pleural and pleural-tailed types (some do not have nor small nor large connection to the pleural wall) [21]. These nodules with intermediate structures complicate the labelling process and increase the overall classification error.

2.2 CAD systems for lung nodule classification

The main objective of Lung Nodule Classification using CAD systems is to differentiate benign from malignant lesions as accurately as possible. Many works have been developed in this area, however, the results have not been fully satisfactory. In this dissertation, only recent publications will be discussed.

Methodologies

A CAD system that is used for differentiating benign from malignant nodules is normally structured in five main steps: pre-processing, feature measurement, feature selection, classification and validation. Figure 2.3 shows the block diagram representing all of these stages.

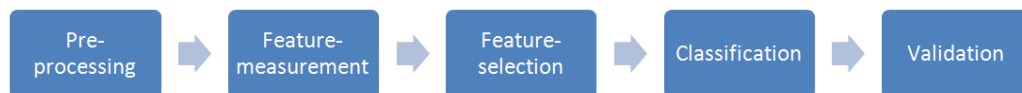


Figure 2.3: General scheme of an automatic pulmonary nodule classification system using CAD.

The following sections describe each one of these stages and the corresponding revised strategies for nodule classification, employed in the recent years.

2.2.1 Pre-processing

In the pre-processing stage the images are prepared for computation in the following stages. The common processing steps in this stage include resizing of the images (reducing or increasing the size of the images), image enhancement, noise removal, segmentation of the ROI, among others. In nodule classification, the pre-processing stage is composed almost entirely by the segmentation of the nodule. Nodule segmentation is important since the objective is to accurately define the region where to measure the best information using an adequate set of features [22].

There are many different ways to perform nodule segmentation. While some authors [23–26] use simple approaches like successive thresholds until the optimal nodule segmentation is achieved, others, [27–29], use region growing or 3D active contours [30]. These techniques can also be combined to refine the segmentation and achieve better results. Yanjie *et al* [27] and Lee *et al* [28] combine region growing with a snake technique and distance transformations, respectively. Ted *et al.* [30], uses a 3D active contour based on k-means clustering with a morphological opening operation. Although nodule segmentation is no trivial task, the ultimate goal must be to achieve an acceptable result that allows a good feature measurement with minimum error.

2.2.2 Feature measurement

The next step is feature measurement. This stage is important because good characteristics lead to good nodule differentiation. In literature we find different sets of features that can be grouped into major categories. The most common are texture, gradient, shape and gray-level or intensity. Some authors, [23] [24] [28], also include radiological and demographic information to analyse their discrimination strength. Information about the nodule’s location is also reviewed in the following paragraphs.

The objective when measuring features is to match, as closely as possible, the radiological characteristics, listed in subsection 2.1.2, with the image analysis characteristics or, in other words, all information that can be acquired using a computer.

In nodule classification, internal characteristics (air component, intra-nodular fat, cavitation, calcification) are very important as many help distinguish benign from malignant nodules. For this purpose, the most helpful and differentiating features are texture features [30] [31]. Gray Level Co-occurrence Matrix (GLCM) is often used [23] [24] [32], as it provides several properties from the spacial dependence of the gray values. Run-length statistics are used in Ted *et al* [30], to analyse the number of runs of a gray level in an image.

Gradient features are also important since the obtained gradient vector and its momentum can help describing the nodule’s boundary sharpness, overall smoothness [33] and density.

Many authors, [25] [30] [32] Yeh, use geometric features as they provide an overall spacial description of the nodules. Some of the most common are volume, diameter and circularity. Many others can be derived from these ones to measure shape and margin properties. This can be achieved by computing the sphericity, second central moment, [25], or performing a *Principal Component Analysis* to measure the main orientation of the nodules in the image space.

Armato *et al* [35] uses gray-level information to compute the overall intensity of the nodule. This feature is important to detect the presence of central opacity in the nodule, for example.

Haifeng, Wang, and Lee *et al.* [23] [24] [28] included radiological features, for distinguishing between benign and malignant nodules. They use them as binary values to indicate if they are present or absent. For example, if there is lobulation then the feature has the value 1, otherwise 0. In their work we are able to find features like: nodule diameter, lymph node status (positive or negative), density (homogeneous or not), solidity (yes or no), ground-glass opacity (yes or no), margin (smooth or speculated), lobulation (present or absent), air inside of the nodule (present or absent), calcification (present or absent), cavitations (present or absent), pleural indentation (present or absent), and pleural effusion (present or absent). The problem when using this information in a CAD system is that the nodules must be previously analysed by a radiologist, thus the measured features obtained by the review are subjective to an erroneous analysis which can mislead the system in classifying the nodules.

Ted *et al* [33] and Huan *et al* [24] used demographic features to assess their effect in the classification. For this, two features were used: age and smoking history in number of cigarette packs per year. However, after analysing the results, the conclusion was that these features did not significantly affect the performance of the CAD system, thus being unhelpful for the differentiation problem.

Another type of feature is location. In the reviewed articles there is no record of this information being used in malignant versus benign classification. However, as stated before, it could prove to be very important in nodule characterization as some nodules seem more likely to be malignant if they are located in the intra-parenchymal region and are well-circumscribed. At this point there are two ways this information can be acquired and used. The procedure can be to use the position of the nodules in the image, or use the information usually computed for determining the nodule's location as features for the malignant versus benign problem. Both have downsides. The first relies on the ability of determining the position of the nodules, thus giving wrong information to the classifier if the nodule is labelled incorrectly. The second can provide redundant information as other features, like texture or gray-level features, can give the same information.

Some work has been done in order to determine the position of the nodules. In the following paragraphs there is a description of the most relevant methodologies in what concerns the construction of the ROI, that is the volume containing the nodule, the methods used for feature measurement and the performance of the systems.

Methods for estimating nodule's location

One of the first to address the estimation of the nodule's location was Faraq *et al* [19] which obtained an average 78.57% correct classification using the Scale Invariant Feature Transform (SIFT). The SIFT descriptor uses several Gaussian filters to identify different scales and locations that can, in a repeatability way, be assigned under different views of the same object. After applying the Gaussian filters, the maximum and minimum values of the filtered image are obtained by

comparing each pixel to its neighbourhood. Gradient orientation and magnitude are calculated for each pixel around the SIFT keypoint location, and stored in a vector. These vectors are then concatenated into oriented histograms representing the final Keypoint descriptor, which is the sum of all gradient magnitudes in that direction and location. The resultant vector has the overall intensity, texture and gradient information of the ROI, being invariant to image translation, scaling, rotation and illumination changes [36].

Because this type of characterization focus on the nodule’s position, Zhang *et al* [37] concludes that including the information from the nodule’s neighbourhood into the classification was vital to capture a more comprehensive description of the lung image. To attain this, a patch-based approach was used, which is performed in order to obtain several square or circular sectors from an initial image. By doing this, much more local detail can be measured as each segment is analysed individually. However, because the information concerns one sub-window, it unavoidably groups unrelated objects. This problem becomes apparent when two or more sub-windows include the nodule and its information is mixed with the surroundings. The approach used by Zhang *et al* [37] solves this due to its adaptive nature. The patches are separated into different levels according to their distance to the nodule’s centroid, being the nodule at the level 0 and the remain neighbourhood at the level 1. This result is achieved by applying the quick shift algorithm [38], which, by mode seeking, constructs a tree of links between a set of superpixels and its neighbours increasing an estimate of the density. Then, using the nodule’s centroid as the center of the image, a circular partition is obtained starting at that position where all pixels, in a given distance, are grouped together. This is done for several distances, K , until the end of the image is reached. The measured feature, for each section, was the foreground ratio which is calculated by counting the pixels belonging to the foreground and dividing it by the total number of pixels in that section. The final result is then concatenated to build the context curve, which is a histogram representing the number of pixels for every section. A SVM classifier was used to conduct the classification and the result had an average classification rate close to 90%, using data ranging from 10 to 90% of the data as training sets.

In 2014, Zhang *et al* [31] used tree different strategies for feature measurement and performed a context (nodule’s neighbourhood) analysis classification.

The same image partition found in Zhang *et al* [37] is used here, however, level 0 is representative of the nodule, level 1 is the nearest neighbourhood and level 2 is the remain neighbourhood.

For feature measurement, density, gradient and intensity information are measured by the use of three methods: a SIFT descriptor; a filter based technique applied to the original image, combined with a rotation-invariant local binary pattern (LBP) analysis, which is applied to the filtered image. The applied filter corresponds to the maximum response 8 (MR8) filter, that uses two anisotropic filters for each of three different scales and two isotropic filters to help discriminate

textures that looks very similar and to achieve rotation invariance; a multiorientation histogram of oriented gradients (MHOG) descriptor, which is a rotation-invariant HOG. This version is able to represent objects by occurrences of gradient orientation, compiled in several histograms for different local proportions in eight different directions. The resultant histograms are concatenated.

This information is then used in two different classification steps. In the first, a classifier is used to estimate the probability of a nodule being of some type. For this, only the information retrieved from level 0 (nodule’s patch) is employed. In the second step, a context analysis information is performed, where levels 1 and 2 give the information. Here, a probabilistic latent semantic analysis, which measured the implicit latent information hidden in the relationship between the images and their categories, with contextual voting is employed. The final stage is designed to combine those two resultant probabilities, thus combining context and nodule analysis. A variable λ (ranging from 0 to 1) is deployed to give more or less weight to nodule probability estimate. With a $\lambda = 0.7$, a 89% correct classification was achieved in Zhang *et al* [37].

2.2.3 Feature selection

Feature selection is an important component of a CAD system when the feature space is large and there are redundant and irrelevant features. Its aim is to reduce that space from a set of N features to a subset of m features, with $m \ll N$, keeping the ones that best characterize the target classification, without compromising class separability and saving computation time [7].

The goal of a selection algorithm is, therefore, to select the features which give a large distance between classes and a small variance within classes. There are numerous processes that perform this selection and each one has its own strategy and use several separability measures.

We encounter three different model searches, each one working in different levels of proximity to the classifiers.

The *Filter* based approach examines the intrinsic properties of the data and calculates the feature relevance score, eliminating the ones who present the lowest results. Although this is a fast computing model, even for high number of features, the main problem is that it is independent from the classifier, thus not working in the hypothesis space.

Wrapper methods work in the hypothesis space searching for the best subset of features. These subsets, obtained from the initial set, are defined by a search procedure and evaluated by training and testing a specific classification model, adapting the approach to a specific classification algorithm [39] [40]. Ideally, the best process would be to analyse every subset obtained by combining all features in the feature set N . That means that for a desired number of features d , from a subset with cardinality j , the procedure would be to examine all $\binom{j}{d}$ possible subsets of the feature set N . This would lead to a very large number of possibilities that would not be practical for a CAD application [41]. Two methods that analyse all

possible subsets are the *Exhaustive Search* and *Branch and Bound Search*. Though they are computationally heavy, the *Branch and Bound Search* uses implicit enumeration to search all subset of features, meaning that it does not searches for every combination of features (branches), but instead ignores branches that have features with low discriminant capacity, making it much faster.

Several, faster, sub-optimal algorithms have been developed to find the best subset of features without searching all subsets. In this group are included deterministic methods like Best Individual and Stepwise Feature Selection algorithms, and non-deterministic or stochastic like the Genetic Algorithms (GA) [41].

Embedded techniques search for optimal subset of features that are included into the classifier, being specific to a given learning algorithm [40] [39]. Like the *Wrapper* methods, they interact with the classification model but are less computationally intensive.

Table 2.3 compiles all feature selection algorithms used by the most cited authors that worked in lung nodule classification. In the reviewed methods, *Wrapper* based are the most used, nowadays.

In the following sections there will be a description of four feature selection methods being used in lung nodule classification. They are two filter based and two wrapper based approaches. The first two were successfully used in recent works and should be further tested to analyse their performance, and the last two are commonly used in research papers.

Table 2.3: Comparison of feature selection methods and classifiers of the most relevant lung nodule classification works, taking into account the number of citations.

	Feature Selection Algorithms	Classifier	Cites (December 2014)
Armato, 2003 [35]	Random Selection and experimentation	Rule-based Classifier and LDA	61
Ted, 2006 [30]	Sequential Forward Selection	LDA	102
Iwano, 2008 [25]	Not used	LDA	24
Ted, 2009 [33]	Stepwise Feature Selection	LDA	38
Lee, 2010 [28]	Genetic Algorithm	LDA	31
Yanjie, 2010 [27]	Genetic Algorithm	SVM	28
Chen, 2010 [32]	Kruskal-Wallis test	3 ANNs	14
Krewer, 2013 [42]	Correlation-based Feature Selection	k-NN	2
Haifeng, 2013 [23]	Penalized logistic regression framework via the lasso-type regularization	ANN	8

Filter based methods

Correlation-based attribute subset evaluator The Correlation-based Attribute Subset evaluator or CFS [43], analyses the strength of a feature in predicting the class of the object, but tends to give little importance to the inter-correlation of the features. It adds features that have high correlation with the class, but only if the set does not have a corresponding high correlated attribute. The CFS’s feature subset evaluation function is given by the following equation:

$$M_b = \frac{n\overline{r_{cf}}}{\sqrt{n + n(n-1)\overline{r_{ff}}}} \quad (2.1)$$

where M_b is an heuristic *merit* of a feature subset b containing n features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in b$), and $\overline{r_{ff}}$ is the average feature-to-feature inter-correlation. The numerator of the equation can be seen as how predictive of the class a set of features are, as the denominator gives information of how much redundant are the features. This equation attributes a ranking on feature subsets in the search space of all feature subsets. [43].

Usually, CFS does not analyse all possible feature subsets in the search space as it is computationally expensive. However, several search strategies can be used to guide the CFS. Some include Sequential Backward Selection or Sequential Forward Selection. These strategies can also be used as *wrapper based methods* if coupled with classifiers. They are described in the section 2.2.3.

Relief F Relief samples instances randomly and checks the distance between them and neighbour instances that have the same and different classes. It calculates the feature weight W by computing the difference between the probability of finding one instance of a different class in the nearest neighbourhood and the probability of finding one instance of the same class in the nearest neighbourhood [43]. W is calculated as follows:

$$W[A] = P(\text{diff. value of } A | \text{nearest inst. from diff. class}) - P(\text{diff. value of } A | \text{nearest inst. from same class}) \quad (2.2)$$

Relief F is an improvement from Relief and is described in detail by Kononenko *et al.* [44]. Contrary to Relief, it searches for k nearest hits and misses from each class, and averages their contributions updating W . This way it is more robust and noise tolerant. It also modulates the probability that the predicted values of two instances are different with the relative distance of those instances. This way, the context sensitivity provided by the “nearest instance” condition is removed and we can rewrite equation 2.2 by recalculating the probabilities P as:

$$P_{diffA} = (\text{diff. value of } A | \text{nearest instances}) \quad (2.3)$$

$$P_{diffC} = P(\text{different prediction} | \text{nearest instances}) \quad (2.4)$$

$$P_{diffC} | P_{diffA} = P(\text{different prediction} | \text{diff. value of } A \text{ and nearest instances}) \quad (2.5)$$

and applying Bayes rule, W is defined as:

$$W[A] = \frac{P_{diffC|diffA} P_{diffC|diffA}}{P_{diffC}} - \frac{(1 - P_{diffC|diffA}) P_{diffA}}{1 - P_{diffC}} \quad (2.6)$$

The number of neighbours k to be checked must be defined, such as the number of instances to be analysed [44].

Wrapper based methods

Stepwise feature selection The basic principle of this type of methods is beginning with a single solution and add or remove one feature at a time until the number of desired features is reached. In the forward methodology, it is added, in each iteration, the feature that improves the model the most. The process ends when the desired number of features is reached. The backward methodology is the opposite, as it starts with all features and, at each iteration, deletes the least significant feature. Although this type of algorithms are very fast, the main problem is that they don not examine all possible subsets, thus is not guaranteed to give the optimal result. As a *wrapper* based method, the Stepwise Feature Selection is coupled with a classifier which calculates the merit of the feature subset.

The genetic algorithm The GA approach mimics the evolution process in biology by representing an arrangement of characteristics as the genotype of an individual that gives it an advantage to survive when competing with other individuals [27]. A feature subset is represented by a binary string of a *chromosome* of length n (total number of available features), with a zero or one in the position of the features being suppressed in a particular evaluation. For each chromosome, is determined its *fitness*, which is estimated by some weight function. If the fitness is high, then the chromosome will *survive* and *breed*, thus creating a new generation. This new generation of chromosomes is created by means of two combinatorial processes:

- crossover - two surviving chromosomes are mixed, with the goal of providing better results.
- mutations - the features index of a single chromosome are randomly altered, depending on a certain probability of mutation, which allows exploring the feature solution space.

After a number of predefined generations, the algorithm yields an acceptable number of solutions to be used in classification [41].

2.2.4 Classification

The next stage in a CAD system is classification. In this work, the purpose of the classification is to distinguish malignant from benign nodules as best as possible using a classifier. It outputs the probability of a certain object, in this case, a benign or malignant nodule, of belonging to a certain class. A classifier uses input data to train a set of rules that are evaluated in the output test data using machine-learning methods [5] [45]. In lung nodule classification, all features obtained from the Feature Selection stage are used as representatives of the nodules and serve as inputs to the classifiers. The purpose is to build a classifier that can be generalized to different data concerning the same problem.

In table 2.3, we identify the classification methods used more recently in lung nodule classification. The most frequently used are the Linear Discriminant Analysis (LDA) or the Artificial Neural Network (ANN), but the one that is showing more promising results is SVM [31] [27] [46]. In [42], the k-nearest neighbours algorithm (k-NN) with $k=5$ was also used with good performance. In the following sections we will present a brief description of two classifiers, the SVM classifier and the k-NN classifier, that are used in our lung nodule classification systems.

Support vector machines

The SVM constructs an hyperplane that best separates two data classes, thus creating a margin that is in both sides of that hyperplane. That margin is the distance between the data points and the hyperplane. Finding the optimal hyperplane that maximizes this margin, is finding the largest distance between each class (figure 2.4), thus increasing the classification accuracy.

The data is linearly separable when a pair (\mathbf{w}, \mathbf{b}) exists in the form of

$$w.x_i + b \geq +1, \quad \text{for } i = 1, \dots, N; \quad x_i \in \text{Class1} \quad (2.7)$$

$$w.x_i + b \leq -1, \quad \text{for } i = 1, \dots, N; \quad x_i \in \text{Class2} \quad (2.8)$$

with a decision rule

$$f_{w,b}(x_i) = \text{sign}(w.x_i + b) \quad (2.9)$$

where w is the weight normal vector to the hyperplane, b the bias or offset value and N the number of data points in the dataset. To find an optimum separating hyperplane, its squared norm must be minimized by a convex quadratic programming problem, where

$$\underset{(w,b)}{\text{argmin}} \Phi(w) = \underset{(w,b)}{\text{argmin}} \frac{1}{2} \|w\|^2 \quad (2.10)$$

subject to

$$y_i(w.x_i + b) \geq 1, i = 1, \dots, N \quad (2.11)$$

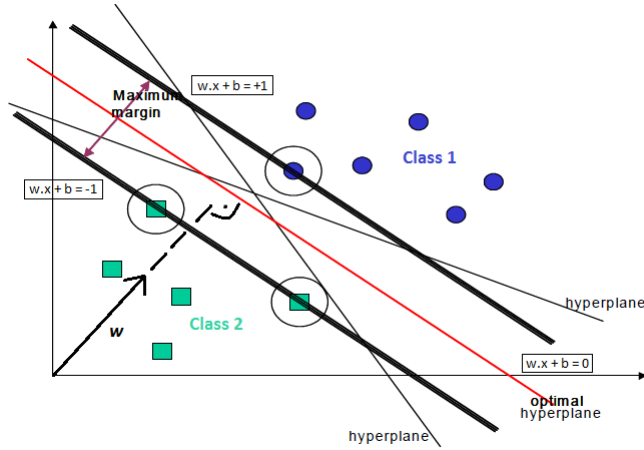


Figure 2.4: A classification example taken from [47] and [45].

When the optimal hyperplane is encountered, data points that lie on its margin are known as support vector points which define a linear solution to the training classifier.

When data contains misclassification instances, the SVM may not be able to find any separating hyperplane. To deal with this problem it is included some level of acceptance on misclassifications of the training instances. This is done by introducing some slack variables ξ_i , $i=1, \dots, N$ in the constraints, thus equations 2.7 and 2.8 become:

$$w \cdot x_i + b \geq +1 - \xi_i \quad \text{for } y_i = +1 \quad \text{and } \xi_i \geq 0 \quad (2.12)$$

$$w \cdot x_i + b \leq -1 + \xi_i \quad \text{for } y_i = -1 \quad \text{and } \xi_i \geq 0 \quad (2.13)$$

and equation 2.10 becomes:

$$\operatorname{argmin} \Phi(w) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (2.14)$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, \dots, l, \xi_i \geq 0 \quad (2.15)$$

where $C > 0$ is a parameter chosen by the user to penalize decision errors. In reality, the ability of separating data does not always exist, which means that an hyperplane cannot be found. To address this problem, the data is mapped into a higher dimension H (see figure 2.5) and try to find a hyperplane there. This higher-dimension is called *Transformed Feature Space*, to distinguish from the original *Input Feature Space*. A linear separation in this transformed space corresponds to a non-linear separation in the *Input Feature Space*. This would

mean that the training algorithm would only depend on dot products in the form $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. This *kernel* K is used to map new points into the feature space for classification. Because there are several *kernel* functions that give different *Feature Spaces*, where the training sets will be classified, it is important to select the most appropriate one. An example is presented in figure 2.5, where different *kernels* K , as the ones in equations 2.16, 2.17 and 2.18.

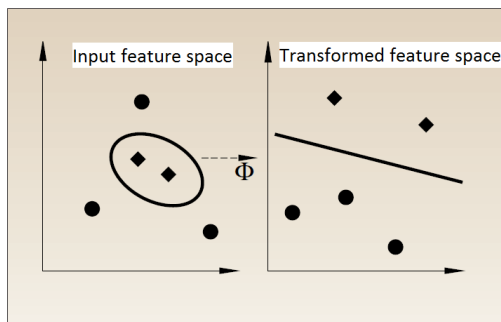


Figure 2.5: Non-linear mapping via Φ , converting a nonlinear into a linear decision boundary [47].

$$K(x, y) = (x \cdot y + 1)^P \quad (2.16)$$

$$K(x, y) = e^{-\|x-y\|^2/(P^2)} \quad (2.17)$$

$$K(x, y) = e^{-(\|x-y\|)/P} \quad (2.18)$$

where x and y are vectors in the *Input Feature Space*, P is the parameter θ of the *kernel* and k is a constant.

The SVM training is finished when it is solved the N^{th} dimensional quadratic programming problem [45] [47].

K-nearest-neighbour

The K-nearest-neighbour, or k-NN, is a non-parametric classification method used in pattern recognition. It is a simple but effective classification method that tries to group samples based on their proximity. Despite its simplicity, it yields good results, mostly when there is little or no prior knowledge about the data distribution.

This classifier is based on a rule of proximity between the starting prototypes and the remaining. The proximity is commonly obtained by computing the Euclidean distance between a test and training samples. If x_i is the input sample with p features $(x_{i1}, x_{i2}, \dots, x_{ip})$, n the total number of input samples ($i=1, 2, \dots, n$) and p the total number of features ($j=1, 2, \dots, p$), then the Euclidean distance between sample x_i and x_l ($l=1, 2, \dots, n$) is defined as:

$$d(x_i - x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \dots + (x_{ip} - x_{lp})^2} \quad (2.19)$$

The decision rule for the prediction of a test sample is set equal to the most frequent class among the k nearest training samples. For instance, if one object is close to two objects of class 1 and one of class 2, for a k value of 3, that object is labelled as class 1 due to the majority of objects in the vicinity [48] [49].

2.2.5 Validation

Validation is the stage where the evaluation of the system is made. There are several aspects that can be examined, and although the most important is the performance of the classification, it is also important to understand what kind of data is used and what is the system's efficiency.

To evaluate the performance of the classification, there are some measures that can be used. Table 2.4 shows the validation methods used by several authors and the corresponding system performances in lung nodule classification. Iwano *et al.* [25] used an evaluation method based on sensitivity and specificity analysis. The accuracy of a system can also be used. Sensitivity and specificity give the correct classification rate of true-positives and true-negatives, respectively, whereas accuracy presents the overall performance of the algorithm in what concerns correct decisions. They are given by the following equations:

$$Sensitivity = \frac{TP}{TP + FN}^1 \quad (2.20)$$

$$Specificity = \frac{TN}{TN + FP}^2 \quad (2.21)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.22)$$

The receiver operating characteristic (ROC) curve, presented in figure 2.6 is the trade-off between true-positives and false-positives rates, and is normally plotted with the sensitivity in the Y-axis and 1-specificity in the X-axis. This trade-off, however, depends on the class we want to make a priority, meaning that we can set the classifier to save more objects from class 1 than class 2 by defining a threshold probability. Normally, this probability is set to 0.5, but to build the ROC curve, all probabilities from 0 to 1 are used.

Almost all of the authors use the Az or area under curve (AUC) value. For comparison purposes, particularly when analysing two classifiers or different systems, it may be the most useful parameter, being a value that is the performance of the classifier when using all threshold probabilities.

By using the Az value, the results can be represented by one parameter that gives the overall discrimination achieved from the system. For comparison purposes, the performance of each methodology is organized in table 2.4. We may

¹TP - True Positives, FN - False Negatives.

²TN - True Negatives, FP - False Positives.

notice that Chen *et al.* [50] based his work on data with ground truth from the radiologists' assessment on the nodule's malignancy, contrary to the others, where the systems were build and evaluated using biopsy or surgically diagnosed nodules. The objective of Chen *et al.* [50] was to build a system that could classify the nodules similarly to radiologists. He compared the system's capacity to guest radiologists. The system achieved an AUC value of 79%, which was higher than the performance of the radiologists.

The remaining authors used nodules with biopsy or surgical confirmed malignancy. It is difficult to say which one presents better results as the number of nodules is different from work to work and, also, as addressed in sub-section 2.2.2, some authors, [23] [24] [28], use clinical features, meaning that these systems depend on the clinical assessment. The best performance is obtained by Haifeng *et al.* [23] that uses a considerable number of nodules and presents a high AUC value.

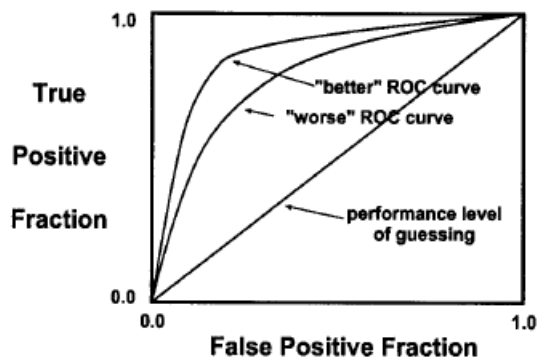


Figure 2.6: Example of a ROC curve. The ideal operating point would be in the upper left corner where sensitivity is 1 and 1-specificity is 0 [51].

When a system is being analysed, it is important to look for high AUC values (ideally, equal to one). However, as seen in table 2.4, these results may be deceiving as some authors use small databases and/or poor nodule variation that may lead to better results. It is important to use a large database with nodules that have different characteristics, namely, in size, opacity, shape and location. By doing this, the authors ensure that the algorithm has good generalization capacity and presents more confidence in classification [51].

Finally, it is also important to examine the system's efficiency. Complex algorithms tend to perform slower than simpler ones and may not translate into better results. It all depends on the strategy employed, which also includes the pre-processing stage. Sub-sampling the images, for example, can greatly reduce the computation time, although decreasing the amount of information.

Table 2.4: Comparison of methods taking into account the number of used nodules in both benign and malignant, and system validation.

	Malignant Nodules	Benign Nodules	Performance
Armato, 2003 [35]	69	401	Az of 79%
Ted, 2006 [30]	44	52	Az of 83%
Iwano, 2008 [25]	52	55	76.9% sensitivity and 80% specificity
Ted, 2009 [33]	124	132	Az of 85.7%
Lee, 2010 [28]	62	63	Az of 88.9%
Yanjie, 2010 [27]	43	34	Az of 87.48%
Chen, 2010 [32]	19	13	Average of 3 classifiers: Az of 79%
Krewer, 2013 [42]	14	19	Accuracy of 90.91%
Haifeng, 2013 [23]	116	86	Az of 91%

2.3 Concluding remarks

Image analysis using CAD systems is becoming increasingly important in diagnosis when used as a second reader. Although their performance is not yet truly satisfactory, it has already been shown that it can be useful to increase the overall classification performance. As stated before, the ability of differentiating benign from malignant nodules is very difficult due to lack of distinguishable characteristics. These systems, however, have the capability of using different features, which combined can result in a reliable classification. There are many characteristics that can be measured and it is important to use different techniques to obtain a good nodule representation.

Many authors consider texture as the most powerful feature for nodule classification, so it is important to consider the various approaches used to measure it. The GLCM seems to be a good tool for this purpose. Also, it is clear that the SIFT descriptor gives important features as it is used by many, with good results. The combination of SIFT with LBP and HOG, increased the overall performance in Zhang *et al* [31], indicating that further texture and gradient analysis is necessary. Shape analysis is also necessary since it was noticed that shape features can have good differentiation capability. Another interesting fact is that including the analysis of the nodule's neighbourhood also helped differentiating the nodules.

After measuring the features, it is important to eliminate the ones that are redundant. A good wrapper or filter based algorithm for feature selection must be employed to find a subset of features that is closely approximated to the optimal solution.

Finally, a good classifier must be employed to provide good class differentiation. The SVM is a powerful classifier and, according to Zhang *et al* [31] and Yanjie *et al* [27] is the one that shows more promising results.

Chapter 3

The LIDC-IRDI Database

The LIDC-IDRI database contains the Computed Tomography (CT) images used in this work. The following sections describe a typical CT image and properties and the general properties of the database. A database analysis is also presented in what concerns the radiologist's annotations and the intensity ranges of the images.

3.1 Computed tomography images

The most commonly used 3D medical imaging is X-ray CT. This modality uses X-ray beams to produce an image, taking advantages of the physical processes of beam generation, dispersion, absorption, and scatter. However, unlike conventional projectional radiography, the CT-image is not acquired in a single, complete form, as it is not projected in a standard X-ray film. The CT-scanner is composed by a moving bed to translate the patient through the scanner, a rotating X-ray tube to emit photons, a collimator, which is a part composed by several pairs of lead plates, to form a flat beam and an array of X-ray detectors to absorb the photons. The simultaneous movement of rotation and translation allows the formation of multiple slices that together compose a 3D image. This process takes usually a few seconds (1 to 5 seconds) to complete.

The resolution plays an important role when acquiring the images. Typically, the sampling resolution in the transaxial plane is either 256×256 or 512×512 voxels square and in the longitudinal direction is dependant on the number of slices acquired. The sampling resolution of the voxels can also be defined and it is dependent on the field of view and pixel dimensions. The spacial resolution corresponds to the vertical width of the voxel and it is limited by the physical properties of the scanner, as it is dependant on the dimensions of the collimator. In theory, the narrower the collimation, the better. However, this requires extremely amounts of X-ray flux to image, so the physical limit is set to 1 mm.

Thin-section CT scanning or High Resolution Computed Tomography (HRCT) is used when the goal is to diagnose potential diseases that are difficult to assess when using conventional CT imaging. In lung HRCT they can include pulmonary

fibrosis, emphysema or bronchiectasis. Although this modality is performed using a conventional CT scanner, it uses different imaging parameters to maximize spatial resolution.

The technical features of the HRCT normally include a thin collimator of 1 mm and the use of a high spatial frequency algorithm. Also, the spatial resolution of modern scanners can reach up to 0.25 mm in the longitudinal direction. However, the slices are normally acquired over broader distances (for every few centimetres or dozens of millimetres). This is performed as HRCT delivers 10-20 times more radiation than conventional CT screening, so it is important to reduce the patient's exposure [52].

A CT image has different levels of gray values along the voxels which form the objects. These levels correspond to different number of photons arriving to the detectors, which depends on the density of the tissues of the patient. Bones are denser so they absorb more photons when compared to soft tissue, for example.

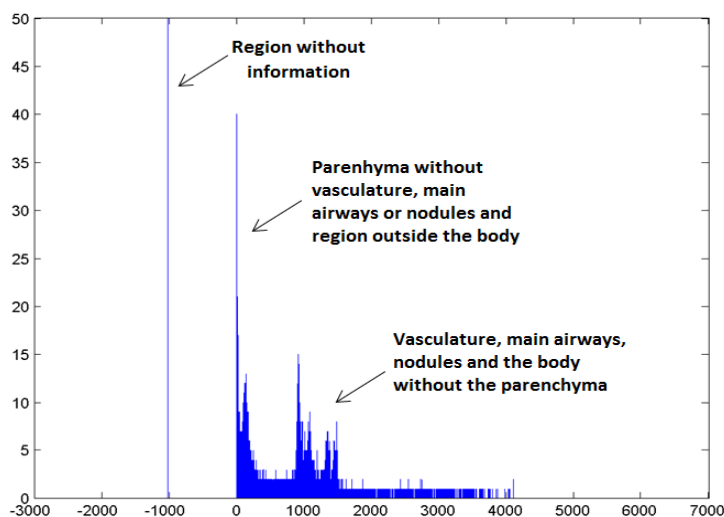


Figure 3.1: The histogram of a 2D CT image.

In CT imaging, the values of attenuation are measured in Hounsfield units, named after the inventor of CT scan, Godfrey Hounsfield. The scale is calibrated using the attenuation coefficient of the water, 0 HU. The lowest value is air, measuring -1000 HU, then comes fat tissue, that ranges between -300 and -100 HU, muscle tissue 10-70 HU, and bone above 200 HU [53]. Figure 3.1 shows the histogram of a chest CT image and it is visible the intensity distribution of the lungs and region outside the body (air), body and the region without information.

As stated previously, different tissues have different levels of attenuation, so chest CT images also present the various structures of the lung with different HU values. The main structures of the lung are: the bronchi, parenchyma, bronchus-associated lymphoid, vasculature, thorax, diaphragm and mediastinum [54]. Figure 3.2 shows an example of a chest CT image. The most visible structures are the

bronchi, thorax and diaphragm because of their volume and higher attenuation capacity. The parenchyma has low attenuation values, so, normally, some intra-parenchymal vasculature can be seen, such as lung nodules or other pathologies.

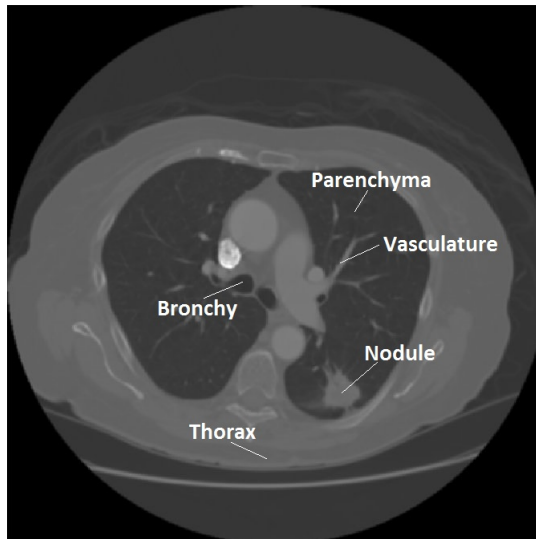


Figure 3.2: Example of a CT image [15].

3.2 Database. Overall description

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions. It is an international resource started by National Cancer Institute, further advanced by the Foundation for the National Institutes of Health, and accompanied by the Food and Drug Administration. It is available via Internet for development, training, and evaluation of CAD systems for lung cancer detection and diagnosis [15]. Specifically, the LIDC-IDRI initiative aims are to provide:

- A reference database for the relative evaluation of image processing or CAD algorithms.
- A flexible query system that provides researchers opportunities to evaluate a wide range of technical parameters and identified clinical information within this database that may be important for research applications.

Importance of LIDC-IDRI database

This database has proven to be very useful in validating system performance. Its size, its variability in nodule type and great discrimination allied with the

fact of being available to public, gives scientists a splendid resource and helps improving their area of research. This database has already been used by several authors, namely, in lung segmentation [55], nodule segmentation [46, 56–58], nodule detection [59–61], nodule characterization [42, 62–64] and nodule’s subtlety assessment [65].

3.2.1 Construction and data

According to Armato *et al.* [66], seven academic institutions participated in the database construction by providing all scans, but only five participated in the interpretation process. From these institutions, a total of 12 radiologists/observers performed the image annotation process, resorting to three different segmentation software tools, including nodules segmentations and the nodules characteristics, presented in a XML file, described bellow. One of the software tools used a semi-automated process to create the outlines, while the others were fully manual. Although the images were presented with a standard brightness/contrast, the radiologists were allowed to adjust those properties and the magnification to enable a more comprehensive interpretation of the scan.

The LIDC-IDRI contains a total of 243958 thoracic CT scan images from 1018 patients. From those 1018 cases, 311 have enhanced images with a minimum, maximum and medium number of slices of 81, 501 and 157, respectively, 704 have unenhanced images with a minimum, maximum and medium number of slices of 65, 764 and 279, respectively. All images have a sampling resolution of 512×512 . The spacial resolution, in both transaxial plane and longitudinal axis, varies from scan to scan as the images were obtained from different machines.

Each annotation is made independently by four radiologists and each one of them review the annotations of the others after the first assessment. Then, the reviews are returned to the corresponding radiologists and are revised, this time knowing the opinion of his pairs. The database has 7371 lesions marked as nodule by at least one radiologist. 2669 of these lesions are bigger or equal to 3 mm which is the main focus of clinical practice and CAD research. It is visible in table 3.1 the range of diameters and volumes contained in LIDC-IDRI [66].

XML file

Each LIDC-IDRI case includes an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists at a time. It contains several code lines representing the radiologists’ identification, the nodule’s identification, its characteristics, the coordinates of its position and marked contours. The annotation only starts at reading section. A reading session consists of a set of predefined objective markings created to homogeneously characterize all different lesions.

In a reading session, if a nodule is detected, the radiologist indicates the version of the annotation and his identification. If the nodule is bigger than 3 mm an identification number is also assigned for that nodule. The nodules are then defined

Table 3.1: Size of nodules presented in LIDC-IDRI database.

	Mean	Minimum	Maximum
Diameter (mm)	7.24	2.03	68.43
Volume (mm^3)	93.08	4.39	1667789.82

as in table 3.2 which presents the characteristics and corresponding degrees used by the radiologists to describe the nodules. They include Internal Structure, Calcification, Sphericity, Margin, Lobulation, Spiculation, Texture and Malignancy. Additionally, the radiologist registers the localization, which is the central position of the nodule, and contour, done for every slice where the nodule is represented, and writes its file name. The coordinates are written as X and Y coordinates and the corresponding slice Z .

Diagnosis data file

The database also presents a small number of cases with known diagnosis. These nodules were evaluated by one of the following methods: review of radiological images to show 2 years of stable nodule, biopsy, surgical resection and progression. A *.xls* file containing all the information can be acquired in [67]. It contains a total of 157 cases with the corresponding malignancy and origin (some nodules are metastatic) for one to two nodules per case. The nodules are labelled as Benign or Non-malignant disease, Malignant (primary lung cancer), or Malignant (metastatic).

Table 3.2: Different characteristics and the corresponding degree of appearance used to characterize the nodules by radiologists.

		Degree					
		1	2	3	4	5	6
Type	<i>Subtlety</i>	Extremely subtle	Very subtle	Subtle	Relatively obvious	Obvious	-
	<i>Internal Structure</i>	Soft Tissue	Fluid	Fat	Air	-	-
	<i>Calcification</i>	Popcorn	Laminated	Solid	Non-Central	Central	Absent
	<i>Sphericity</i>	Linear		Ovoid		Round	-
	<i>Margin</i>	Poorly Defined				Sharp	-
	<i>Lobulation</i>	No lobulation				Marked	-
	<i>Spiculation</i>	No Spiculation				Marked	-
	<i>Texture</i>	Non-Solid		Part Solid		Solid Texture	-
	<i>Malignancy</i>	Highly Unlikely	Moderately Unlikely	Indeterminate	Moderately Suspicious	Highly Suspicious	-

3.3 Database analysis

To fully understand and to reliably use the database, the following sub-sections present an analysis of the data concerning the inter-agreement of the radiologists' segmentations and, additionally, the intensity ranges of the images, whose acquisition process has great variability.

3.3.1 Agreement of inter-observer segmentations

The method for building this database led to several limitations, though many of them have already been mentioned in [66]. Some include the inability to perform reader studies due to the continual change of the radiologists' identification and the fact that the annotation is not performed by all same four radiologists. Also, the segmentations performed by the radiologists, saved in the *XML* file, can diverse greatly from each other. Despite these limitations, there is still a need to analyse the nodules segmentations created by the radiologists by performing a inter-agreement study to determine the overall agreement of the observers. The objective is to assess if it is important to use developed segmentation tools and compare their efficiency to the nodule masks obtained from the segmentations of the database. In the following section we discussed the method used for the agreement analysis and corresponding results.

Methods for estimating the radiologists agreement

To assess the overall agreement in the database, some performance measures are used to analyse the divergence between the segmentations created by the radiologists. In several studies, [68–71], the Jaccard's index is used to compare segmentations, while in others, [72] [73], is the Dice's index. Although these measures are similar, others were also implemented by the authors to support their analysis like: sensitivity, specificity, conformity, percentage of area difference, Hausdorff distance and Williams' index. All, but one, of these studies were conducted to assess the performance of a particular segmentation tool, having as ground truth a manual or semi-automatic segmentation performed by one or several specialists. There was one study, [68], whose goal was to analyse the inter-observer variability in segmenting the left ventricle of the heart using a given segmentation tool, but did not have a ground truth. Two others, [69] [71], also analysed the inter-observer segmentation variability of their ground truth, the first for pulmonary sub-solid nodules and the other in Intravascular ultrasound.

Unlike the studies mentioned above, that assess the agreement between the same observers, which in turn are properly identified, this study is designed in a way so that an overall agreement analysis is obtained. Additionally, the goal is to analyse the average difference between observers for the segmentation of small, medium and big solid nodules (texture value of 4 and 5), and, separately, sub-solid nodules (texture value of 1 and 2). This is done so there is an understanding of which nodules are more difficult to outline. Only nodules that have an agreement

between all four radiologists are accounted for analysis. Altogether, 599 nodules are used in this study, from which 336 are small nodules, 90 are medium sized nodules, 143 are big nodules and 30 are sub-solid.

Agreement measure

Due to the similarities between this study and the studies of Silva *et al.* [68], Lassen *et al.* [69] and Balocco *et al.* [71], the agreement measure used here is the Jaccard's index. If the Jaccard's index is close to 1, a good agreement is present [68] [69]. The equation for calculating this index is given by equation 3.1,

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \quad (3.1)$$

where, A and B represent the two areas of segmentation obtained from the outlines traced by the radiologists.

Agreement analysis between radiologists

The masks of the nodules were first obtained by filling the region inside the outlines. This procedure was done in 2D slices and then concatenated to build a 3D mask. Each mask was compared, individually, with the remaining 3 and an average Jaccard value was calculated using the resulting three Jaccard values so a single agreement value per segmentation is obtained.

To provide a better understanding of the results, all data is presented in two ways. The graphic representation of the data is located in Appendix A in figures A.1 to A.4. They are calculated for each of the nodules and observers. Table 3.3 presents the overall Jaccard variability, that is the mean, standard deviation and median values for solid nodules with three different sizes and GGOs. The mean value agreement was calculated by averaging all Jaccard values of the segmentations and the standard deviation and median values were also obtained from all those values. The reason why the mean, standard deviation and median Jaccard values for the radiologists segmentations are used is because the unknown identity of the radiologists for different cases makes it impossible to link one annotation to another, so this is the only way to compare the agreement for different nodules.

The results from solid nodules suggest that there is a low level of agreement between observers. Looking only to figures A.1-A.4, it is visible that there is a great dispersion, with great variation between values. It is also visible that all values stay below 0.9 and only a few are close to that value.

Now, looking to table 3.3, it is conclusive that the agreement between observers is low. The highest mean Jaccard's value arises from segmentations of big nodules and it is only 71%. The fact that this type of nodules present the highest value can be surprising, as many are difficult to outline due to their connection to vasculature and other structures. However, due to the fact that the majority of their volume is clear, it can be assumed that all radiologists equally outlined most of the nodule, as visible in figures 3.3(c)(f). The low agreement probably arises from places where

the nodules are connected to other structures, as seen in figure 3.3(f), meaning that some radiologists included parts of those structures and others did not.

Table 3.3: Results for the mean inter-observer agreement in terms of Jaccard’s agreement index.

Nodules	Volume (mm^3)	Mean	Median
<i>Small</i>	<500	0.64 ± 0.10	0.65
<i>Medium</i>	$\geq 500-1000$	0.68 ± 0.10	0.69
<i>Big</i>	>1000	0.71 ± 0.10	0.73
<i>Sub-solid</i>	all sizes	0.62 ± 0.09	0.62

The lowest mean Jaccard’s value for solid nodules is 64% and is found on the small sized ones. The reason behind this can derive from the fact that, although these nodules are solid, their margins are not well defined, so the segmentations can diverge greatly between observers. Because they have a small size, what appears to be small variations, can actually translate into segmentations that occupy a much bigger percentage of volume, resulting in a low agreement.

For medium sized solid nodules, the mean Jaccard’s value is 68%. This result was expected giving the results obtained for small and big nodules, as it stays between the values of those sizes. The fact that their boundaries are relatively well defined and have a significant size, makes the agreement between observers higher when comparing to small nodules. However, they are still small, thus presenting the same outlining problems as of small nodules. Figure 3.3(b) presents a good example of outlines for medium sized nodules. It is visible that its shape is round and its boundaries are somehow well defined. However, the outlines differ greatly which indicates low agreement between radiologists. In another example, 3.3(e), is it visible that there is more agreement, with only one radiologist (red outline) veering from the others.

As expected, due to the subtle appearance in the CT images and poor defined boundaries, the lowest agreement between observers occurs in sub-solid nodules. Although all nodules in this group have different sizes (small, medium and big), the mean Jaccard’s index stays at 62%. Figures 3.3(g)(h) are two examples of sub-solid outlines, where 3.3(g) shows good agreement between observers (the outlines are close to each other), and 3.3(h) shows poor agreement, particularly for the red marking.

Analysing the mean Jaccard value gives good insight of what level of agreement is present in the database, however, the standard deviation provides an instructive understanding of the degree of variability. As visible in table 3.3, the standard deviation in Jaccard’s varies from $\pm 9-10\%$, indicating that the agreement between observers in different nodules has a significant variation.

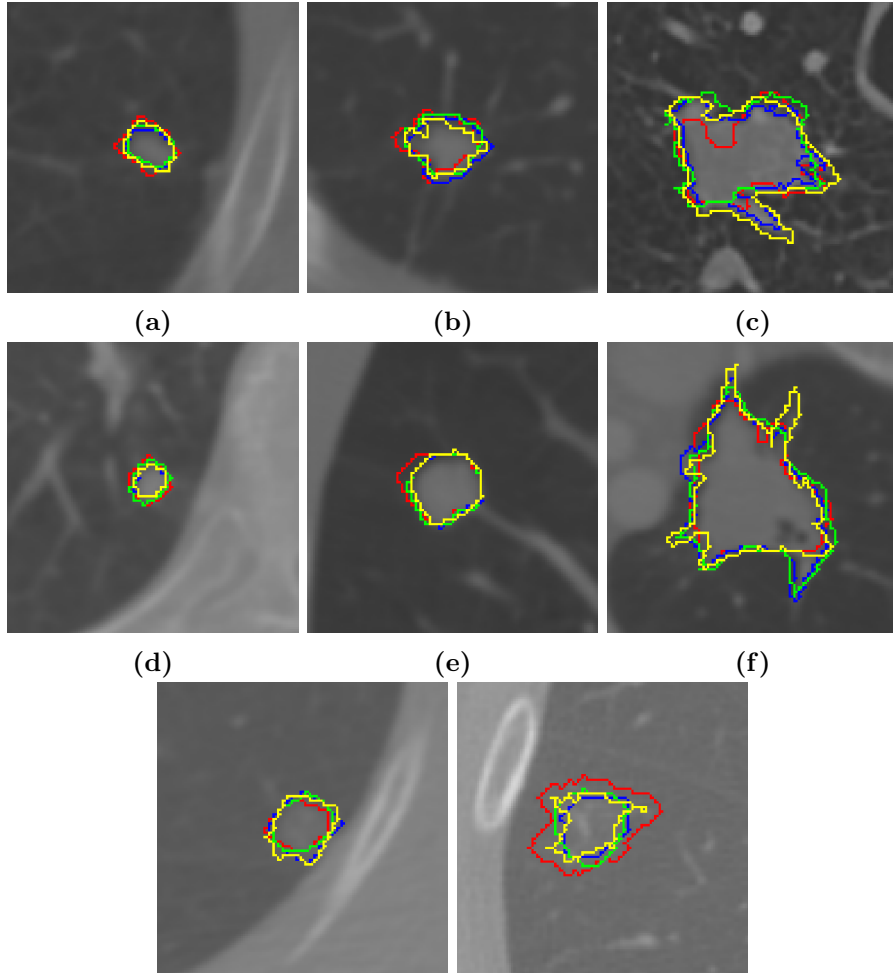


Figure 3.3: Outlines of the observers in red, blue, yellow and green. (a) and (d) are two small nodules, (b) and (e) are two medium sized nodules, (c) and (f) are two big nodules, (g) and (h) are two sub-solid nodules.

3.3.2 Intensity ranges of the images

The LIDC-IDRI images present a variety of intensity ranges from -1024 to 32767 HU. Depending on the type of acquisition method, the ranges can be significantly different, which is a result of using different CT scanners and KeV values that affect the attenuation coefficient of the tissues due to the physical properties of the X-Rays produced by one specific scanner [74]. As the HU values are obtained using these attenuation coefficients, they also diverge from image to image.

As the objective of this work is to label nodules into benign and malignant, and thus intensity properties give crucial information to achieve a good accuracy, it is of the most importance to perform a study that analyses the ranges in intensity of the images and, consequently, of the nodules. Normally, the images are normalized to obtain a coherent and homogeneous measure of the image properties, but in this case, as the image intensity ranges are very different, this normalization can be harmful for the classification step. The objective is, therefore, to know if the intensity ranges between the images and nodules are in fact significantly different and, if so, how much.

Intensity range of the nodule

In order to perform an analysis on the images and nodule's range of intensity, a set of 5 different scans were chosen, each one having large nodules so more information could be acquired. These images represent each of the type of images in the database. The set of 5 images is divided in two main groups, enhanced and unenhanced. The first group has two images, where the first ranges from -1024 to 4095 HU and the second from -1024 to 6916 HU. The second group has three images, where the first ranges from -1024 to 4095 HU, the second from -2000 to 4095 HU and the third from -12209 to 32767 HU, this last one being chosen due to its singularity in intensity range. A window of size $41 \times 41 \times Z$ was chosen to comprehend the nodules and an intensity normalization (0-1) was performed for each image to better visualize the ranges discrepancy. Z is the length of the nodule in the z direction. 2D images (an example slice with the nodule) and the corresponding 3D histograms are represented in figures 3.4 and 3.5.

The histograms of each image are represented in the same scale so it is visible the effect of the normalization when comparing to the other images. As visible in figure 3.4(a), the main mode resides in the same range for every image and is where the important information is located. All images have a high number of counts in the lower intensity, which represents the black area where no information is present. Images in the first, third and, less prevalent, fourth row, also have a high counting in the upper limit. This comes from high density material inside the body like bones. In the second row, we find an image that spreads these high intensity values through a high intensity range instead of confining it to a single value. The last image, fifth row, is the most intriguing. Its upper and lower limits are way disproportional for what is common in CT scanning. Like in the other images, it presents a high peak around -1000 HU, but lower intensity values

continue to appear after that. The same happens as in the image of the second row for the upper limit, but this time the limit reaches the value of 32767 HU which is far away above any other. Naturally, the images that show broader HU intensity ranges present a narrower normalization range and the inverse happens to the ones presenting narrower HU intensity ranges. This is visible in figures 3.4(b). Also, the normalized ranges now depart from each other. The result of the normalization can be visualised in figures 3.4(c), where the broader ranged images are darker and fuzzy and the others more contrasted.

Similar comments can be given for the nodules. Their intensity values are located inside the main peak of the images, thus the range of intensity HU values is the same for every nodule. This similar behaviour was observed for other images proving that the intensity of the nodules does not differ between images, thus feature measurement, in what intensity and texture information concerns, can be performed in the raw non-normalized images, so a good nodule presentation is achieved.

3.4 Concluding remarks

The LIDC-IDRI database can be very useful in validating different system performances. Its size, its variability in nodule type and good discrimination, allied with the fact of being available to public, gives scientists a splendid resource to study this medical field. However, some problems, addressed in this report, were noticed when analysing its data, although they are also mentioned in [66] by the creators. The study is conclusive in what concerns the overall agreement between observers. The higher mean Jaccard's index is obtained for big nodule. Globally the agreement is 71%.

Some reasons may point to why the overall agreement is low. One can be the use of three software tools to outline the nodules, one semi-automatic and two manual. Giving the size of the database, it is natural to assume that the difficult task of analysing every case resulted in fatigue and consequently, sometimes, in poor accuracy. The principal reason, however, can be the fact that the database was analysed by 12 radiologists. If it is assumed that the experience between each of them is different, the results can also vary between each other.

The analysis on the intensity range of the images shows that there is a great diversity between them and that they must be treated individually. Although this is true for the global range of the images, the intensity range of the nodules is the same for every case indicating that the images must be used in the original intensity values, namely when performing feature measurement.

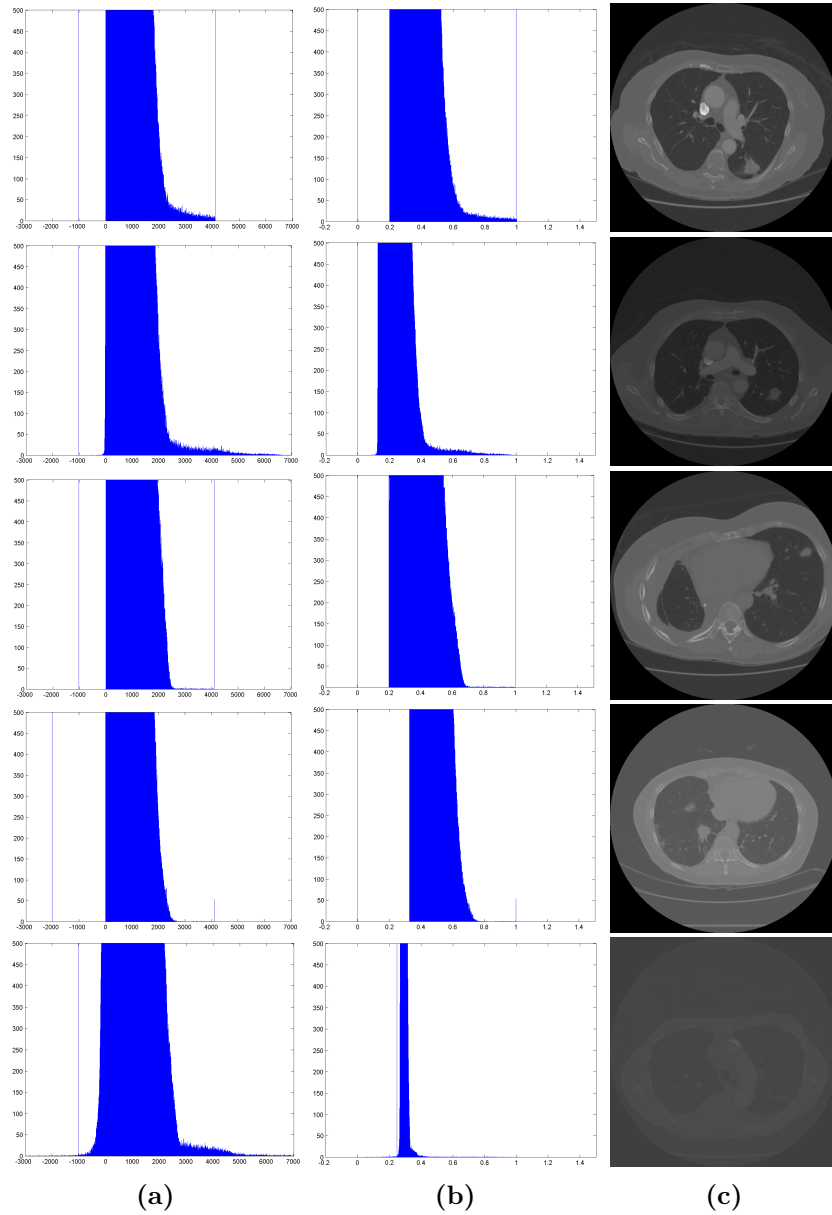


Figure 3.4: Histograms of the 3D CT images for different cases. (a) Image histograms, (b) Normalized histograms, (c) Single slices for each case. First and second rows represent two enhanced cases and third and fourth rows two unenhanced. Fifth row represents a singularity in the database.

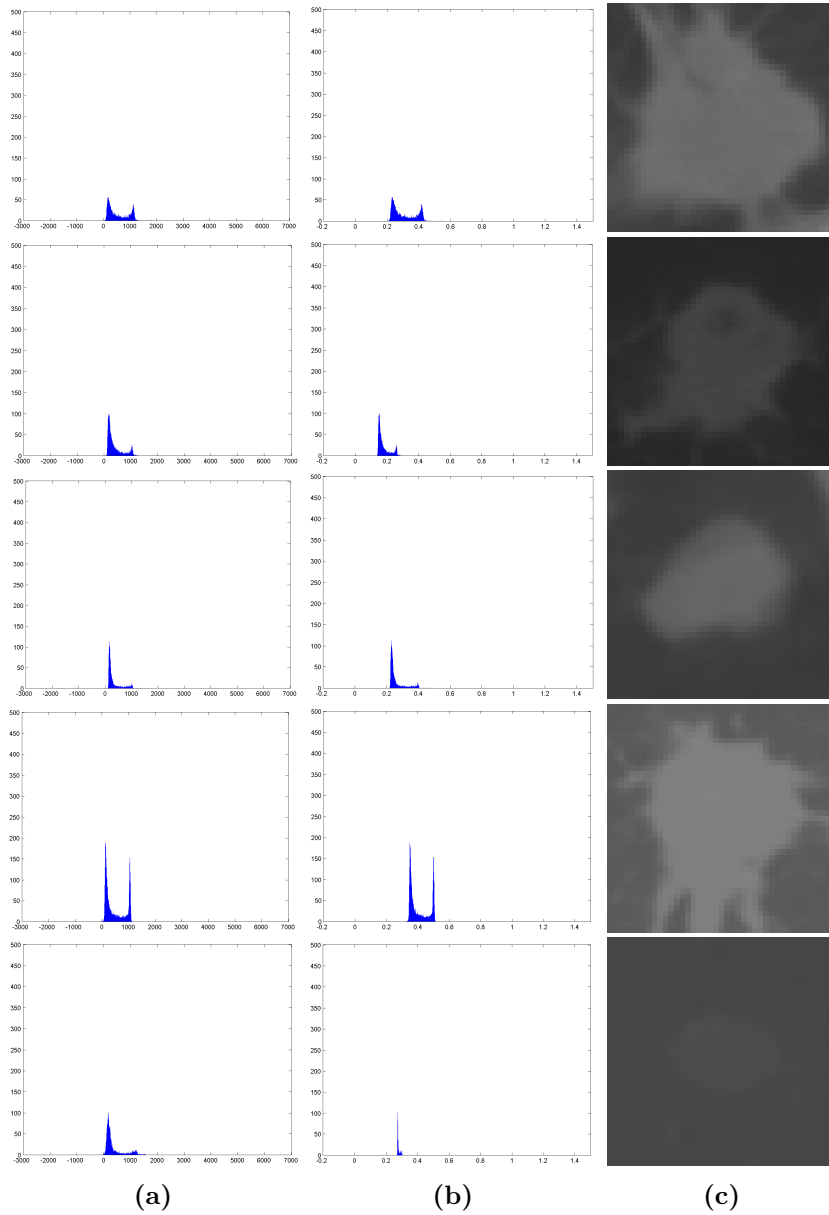


Figure 3.5: Histograms of nodules included in a 3D ROI. (a) Histograms of the ROIs, (b) Histograms of the normalized ROIs, (c) Single slices of the windows containing the nodule. First and second rows represent nodules from enhanced images and third and fourth rows from two unenhanced.

Chapter 4

Methodology for Lung Nodule Segmentation and Classification

Nodule classification is a hard and tiring task to any given radiologist. The diversity of nodules in size, texture and shape makes it a big challenge when trying to define the set of characteristics useful for classifying a nodule in benign or malignant. This way, CAD systems can prove to be very important by assisting the radiologists in the classification task.

As the database has two different types of data, one being the nodules characterized by the radiologists and the other diagnosed nodules, the proposed system is divided in two different classification processes. The first is a system that mimics the classification of the radiologists and the second is a system that diagnosis the nodules similarly to a real biopsy, surgery or follow up exam.

Though both parts of the system are intended for different purposes, the process of diagnosis consists of a set of 5 steps that are common to both. They include nodule segmentation, feature measurement, feature selection, classification and validation. Each one of these steps is very important and determine the overall performance of the system. This work, however, is centred in the lung nodule segmentation and classification stages.

4.1 Lung nodule segmentation

The objective of this work is to classify the nodules into benign or malignant as accurate as possible. In order to do that, a reliable nodule segmentation must be used, so that a good set of nodule characteristics is obtained. Giving the results acquired from the inter-observer segmentation analysis, it becomes clear that the database does provide a nodule outlines' ground truth that varies from radiologists to radiologist. Also, there is no way of knowing what segmentation should be used or if the combination of segmentations is the best option even if the agreement is low. The most effective and efficient way to do lung nodule segmentation is by using an automatic segmentation methodology, particularly when the number

of nodules is large. Therefore, a new strategy was employed following the work developed by Novo *et al.* [75]. This methodology uses the 3D Hessian matrix to calculate the eigenvalues of the images for every voxel at different scales of σ . All process was performed using a ROI with 100 voxel length in the x and y directions in order to accommodate the biggest sized nodule, and for each particular nodule the length in the z direction plus one voxel in the upper and lower limits of the nodule to attain intensity variation in this direction.

In the following is a description of the methodology along with a graphic visualization of every step in figure 4.1.

4.1.1 Nodule segmentation using Hessian matrix

The 3D Hessian matrix, presented in equation 4.2, gives information about the gradient changes between one voxel $p = (x, y, z)$ and its neighbours. The three eigenvalues, λ_1 , λ_2 and λ_3 , which are the magnitude of the eigenvectors of the 3D Hessian matrix, correspond to changes in intensity in the three principal directions or, in other words, where the local intensity change is more prominent. This means that a drastic change in intensity produces a high positive or negative eigenvalue in the corresponding direction depending if the change occurs from a low to a high intensity voxel or the opposite, respectively. Because this information can be used to enhance structures that are located inside the parenchyma and are significantly brighter than their neighbourhood, the result is the enhancement of vasculature and other lung structures. For this purpose, and using an approach described in Novo *et al.* [75], the second order partial derivatives in Hessian matrix H are calculated for several images I , using a Gaussian smoothing filter G with a particular scale of σ . L is the result of the convolution between the images and the Gaussian filter. As in Novo *et al* [75], a set of 7 σ s was used to address the size of nodules in the range from 0.5 to 3.5, increasing 0.5. This multiscale Gaussian smoothing is illustrated in figure 4.1, stage A.

$$L(I, \sigma) = I(p) * G(p, \sigma) \quad (4.1)$$

$$H(I)_\sigma = \begin{bmatrix} \frac{\partial^2 L}{\partial x^2} & \frac{\partial^2 L}{\partial x \partial y} & \frac{\partial^2 L}{\partial x \partial z} \\ \frac{\partial^2 L}{\partial y \partial x} & \frac{\partial^2 L}{\partial y^2} & \frac{\partial^2 L}{\partial y \partial z} \\ \frac{\partial^2 L}{\partial z \partial x} & \frac{\partial^2 L}{\partial z \partial y} & \frac{\partial^2 L}{\partial z^2} \end{bmatrix} \quad (4.2)$$

After obtaining the Hessian matrix H for a particular σ , three eigenvalues (λ_1 , λ_2 and λ_3) are computed (figure 4.1, stage B) and used in different enhancement methods (figure 4.1, stage C), giving a response $V_\sigma(p)$. To combine the responses of different scales, the maximum response at voxel p is calculated for every voxel in the image:

$$V(p) = \max_{\sigma_1 \leq \sigma_j \leq \sigma_n} V_{\sigma_j}(p) \quad (4.3)$$

where j is the scale of the σ and n the total number of σ s. The eigenvalues information is first employed in an adaptation from Murphy *et al* [12]. It uses the maximum (λ_3) and minimum (λ_1) eigenvalues of the Hessian matrix to calculate two indexes, the Shape Index and Curvedness, presented in equations 4.4 and 4.5.

$$SI = \frac{2}{\pi} \arctan \left(\frac{\lambda_3 + \lambda_1}{\lambda_3 - \lambda_1} \right) \quad (4.4)$$

$$CV = \sqrt{\lambda_3^2 + \lambda_1^2} \quad (4.5)$$

The central adaptive medialness principle is the second method, and was adapted from Krissian *et al* [13]. It was firstly used to detect 3D tubular structures and, more recently, for lung vessel extraction. It uses the maximum (λ_3) and medium (λ_2) eigenvalues to calculate the response $V^{med}(\sigma, p)$, in equation 4.6, and enhances both vessels and blob like structures.

$$V^{med}(\sigma, p) \begin{cases} 0 & \lambda_1 + \lambda_2 + \lambda_3 \geq 0 \\ -\frac{\lambda_2}{\lambda_3} \cdot (\lambda_2 + \lambda_3) & otherwise \end{cases} \quad (4.6)$$

Equation 4.3 is used to obtain the maximum response for each SI , CV and V^{med} , calculated for every σ . The results are illustrated in figure 4.1, stage D. It shows that the nodule is clearly enhanced, although in the method adapted from Murphy, particularly in the SI index, there are many small structures that also have high responses. The construction of the mask is done by thresholding the indexes of the methods to select the voxels which present high response values. The threshold values are presented in table 4.1 and were set empirically by performing various experiments in order to build a mask that would include, as precise as possible, the region occupied by the nodule and, consequently, a segmentation with high precision. The resultant image are two binary masks, one for each method, that include the nodules and other lung structures like vessels, pleura or bronchi.

As the main objective is to classify the nodules in benign and malignant, because the image mask presents many other structures, we need to obtain only the mask of the nodules for each method. A procedure was performed to save only the region occupied by the nodule. It consists of using the union of the radiologists segmentations to construct a mask and then multiplying it by the ones generated by the methods. A closing operation was conducted using a disk as a structuring element with a radius of size 5. An example of the final output mask is visible in figure 4.1, stage E. As the methods produce different masks, an agreement analysis between the new masks and the masks from the radiologists must be made. This study will be presented in the experimental results, chapter 5.

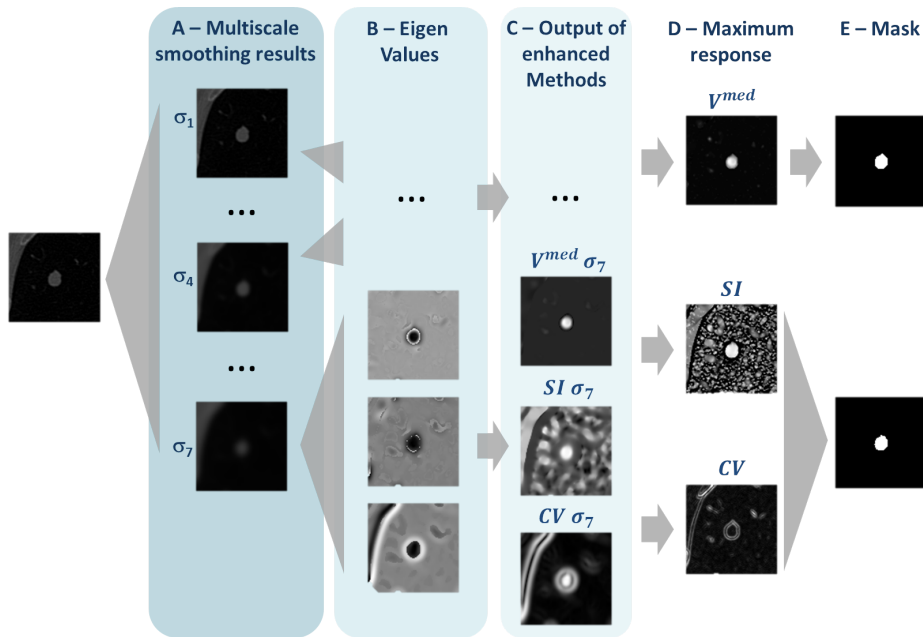


Figure 4.1: Nodule segmentation. A - Multiscale Gaussian smoothing using σ of 0.5 to 3.5, step 0.5. B - Eigenvalues computed from the Hessian Matrix. C - Nodule enhancement for each method for every size of σ (we show only the response for $\sigma = 3.5$). D - Maximum response. E - Final mask.

Table 4.1: Threshold values.

	<i>SI</i>	<i>CV</i>	V^{med}
Threshold	0.3	0.08	0.3

4.2 Feature measurement

Feature measurement is an important stage as we need to obtain good characteristics that lead to good nodule differentiation. This is particularly important in nodule classification, as malignant and benign nodules share many similar characteristics. A generalized set of features were included in order to collect as many properties as possible. They include shape, intensity and texture properties and a descriptor that provides shape information as well as the orientation predominance.

4.2.1 Shape features

The geometric properties can provide important information about the nodule's margin, shape and volume. These features are obtained from the 3D mask of the nodule that due to differences in the sampling resolution between images are normalized using a scale factor. This scale factor is obtained by calculating the

product between the length of the voxels in the x , y and z directions, provided in the *XML* file, for a particular image. The scale factor is given by equation 4.7, where $Voxel_Length_l$ is the length of the voxel in direction $l \in [x, y, z]$. It is defined as $Voxel_Length_l = [Voxel_Length_x; Voxel_Length_y; Voxel_Length_z]$.

$$Voxel_Unit_Volume = Voxel_Length_x \times Voxel_Length_y \times Voxel_Length_z \quad (4.7)$$

The first feature is the volume of the nodule and it is obtained by summing all voxels of the nodule's mask (I_{mask}), multiplied by the $Voxel_Unit_Volume$:

$$Volume = \sum (I_{mask} \times Voxel_Unit_Volume) \quad (4.8)$$

Compactness

The compactness gives information about the 3D projection of the nodule on the x , y and z image planes [12]. This means that the more scattered is the nodule, the lower will be the compactness. If the nodule is somehow round, then the compactness will be high. The compactness is given by equations 4.9 and 4.10, where dim_l is the dimension of the nodule in the x , y and z directions. It is defined as $dim_l = [dim_x; dim_y; dim_z]$ and max is the maximum value.

$$Compactness1 = \frac{Volume}{dim_x \times dim_y \times dim_z \times Scale_Factor} \quad (4.9)$$

$$Compactness2 = \frac{Volume}{max(dim_l \times Voxel_Length_l)} \quad (4.10)$$

Aspect ratios

The aspect ratios can give good insight on how elongated or flat is the nodule. The nodule is flatter if the ratio between the maximum and minimum (min) lengths (equation 4.11) is high and the ratio between maximum and median (med) lengths (equation 4.12) is close to 1. If the ratio between maximum and median lengths is also high, then the nodule is probably elongated.

$$Ratio1 = \frac{max(dim_l \times Voxel_Length_l)}{min(dim_l \times Voxel_Length_l)} \quad (4.11)$$

$$Ratio2 = \frac{max(dim_l \times Voxel_Length_l)}{med(dim_l \times Voxel_Length_l)} \quad (4.12)$$

These ratios do not always provide good information as some nodules are considerably spiculated or lobulated. To address these properties, several properties derived from a principal component analysis are defined.

Principal component analysis

Similarly to *Aspect Ratio* features, the principal component analysis (PCA) analyses the elongation and flatness of the nodules. It computes (equation 4.13) three principal eigenvalues ($\lambda_1, \lambda_2, \lambda_3$) which are the representation of the object's spacial projection in the three main directions. Several ratios were defined in equation 4.14. *Eigen_ratio1*, *Eigen_ratio3* and *Eigen_ratio4* give information about the nodule's flatness. *Eigen_ratio2* calculates the nodule's elongation. Note that λ_1 is the main projection, λ_2 the intermediate projection and λ_3 the smaller projection.

$$[\lambda_1, \lambda_2, \lambda_3] = PCA(Nodule \times Voxel_Length) \quad (4.13)$$

$$Eigen_ratio(1, 2, 3, 4) = \frac{\lambda_3}{\lambda_1}, \frac{\lambda_3}{\lambda_2}, \frac{\lambda_2}{\lambda_1}, \frac{(\lambda_3)^2}{(\lambda_1)^2} \quad (4.14)$$

Sphericity

These properties defined below have the purpose of acquiring the nodule's spherical properties. If the ratios are close to 1, then the nodule presents a spherical shape, otherwise, it is elongated or spiculated. It can give similar information as compactness, however, compactness would be high if the nodule presents a polygonal shape, while sphericity would be low. This is then an important feature to obtain information about the shape of the nodule. $I_{mask} \cap Sphere$ is the resultant mask when multiplying the nodule's mask with an equivalent sphere with radius of the nodule.

$$Sphericity_ratio1 = \frac{\sum(I_{mask} \cap Sphere)}{\sum Sphere} \quad (4.15)$$

$$Sphericity_ratio2 = \frac{\sum(I_{mask} \cap Sphere)}{\sum I_{mask}} \quad (4.16)$$

Calculation of the nodule's equivalent sphere radius:

$$Eq_Sphere_Radius = \frac{\max(dim_l \times Voxel_Length_l)}{2} \quad (4.17)$$

$$Sphericity_ratio3 = \frac{Eigen_ratio3}{(Eq_Sphere_Radius)} \quad (4.18)$$

4.2.2 Intensity features

Intensity properties tend to represent the nodule's degree of calcification. A high presence of calcification gives rise to high HU values in the CT image. Also, the calcification distribution varies between benign and malignant nodules. For example, it allows to evaluate if the calcification is located in the center of the nodule or in the periphery. For this purpose, the intensity features are obtained using different formulas.

Overall intensity

The overall intensity gives information about the subtlety of the nodules and the degree of calcification. The maximum, minimum, mean, median and standard deviation (*std*) of the nodule's intensity (I_{Nodule}) are computed, and I_{Nodule} is obtained by multiplying the initial image I that contains the nodule by the mask I_{mask} : $I_{Nodule} = I \cap I_{mask}$.

$$Max_Intensity = \max(I_{Nodule}) \quad (4.19)$$

$$Min_Intensity = \min(I_{Nodule}) \quad (4.20)$$

$$Mean_Intensity = \text{mean}(I_{Nodule}) \quad (4.21)$$

$$Median_Intensity = \text{med}(I_{Nodule}) \quad (4.22)$$

$$Std_Intensity = \text{std}(I_{Nodule}) \quad (4.23)$$

Intensity over spheres

The Intensity Over Spheres features are computed similarly to the *Overall Intensity* features, however, they are designed to acquire the degree of intensity from the center to the periphery. Their main objective is to provide information about the central calcification. Spheres with radius (r) of 1, 3 and *Eq_Sphere_Radius*, centred in the nodules' location, are used as masks to calculate different properties of the nodule.

$$Max_Intensity_Overlap = \max(I_{Nodule} \cap Sphere_r) \quad (4.24)$$

$$Min_Intensity_Overlap = \min(I_{Nodule} \cap Sphere_r) \quad (4.25)$$

$$Mean_Intensity_Overlap = \text{mean}(I_{Nodule} \cap Sphere_r) \quad (4.26)$$

$$Median_Intensity_Overlap = \text{med}(I_{Nodule} \cap Sphere_r) \quad (4.27)$$

$$Std_Intensity_Overlap = \text{std}(I_{Nodule} \cap Sphere_r) \quad (4.28)$$

4.2.3 Texture features

Texture features are very important as they mostly measure internal characteristics (air component, intra-nodular fat, cavitation, calcification), that in turn are essential in distinguishing benign from malignant nodules. They can be defined as first, second and higher order statistics. Each feature was measured in the slice where the nodule was most represented (slice where the nodule has the largest area).

First order statistics: gray-level intensity histogram

First order statistics are obtained by computing the probability of finding a particular intensity in a random location on the image. Gray-level Intensity Histogram (GLIH) features are computed by estimating all intensity probabilities and then analysing the resultant histogram.

As stated previously, I_{Nodule} is the segmented nodule. $quant(P_{x,y,z})$ is the function that returns the number of occurrences of a given intensity i and $P_{x,y,z}$ the intensity i of $Voxel_{x,y,z}$. We can then calculate the histogram of the image by equation 4.29 [76],

$$h(I_{Nodule}) = \{quant(P_{x,y,z})|P_{x,y,z} = I_{Nodule}\} \quad (4.29)$$

After obtaining h_i , we can then calculate the mean value, μ (equation 4.30), and the variance, σ^2 (equation 4.31), which is also the second angular moment of the image I . N is the maximum intensity of the image.

$$\mu = \sum_{i=0}^{N-1} i.h_i \quad (4.30)$$

$$\sigma^2 = \sum_{i=0}^{N-1} (i - \mu)^2 .h_i \quad (4.31)$$

Given μ and σ , we calculate the third and fourth, $n=[3,4]$, angular moments for each nodule:

$$M_n = \sum_{i=0}^{N-1} (i - \mu)^n .h_i \quad (4.32)$$

The third moment is used to compute *Obliquity*. The *Obliquity* is given by equation 4.33 and measures the asymmetry of the histogram probability distribution. If the histogram tends to have higher occurrences in darker intensities, the value of *Obliquity* tends to positive. On the other hand, if the occurrences are higher in brighter intensities, then the *Obliquity* tends to negative.

$$Obliquity = \frac{M_3}{\sigma^3} \quad (4.33)$$

The fourth moment is used to compute *Kurtosis*. The *Kurtosis* is given by equation 4.34 and measures the distribution of the histogram. The sharper the h_i , the higher is the value of *Kurtosis*.

$$Kurtosis = \frac{M_4}{\sigma^4} \quad (4.34)$$

Other two features can be computed using h_i . They are *Energy* and *Entropy*, given by equations 4.35 and 4.36. The *Energy* represents the intensity variation in the region and *Entropy* the distribution of the histogram.

$$Energy = \sum_{i=0}^{N-1} (h_i)^2 \quad (4.35)$$

$$Entropy = - \sum_{i=0}^{N-1} h_i \cdot \log(h_i) \quad (4.36)$$

Second order statistics: gray-level co-occurrence matrix

The Gray-Level Co-occurrence Matrix (GLCM) [77], given by equation 4.37, is a measure of the simultaneous occurrence of gray-levels i and j in pairs of pixels (p_1, p_2) separated by a displacement vector $\delta = (\Delta x, \Delta y)$ into a 2D histogram.

$$C_\delta(i, j) = |\{p_1, p_2 \in I : I(p_1) = i; I(p_2) = j; p_2 = p_1 \pm \delta\}| \quad (4.37)$$

Using this measure, we can calculate the probability of a particular intensity to appear in the image using equation 4.38, which can in turn be used to compute several features that are the representation of the relations between pixels and the analysis of the GLCM.

$$p_{\delta ij} = \frac{C_\delta(i, j)}{\sum_{ij} C_\delta(i, j)} \quad (4.38)$$

As proposed by Haralick *et al.* [77], Conners *et al.* [78], Soh *et al.* [79] and Clausi *et al.* [80], many texture features may be extracted from the co-occurrence matrices. In this work 20 texture features were computed for four directions, 0° , 45° , 90° and 135° , and distance of 2. The distance was set to 2 pixels because the database as many small nodules, many having only 4 or 5 pixels. Increasing the distance would not give any additional information. The gray levels were set to 10 and 20 and the symmetry to true. To avoid direction dependency, the angular mean and standard deviation for each textural measure were calculated as proposed in [81].

Second order statistics: gabor filters

A Gabor filter is a sinusoidal plane wave (carrier), composed by a particular frequency and orientation, modulated by a Gaussian envelope. In computer vision and image processing it is used mainly for texture analysis. A 2-D Gabor filter

over the image domain (x, y) , with a particular frequency f and an orientation θ is defined as:

$$\psi(x, y; f, \theta) = \frac{f^2}{\pi\gamma\eta} \cdot e^{-f^2 \left[\frac{x'^2}{\gamma^2} + \frac{y'^2}{\eta^2} \right]} \cdot e^{j2\pi f x'} \quad (4.39)$$

for

$$\begin{aligned} x' &= x \cos(\theta) + y \sin(\theta) \\ y' &= -x \sin(\theta) + y \cos(\theta) \end{aligned} \quad (4.40)$$

In the frequency domain (u, v) it is defined as:

$$\psi(u, v; f, \theta) = e^{-\frac{\pi^2}{f^2} [\gamma^2(u-f)^2 + \eta^2 v^2]} \quad (4.41)$$

for

$$\begin{aligned} x' &= u \cos(\theta) + v \sin(\theta) \\ y' &= -u \sin(\theta) + v \cos(\theta) \end{aligned} \quad (4.42)$$

where γ define the effective width of the filter in the frequency projection and η is the effective width of the filter in the orientation projection.

By varying f and θ , a filter bank can be build to extract different features. The orientation θ_k can be uniformly defined by:

$$\theta_k = \frac{k\pi}{n} \quad k = 0, \dots, n - 1 \quad (4.43)$$

where n is the total number of orientations.

In order to maintain homogeneity spacing between the filters, a logarithmic relation between the frequencies f can be established by:

$$f_c = \frac{f_{max}}{\sqrt{2}^c} \quad c = 0, \dots, m - 1 \quad (4.44)$$

where f_c is the c th frequency, m is the total number of frequencies and f_{max} is the highest frequency desired. The scaling factor was set to $\sqrt{2}$ for a half-octave spacing.

For the purpose of this work, two different banks of filters were set. One with 5 scales of frequencies f ($m = 5$) and another with 8 ($m = 8$). Both have the same number of orientations θ , set to 8 ($n = 8$). The maximum frequency f_{max} was defined as 0.25 and $\gamma = \eta = \sqrt{2}$. This resulted in a total bank with 108 filters.

To obtain the features using the filter bank, each Gabor filter in the space domain was convoluted with the image slice where the nodule is more represented (the are is bigger). Then, the mean and standard deviation was calculated for each of the 108 image responses resulting in 216 features [82–84].

Higher order statistics: Laws' texture energy measures

The Law's Measures provide the amount of variation within a fixed-size window by using a set of convolution masks used to compute the texture energy of each pixel in the image. These convolution masks are obtained by computing the product of the one-dimensional Lattice Aperture Waveform Sets (LAWS) of order 3, 5 and 7. Sets of order 5 are obtained by first convolving two Sets of order 3. The Sets of order 7 are obtained by convolving Sets of order 3 and 5. The Center-Weighted Vector Masks, each representing a particular property of the image of orders 3, 5 and 7 are:

$$\begin{aligned}
 \text{L3 (Level)} &= [1 \ 2 \ 1] \\
 \text{E3 (Edge)} &= [-1 \ 0 \ 1] \\
 \text{S3 (Spot)} &= [-1 \ 2 \ -1] \\
 \\
 \text{L5 (Level)} &= [1 \ 4 \ 6 \ 4 \ 1] \\
 \text{E5 (Edge)} &= [-1 \ -2 \ 0 \ 2 \ 1] \\
 \text{S5 (Spot)} &= [-1 \ 0 \ 2 \ 0 \ -1] \\
 \text{W5 (Wave)} &= [-1 \ 2 \ 0 \ -2 \ 1] \\
 \text{R5 (Ripple)} &= [1 \ -4 \ 6 \ -4 \ 1] \\
 \\
 \text{L7 (Level)} &= [1 \ 6 \ 15 \ 20 \ 15 \ 6 \ 1] \\
 \text{E7 (Edge)} &= [-1 \ -4 \ -5 \ 0 \ 5 \ 4 \ 1] \\
 \text{S7 (Spot)} &= [-1 \ -2 \ 1 \ 4 \ 1 \ -2 \ -1] \\
 \text{W7 (Wave)} &= [-1 \ 0 \ 3 \ 0 \ -3 \ 0 \ 1] \\
 \text{R7 (Ripple)} &= [1 \ -2 \ -1 \ 4 \ -1 \ -2 \ 1] \\
 \text{O7 (Oscillation)} &= [-1 \ 6 \ -15 \ 20 \ -15 \ 6 \ -1]
 \end{aligned}$$

The convolution masks obtained from LAWS of order 5 (5×5) are more appropriate for image analysis as they are more powerful than 3×3 masks (order 3) and simpler than 7×7 (order 7), being also similar. In the lung nodule classification problem, convolution masks of order 7 are also too big for small nodules giving no precise information about the nodules texture.

As suggested in [85], we also consider only 4 LAWS of order 5. The product of the vectors result in 16 energy maps that can, given the symmetry of certain pairs, be combined to build 9 convolution maps, replacing each pair with its average. For our purpose, the convolution masks obtained and used here were: L5E5\ E5L5, L5R5\ R5L5, E5S5\ S5E5, S5S5, R5R5, L5S5\ S5L5, L5E5\ E5L5, E5E5, E5R5\ R5E5, S5R5\ R5S5. Though 3×3 convolution masks are simpler, they are also used in this study. The convolution masks used here are: L3E3\ E3L3, L3S3\ S3L3, E3E3, S3E3\ E3S3, S3S3.

After filtering the image with the convolution masks, the resultant filtered image must be reduced to a single feature array. To obtain this, the mean and standard deviation for each filtered image are calculated resulting in a total of 28 features $((9 + 5) \times 2)$.

4.3 Feature selection and classification

After obtaining all features, it is necessary to select the ones that provide the best discrimination between classes. The following sections present the proposed methodology to study the performance of different classifiers and assess the best feature set/classifier combination that provides the optimal result.

4.3.1 Feature selection

The cloud of features must be reduced by eliminating the ones who give redundant information and saving the ones who best discriminate the nodules in benign and malignant. As stated in subsection 2.2.3, there are three types of model searches: filter based, wrappers and embedded. The most suited for feature selection are the wrapper methods, but are computationally expensive, specially when coupled with SVMs, due to its great deal of computation [86]. The solution would be to implement specially designed feature selection methods for SVM, found in [86–90], or to use filter based methods. The choice was based on two aspects: the complexity involved in implementing a feature selection method for the SVM; k-NN based classifiers are used with the objective to compare both classifiers using the same feature selection methods. Given these reasons, two filter based methods are used to build two different sets of features. The objective is to find the best possible performance and by using two different methods, the probability for that to happen increases and, also, it is possible to assess what are performances of the classifiers using different subsets.

The filter based methods used here are the Correlation-based Attribute Subset evaluator or CFS, [43], and the Relief-F evaluator, [44]. The CFS analyses the strength of a feature in predicting the class of the object, but tends to give little importance to the inter-correlation of the features. It adds features that have high correlation with the class, but only if the set does not have a corresponding high correlated attribute. Relief-F samples instances randomly and checks the distance between them and neighbour instances that have the same or different classes. The number of neighbours to be checked must be defined, such as the number of instances to be analysed. An exponential function is used to determine the weight for a given distance, which in turn will be used to rank the features [43] [44].

4.3.2 Classification

Two different classifiers are used in this work. One is the Support Vector Machine (SVM) and the other is the k-Nearest Neighbour (k-NN). As both classifiers are sensible to different settings, a study was conducted in order to achieve the best performance. For the k-NN different sets of k values were defined and evaluated. For the SVM, three Θ values were defined using an exponential kernel $K(x,y)$ defined in equation 4.45 [91].

$$K(x, y) = \exp\left(\frac{-\|x - y\|}{\Theta}\right) \quad (4.45)$$

4.4 Validation

As stated in section 2.2.5, the validation is the stage where it is made an evaluation of the system. For this purpose, it is important to compare the performance of both classifiers with different sets of characteristics. Several measures can be used to calculate that performance. They are sensitivity, specificity or AUC value. The most used, however, is the AUC value, common to many lung nodule classification problems found in literature [23] [27] [28] [30] [33] [32] [35].

The AUC value is the area under the receiver operating characteristic (ROC) curve, being the trade-off between true-positives and false-positives rates. By using this measure, the results can be represented by one parameter that gives the overall classification achieved for each classifier and a given set.

The classification is performed by 10-fold cross-validation, with 50 repetitions to obtain a solid estimation of the classifiers' performance. As each repetition presents one performance value, all values are averaged to get the overall performance.

Chapter 5

Experimental Results

This chapter presents the implementation of the methodology described in chapter 4. Section 5.1 addresses the construction of the ground truth (GT), section 5.2 nodule segmentation procedure, section 5.3 the results of feature selection and classification, and section 5.4 some tests using the CAD system.

5.1 Construction of the ground truth framework

Before any feature selection or classification can be performed, a GT must first be organized. The LIDC-IDRI database, as addressed in section 3.2, presents two different types of data that can be used to form two separate datasets. One is provided by the *XML* file and represents the radiologists assessment. It characterizes all the nodules in degree of malignancy using a scale from 1 (*Highly Unlikely*) to 5 (*Highly Suspicious*). From all the nodules characterized in the *XML file*, there is a small set of 34 that was evaluated on one of the following grades: review of radiological images to show 2 years of stable nodule, biopsy, surgical resection and progression. This diagnosis is provided by a *.xls* file that labels the nodules as benign or non-malignant disease, malignant - primary lung cancer, or malignant - metastatic.

By using two datasets, it is possible to build a system that performs two classifications, one similar to the assessment of radiologists and another, more reliable, similar to a biopsy, surgery or follow up exam. However, it is visible that the two types of data are labelled differently. The *XML* file contains the degrees of malignancy, in a scale from 1 to 5, given by one to four radiologists, and the *.xls* file distinguishes the nodules in benign or malignant, but the malignancy is defined as primary lung cancer or metastatic. To simplify the problem, the malignancy was set as benign and malignant for both datasets, so the GT only has two classes. To achieve this, we developed a strategy that is described in the following paragraphs. Sub-section 5.1.1 describes the construction of the GT for the first dataset using the radiologists' assessment, from this moment on referred as the Radiologists' data, and sub-section 5.1.2 describes the construction of the GT for the second

dataset, from this moment on referred as the Diagnostic data, provided by the *.xls* file.

5.1.1 Radiologists' data. Ground truth

To build the Radiologists' data using the *XML* file, we first selected solid nodules (*Texture* of 4 and 5) annotated by all four radiologists, meaning that there must be a total agreement between radiologists to include that nodule in the dataset. Only solid nodules were selected because the characteristics associated to the degree of malignancy in nodules are different for solid, mixed or GGO nodules. Given the larger amount of solid nodules compared to the other types, it is advisable to classify each one of them independently. In this work, the focus is only on the solid type, as it is the most common, and a dataset with 579 nodules was created.

The GT for this dataset is the nodules' labelling from four radiologists. Naturally, the opinions among radiologists are not in agreement, so different degrees of malignancy are often given to one nodule. Besides this, some nodules are also labelled as *Indeterminate*, giving no useful information to the classifier. Because of this relatively significant disagreement, the GT for the Radiologists' dataset will be organized in 3 different types, Ground Truth 1, Ground Truth 2 and Ground Truth 3, each one representing a different level of agreement, being Ground Truth 1 the one with the highest agreement and Ground Truth 3 with the lowest agreement. The process is illustrated in figure 5.1.

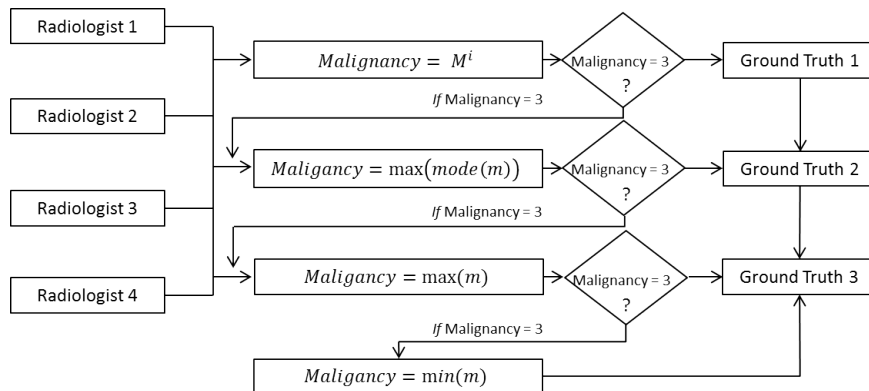


Figure 5.1: Diagram of the construction of the GT using the radiologists assessment. The input is the classification of four radiologists and the output are three different GT, one for each degree of agreement. m is the array of labels for the degree malignancy and p a weight value.

The first level (Ground Truth 1) consists of estimating the degree of malignancy M^i , for each nodule i , given by the expression:

Table 5.1: Weights for the radiologists classification. m is the array of labels for the degree malignancy and p the weight value.

m	1	2	3	4	5
p	0.15	0.15	0.20	0.25	0.25

$$M^i = \alpha \sum_{k=1}^4 \beta_i^k dm^i(k) \therefore \beta_i^k = p(k); \alpha = \frac{1}{\sum_{k=1}^4 \beta_i^k} \quad (5.1)$$

where $p(k)$ is the weight value for each radiologist k and $dm^i(k)$ is the determination of the assigned label m by the radiologist for a particular nodule. Table 5.1 presents the corresponding weights p for all degrees of malignancy m , and it is visible that the weights are bigger for labels 4 and 5, as we give more importance to the opinion of the radiologists that classify a nodule as *Highly Suspicious* or *Moderately Suspicious* of malignancy rather than *Highly Unlikely* or *Moderately Unlikely* of benignity. After estimating M^i for one nodule, it is verified if the returned label is 3 (*Indeterminate*). If it is different from 3, then the label is saved in all three Ground Truths because the agreement between radiologists is high. If the returned value is equal to 3, then that label is saved in Ground Truth 1 and the second level (Ground Truth 2) is used.

The second level consists of verifying what is the most agreed radiologist classification. Many nodules have distinct classifications, but in some cases, two or more radiologists can be in agreement. In those cases it is assumed that the majority of opinions is sufficient to determine the nodule's degree of malignancy. Sometimes, however, there are two or more majority opinions in cases where all radiologists assign different labels or when two radiologists agree in one opinion and the other two agree in another. In these cases the highest value is saved, again due to the fact that we give more importance to radiologists that suspect a nodule to be malignant. The output of this level is saved in Ground Truth 2, but, similarly to the first level, the label is verified to see if it is equal to 3. If it is different, the label is also saved in Ground Truth 3, if not, then the third level is used.

The third level is only used when the majority of opinions is 3 and its objective is to find if at least one radiologist classified the nodule as not 3. In this level, the first step is to find the maximum value of the array of opinions. However, if the returned label is 3, then the minimum value is found. Finally, the output label of this level is saved in Ground Truth 3.

Each GT presents a different number of nodules for each label, particularly for label 3. To have an idea on the distribution of nodules for every label, the output for each level was compiled in three tables, one one for each of the GT, and are visible in appendix B. There, the number of nodules for the labels of each GT versus the labels from radiologists are presented.

To complete the GT set up for all three Ground Truths, if the assigned label is

equal to or higher than 4, then the nodule is considered malignant and the label is changed to 5, if the decision is equal to or lower than 2, then the nodule is benign and the label is changed to 1. The total number of nodules having benign, malignant and indeterminate labels for each of the GTs is presented in table 5.2. It is visible that the number of indeterminate nodules decreases from Ground Truth 1 to Ground Truth 3 as they are labelled as benign or malignant.

Table 5.2: Number of nodules having benign, malignant and indeterminate labels for each of the GTs.

	<i>Benign</i>	<i>Malignant</i>	<i>Indeterminate</i>
Ground Truth 1	121	179	245
Ground Truth 2	205	204	138
Ground Truth 3	255	277	13

5.1.2 Diagnostic data. Ground truth

The GT provided by the *.xls* file contains information about nodules with known diagnosis. Information of a total of 157 cases is provided with the corresponding malignancy and origin (some nodules are metastatic) for one to two nodules per case. Although the number of cases is relatively large there are some problems regarding the identification of the diagnosis for a particular nodule and some cases are actually missing a diagnosis. Regarding the identification of the nodules, the *.xls* file does not actually identifies the nodule or nodules that are diagnosed. This means that when a case has 2 nodules, if they have two different diagnosis, then it is impossible to know whose diagnosis corresponds to. Even if the diagnosis is the same, when the case has 3 or more nodules, there is no way of knowing which nodule was actually diagnosed.

Knowing all the restrictions that this information file presents, to manage this dataset we selected the cases that only have one nodule, or the cases that have two nodules and both have the same level of malignancy. This returns a total of 13 benign nodules and 21 malignant nodules. For the management of the GT for the Diagnostic data, nodules assigned in the *.xls* file as benign or non-malignant disease are labelled as benign and nodules assigned as primary lung cancer or metastatic are labelled as malignant.

5.2 Lung nodule segmentation

Nodule segmentation was performed using the methodology proposed in section 4.1. Two methods were applied in this work and examples with all stages of the segmentation are presented in figure 5.2. Though the whole process was performed in 3D, the results are shown in 2D images.

	<i>Original Image</i>	<i>Gaussian Blurring</i>	<i>Eigen values</i>	<i>Methods Response</i>	<i>Maximum Response</i>	<i>Masks</i>	
a)		$\sigma = 0,5$ 					Krissian
							Murphy
b)		$\sigma = 1,5$ 					Krissian
							Murphy
c)		$\sigma = 3,5$ 					Krissian
							Murphy

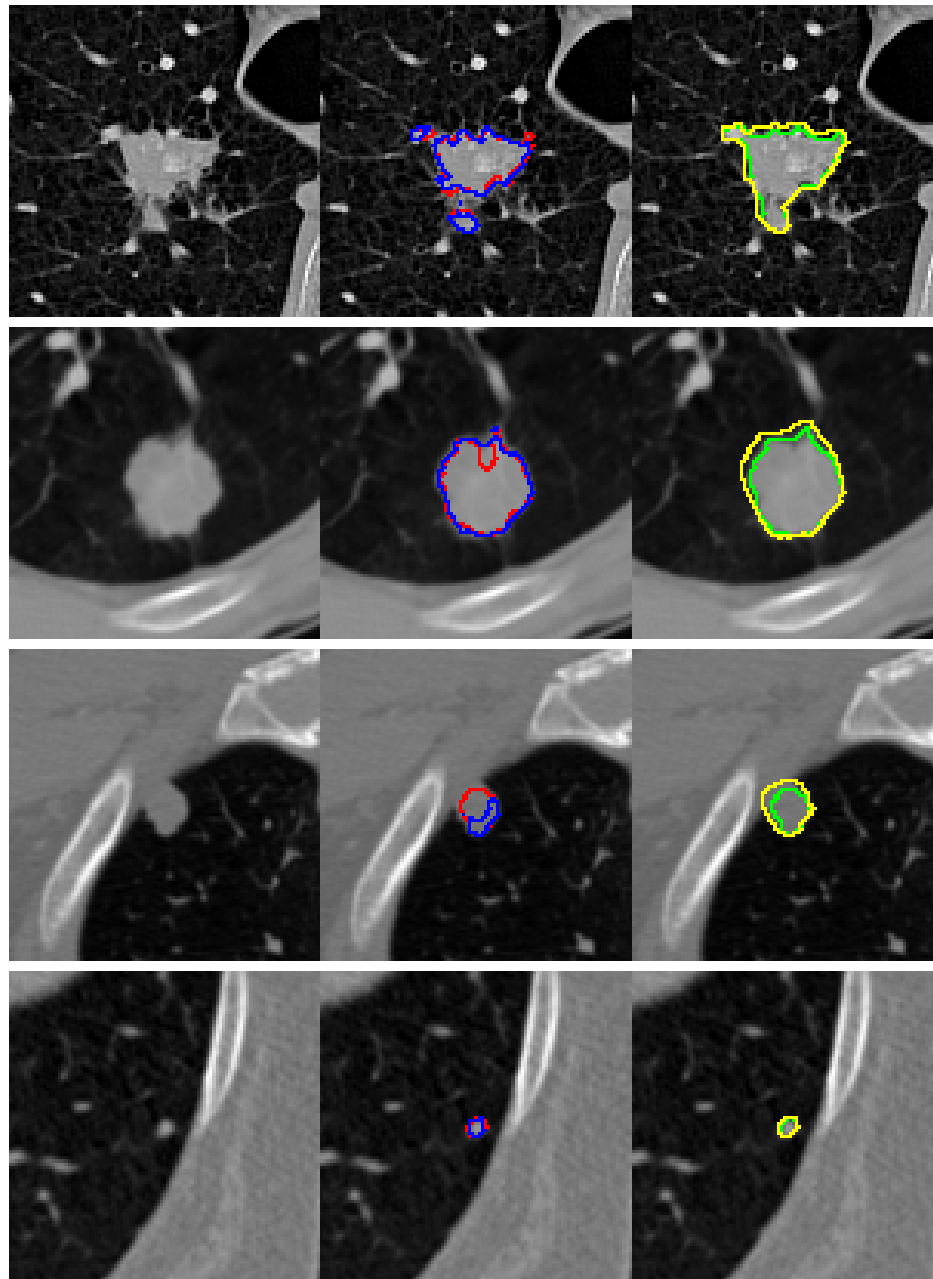
Figure 5.2: All stages of the segmentation procedure are presented for a particular σ . The columns of Maximum Response and Masks give the result for all σ s. (a) Stages for a medium, justa-pleural nodule, using a $\sigma = 0.5$. (b) Stages for a large, justa-pleural nodule, crossed by an airway (hole in black), using a $\sigma = 1.5$. (c) Stages for a large, lobulated nodule, using a $\sigma = 3.5$.

The first four columns of figure 5.2 show the input image, the image blurred with a Gaussian filter, the response of the 3 eigenvalues of the Hessian matrix and each of the methods response (V^{med} , CV and SI) for one σ . The remaining columns present the maximum response after using every σ and the final mask.

Figure 5.2a shows a good segmentation of a medium, justa-pleural nodule, using a $\sigma = 0.5$. It is visible that by using Krissian’s method, a good segmentation is achieved, while the adaptation from Murphy’s does not. Figure 5.2b shows a large, justa-pleural nodule, crossed by an airway (black hole inside the nodule), using a $\sigma = 1.5$. For this nodule, a bad segmentation is seen using both methods. This is probably due to the fact that the nodule is strongly connected to the pleura and there is an airway crossing its center, which has high positive eigenvalues resulting in a low response in V^{med} , CV and SI . Figure 5.2c presents a reasonable segmentation of a large, lobulated nodule, using a $\sigma = 3.5$. It is visible that both methods fail a high response in the connection between the large portion of the nodule and its lobulated part. The outlines of several other segmentations are presented in figure 5.3 in yellow color.

5.2.1 Agreement analysis between methods’ segmentations and radiologists’ segmentations

The lung nodule segmentation procedure has different performance for different types of nodules. The results vary with size, location and texture. Examples of segmentations for different types of nodules can be seen in figure 5.3. Figure 5.3a shows four nodules in ROIs, figure 5.3b the corresponding outlines for both methods, and figure 5.3c the final mask combination and the union of the radiologists outlines. It is visible that the methods have different responses for different nodules. Looking at the nodules in figure 5.3b, first and fourth rows, there is no clear conclusion of which method performs the best. However, for the nodule in second row, Murphy’s adapted method (blue outline) clearly outperforms the method from Krissian (red outline) and the opposite is true for the nodule in third row. Looking at the green outline from figure 5.3c, it is clear that in any case, the best way to segment the nodules is to combine both methods and perform the closing operation, which smooths the margins of the mask and even corrects some poor segmentations like the one presented in figure 5.3a, first row.



(a)

(b)

(c)

Figure 5.3: a) Original ROIs with nodules. b) Outlines of the masks produced by both methods. The red line corresponds to the approach from Krissian and blue line to the approach adapted from Murphy. c) Comparisons between the outlines of the final masks (green line) and the outlines of radiologists (yellow line). First row - A big size, justa-vascular nodule. Second row - A big size, round, justa-vascular nodule. Third row - Medium sized, justa-pleural nodule. Fourth row - A small size, intra-parenchymal nodule.

Observing the segmentation outlines originated from the methods and the union of the radiologists (yellow line), presented in figure 5.3c, it is visible that, in these cases there is a better outline given by the green line than by the yellow line. However, in order to make any conclusion regarding the performance of the proposed lung nodule segmentation algorithm, an agreement study was conducted between the new segmentations and the radiologists' segmentations.

To evaluate the degree of agreement for the segmentations from the developed methods, the Jaccard index was also obtained for the masks generated by both methods and, additionally, the combination of both. The analysis was not extended to GGOs as the agreement between radiologists is very low and they are not used in this work. Each segmentation was compared with the radiologists' segmentations. This resulted in three Jaccard values for each one of these methods so a mean value agreement was calculated by averaging the results. The reason why the mean Jaccard index for the radiologists segmentations was used is because the unknown identity of the radiologists for different cases makes it impossible to link one annotation to another, as referred previously in section 3.3.1.

There are now available two different types of segmentations (methods and radiologists), so it is useful to observe which one of them presents a better overall agreement. The Bland-Altman method, [92], is used to measure the agreement between metrics for different methods. It was applied using the Jaccard's indexes obtained previously, where the vertical axis is the difference between the mean Jaccard value of the radiologists masks and the mean Jaccard value of the masks' segmentation methods (Mean Difference). The horizontal axis is the their Mean Value.

Figure C.1 in Appendix C presents the Bland-Altman results for small nodules, figure C.2 for medium sized nodules and figure C.3 for big nodules. For better comprehension, the Mean Difference for every Bland-Altman plot is presented in table 5.3. Analysing table 5.3, it is visible that the Krissan method presents a mean negative value for every size, indicating that it agrees more with each one of the radiologists than between themselves. The method adapted from Murphy, however, presents positive values indicating that it agrees less. Both results are supported by the examples in figure 5.3, though, sometimes, as visible in figure 5.3b second row, Murphy's method outperforms Krissian's method. The combination of both, however, presents higher negative values than Krissian alone, indicating that together they are more reliable. The results in table 5.3 also show that the agreement is higher for small nodules. This is probably the result of the lower agreement between radiologists for that size. In fact, the mean value tends to be lower as the nodules grow in size.

The Bland-Altman plots also show that, for large and small nodules, the difference value between Jaccard values is farther from the Mean Difference when the Mean Values are lower. In other words, the data is more scattered for lower Mean Values. For medium sized nodules, it appears that the same pattern is present, but the amount of data is insufficient to make any conclusion.

Table 5.3: Results for the Bland-Altman study. Mean of the difference between the mean Jaccard value of the radiologists and the mean Jaccard value of the methods segmentations.

Nodules	Volume (mm^3)	Murphy	Krissian	Combination
<i>Small</i>	<500	0.03	-0.05	-0.053
<i>Medium</i>	$\geq 500-1000$	0.0027	-0.033	-0.038
<i>Big</i>	>1000	0.054	-0.007	-0.017

5.3 Feature selection and classification

A total of 293 features, presented in table 5.4, were measured for each nodule. For both datasets we need to select a subset of features that can differentiate, as efficiently as possible, malignant from benign nodules. Also, it must be assessed which of the different classifiers gives the best results.

Table 5.4: Measured features. (Gabor - the features are described as $Si-Ow$, where S means scale, i is the total number of scales, O means orientation and w is the total number of orientations. Other definitions: *max* - Maximum; *min* - Minimum; *std* - Standard Deviation; *r* - Radius)

Features	
Geometric	[1] - Volume [2,3] - Compactness1; Compactness2 [4,5] - Ratio1; Ratio2 [6-9] - Eigen_ratio(1, 2, 3, 4) [10-12] - Sphericity_ratio1; Sphericity_ratio2; Sphericity_ratio3
	[13-18] - Overall Intensity: Max; Min; Mean; Median; Std [19-32] - Intensity Over Spheres ($r = 1, 2, Eq\text{-Sphere}\text{-Radius}$): Max; Min; Mean; Median; Std
Texture	<i>GLIH</i> [33-37] - Obliquity; Kurtosis; Energy; Mean_intensity (μ)
	<i>GLCM</i> [33-53] - Autocorrelation; Contrast; Correlation; Cluster Prominence; Cluster Shade; Dissimilarity; Energy; Entropy; Homogeneity; Maximum probability; Sum of squares; Variance; Sum average; Sum variance; Sum entropy; Difference variance; Difference entropy; Information measure of correlation1; Informaiton measure of correlation2; Inverse difference; Inverse difference normalized; Inverse difference moment normalized
	<i>LAWS</i> [58-68] - mean Laws 3×3 ; std Laws 3×3 (5 convolution masks) [59-85] - mean Laws 5×5 ; std Laws 5×5 (9 convolution masks)
	<i>Gabor</i> [86-166] - mean Gabor S5-O8; std Gabor S5 ($5 \times 8 = 40$ filters) [167-293] - mean Gabor S8-O8; std Gabor S8 ($8 \times 8 = 64$ filters)

This section, therefore, presents the feature selection and classification procedures, the results for both Radiologists' data and Diagnostic data, as well as the results of a classification where both datasets are crossed with each other, meaning that the training of the classifier is first done on one dataset and the testing

on the other, and vice versa. This procedure is, for now on, called Inter-datasets validation.

5.3.1 Radiologists' data. Procedure and results

For the Radiologists' dataset, from the available GTs, Ground Truth 1 was used as it is the one with the highest agreement between all three. After excluding the indeterminate nodules, the dataset was left with 179 malignant and 121 benign nodules. This dataset was randomly divided in two smaller datasets with similar size, with both containing similar numbers of benign and malignant nodules. The sets were used in different ways. One was used to select the best features and the second to perform the classification. Feature selection was performed by 10-fold cross validation and the results were obtained for each method. As mentioned before, the filter based feature selection methods give the best features but not the best subset for a particular classifier, which means that there must be set a threshold to select the ones who have the highest rank. The CFS gives the number of folds where a particular feature is found and this percentage can vary from 0% to 100%. In this work, the threshold was set to include highly ranked features, whose total would be approximately 10 times smaller than the dataset to sufficiently represent the nodules of this dataset. To achieve this, the features that appeared in more than 80% of the folds were selected and ranked according to the number of times they were selected (100%-80%). The Relief-F simply presents the ranking for each feature, so for a k value of 10, the highest ranked features were chosen until the selected total was equal to the number of features selected by the CFS. This was done in order to compare both feature selection algorithms. The total number of features selected by both methods was 12 and are presented in table 5.5, ordered according to relevance.

Six classifiers were also set, three k-NN based and three SVM based. The k-NN classifiers were set with different k values, defined as 13, 15 and 17, and the SVM classifiers were defined with an exponential Kernel and parameters θ of 1, 2 and 3. The classification was performed by 10-fold cross-validation on the second set with 50 repetitions. The mean and standard deviation were calculated from the 50 values of AUC, using the selected subset of features. The results are presented in figure 5.4 in form of a bar chart. The AUC value was calculated from the ROC curve, obtained after varying the thresholds from 0 to 1, increasing 0.01.

Table 5.5: Selected features using CFS and Relief F for the Radiologists' data and Diagnostic data. (For Gabor, the features are described as $Si-j Ok$, where S means scale, i is the total number of scales, j the scale of the filter, O means orientation and k is the number of the corresponding orientation. Other definitions: std - Standard Deviation; r - Radius).

	CFS	Relief F
Radiologists' data	1 - Compactness1 2 - Difference Entropy (GLCM) 3 - Inverse Difference Normalized (GLCM) 4 - Volume 5 - Entropy (GLIH) 6 - std Gabor S5-5 O4 7 - std Laws L3E3\E3L3 8 - mean Laws E3E3 9 - std Laws S3S3 10 - mean Laws E5S5\S5E5 11 - Cluster Shade (GLCM) 12 - Information Measure of Correlation1 (GLCM)	1 - Cluster Shade (GLCM) 2- Information Measure of Correlation2 (GLCM) 3 - Inverse Difference Normalized (GLCM) 4 - Information Measure of Correlation1 (GLCM) 5 - Sum Entropy (GLCM) 6 - Autocorrelation (GLCM) 7 - Entropy (GLCM) 8 - Difference Entropy (GLCM) 9 - <i>Max_Intensity_Overlap</i> ($r = 1$) 10 - <i>Max_Intensity_Overlap</i> ($r = 3$) 11 - std Laws L5E5\E5L5 12 - <i>Max_Intensity_Overlap</i> ($r = Eq_Sphere_Radius$)

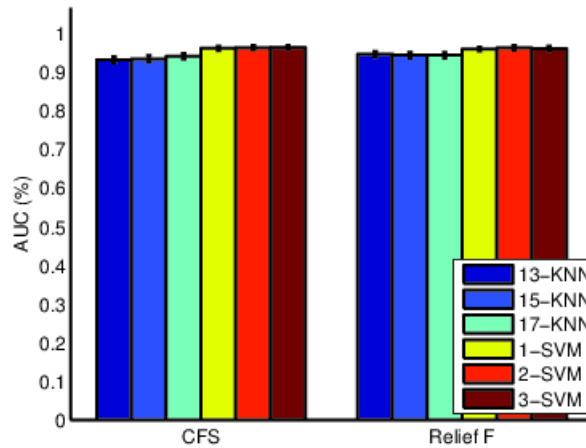


Figure 5.4: Classification results for the Radiologists' data, presented as AUC (%) value, for 12 features selected by two model searches and six classifiers.

The left columns of figure 5.4 present the results using the features selected by the CFS and the right columns the results using the features selected by the Relief F. Table 5.6 presents the values for each of the columns. The lowest value is achieved using 13-KNN and CFS with an AUC of 93.23 %. Looking at the SVMs results with the CFS subset, there is no clear conclusion on which classifier is the best, though the 3-SVM is slightly better, having an AUC of $96.43 \pm 0.5\%$ versus an

96.3±0.6% and 96.20±0.5% from the 2-SVM and 1-SVM classifiers, respectively.

Table 5.6: Classification results for the Radiologists’ data, presented as the mean and standard deviation of 50 AUC (%) values, for 12 features selected by two model searches and six classifiers.

AUC (%)	13-KNN	15-KNN	17-KNN	1-SVM	2-SVM	3-SVM
CFS	93.2 ± 0.8	93.5 ± 0.9	94.1 ± 0.7	96.2 ± 0.5	96.3 ± 0.6	96.4 ± 0.5
Relief F	94.7 ± 0.7	94.4 ± 0.8	94.4 ± 0.7	96.0 ± 0.6	96.3 ± 0.6	96.2 ± 0.6

5.3.2 Diagnostic data. Procedure and results

For the Diagnostic dataset, due to the small amount of nodules, the entire dataset served as input to select the best features and to perform the classification. Similarly to the procedure when using the Radiologists’ data, feature selection was performed by 10-fold cross validation and the results were obtained for each method. For the CFS, the features that appeared in more than 50% of the folds were selected and ranked according to the number of times they were selected (100%-50%). Again, this threshold allowed to select highly ranked features that would sufficiently represent the nodules in the dataset. For the Relief F, the k value was set to 5 and the highest ranked features were chosen until the selected total was equal to the number of features selected by the CFS. The selected features for both methods are presented in table 5.5 and are ordered according to relevance.

Table 5.7: Selected features using CFS and Relief F for the Radiologists’ data and Diagnostic data. (For Gabor, the features are described as $Si-j Ok$, where S means scale, i is the total number of scales, j the scale of the filter, O means orientation and k is the number of the corresponding orientation. Other definitions: *std* - Standard Deviation; r - Radius)

	CFS	Relief F
Diagnostic data	1 - std Gabor S5-5 O4 2 - <i>Mean_Intensity_Overlap</i> ($r = 3$) 3 - Compactness2 4 - std Laws S5R5\R5S5 5 - mean Laws E5R5\R5E5	1 - <i>Sphericity_ratio3</i> 2 - <i>Mean_Intensity_Overlap</i> ($r = 1$) 3 - std Gabor S8-4 O8 4 - std Gabor S5-4 O8 5 - Sum Entropy (GLCM)

Again, and similarly to the subsets obtained from the Radiologists’ data, the majority of the features chosen by both methods were texture features. Both CFS and Relief F also selected one geometric feature and one intensity feature.

Six classifiers were set, three KNN based and three SVM based. The KNN classifiers were set with different k values, defined as 3, 5 and 7, and the SVM classifiers were defined with an exponential Kernel and parameters θ of 1, 2 and 3. The classification was performed by 10-fold cross-validation with 50 repetitions and the mean and standard deviation were calculated from the 50 values of AUC using 5 features. The results are presented in figure 5.5 in form of a bar chart.

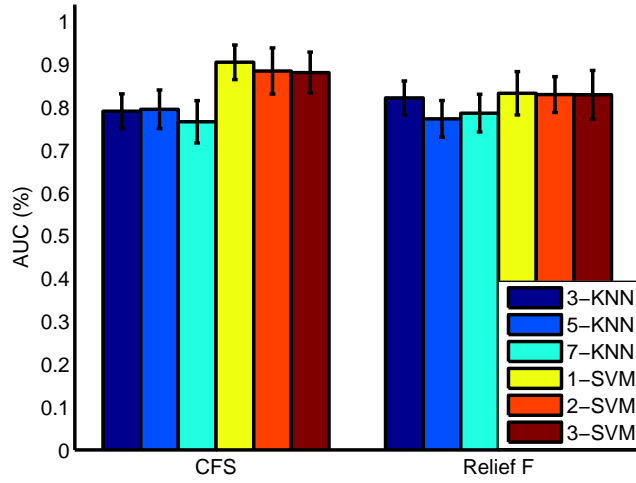


Figure 5.5: Classification results for the Diagnostic data, presented as AUC (%) value, for 5 features selected by two model searches and six classifiers.

The left columns of figure 5.5 present the results using the features selected by the CFS and the right columns the results using the features selected by the Relief F. Table 5.8 presents the values for each of the columns. The lower result is given by the 7-KNN and CFS with an AUC of 76.6%. The best results, again, are given by the SVM classifiers and CFS, particularly the 1-SVM, presenting an AUC value of $90.5 \pm 4.0\%$.

Table 5.8: Classification results for the Diagnostic data, presented as the mean and standard deviation of 50 AUC (%) values, for 5 features selected by two model searches and six classifiers.

AUC (%)	3-KNN	5-KNN	7-KNN	1-SVM	2-SVM	3-SVM
CFS	79.1 ± 4.0	79.5 ± 4.5	76.6 ± 4.9	90.5 ± 4.0	88.5 ± 5.4	88.1 ± 4.8
Relief F	82.2 ± 3.9	77.3 ± 4.3	78.6 ± 4.4	83.3 ± 5.1	82.9 ± 4.2	82.9 ± 5.7

Classification results for the best performance from all 50 repetitions are shown in figures 5.6, 5.7 and 5.8. All 34 nodules are presented with the respective classification and the confidence of the classifier. The confidence is the amount of certainty that a classifier has on labelling a nodule as benign or malignant. The true labels are the columns, where the left column presents the benign nodules and right column the malignant. The contours give the classification results, where the red contours represent nodules classified as malignant and green contours nodules classified as benign. It is visible that all malignant nodules and 8 in 13 benign nodules were correctly classified, though the confidence is generally higher for the malignant.

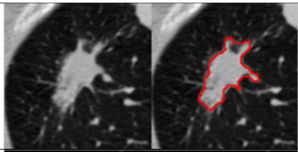
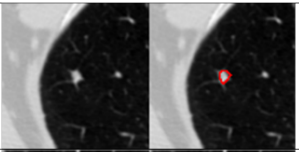
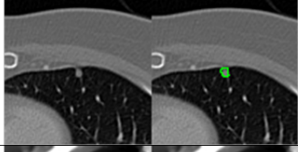
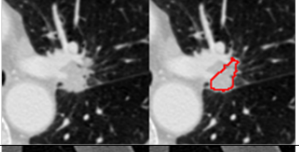
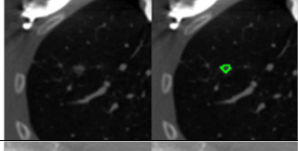
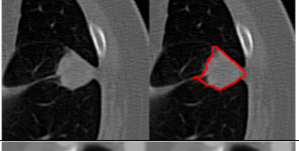
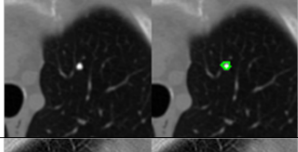
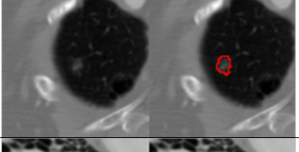
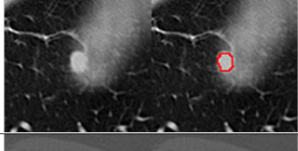
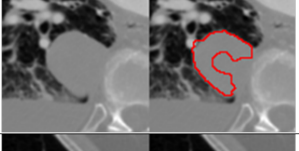
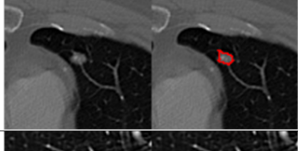
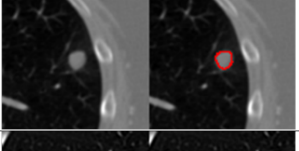
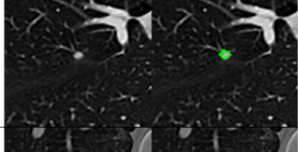
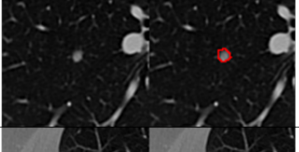
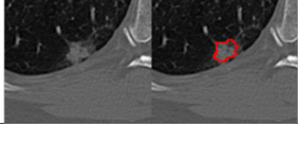
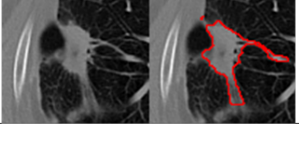
True Label	Benign		Confidence (%)	Case	Malignant		Confidence (%)
Case							
72			93,5	90			85,7
178			75,5	138			88,3
180			72,4	163			94,6
212			72,1	165			87,1
234			88,1	166			88,1
246			71,7	175			87,9
247			75,7	184			76,3
257			80,6	191			91,5

Figure 5.6: Classification results for the Diagnostic dataset. The columns are the true labels of the nodules, where the left column presents the benign nodules and right column the malignant. The red contours represent nodules classified as malignant and green contours nodules classified as benign.

268		79,3	193		79,9
270		55,48	202		80,1
275		55,24	210		70,4
277		68,5	242		92,7
283		88,4	258		80,4
			258		86,8
			265		94,9
			266		93,3

Figure 5.7: Classification results for the Diagnostic dataset. Classification results for the Diagnostic dataset. The columns are the true labels of the nodules, where the left column presents the benign nodules and right column the malignant. The red contours represent nodules classified as malignant and green contours nodules classified as benign.

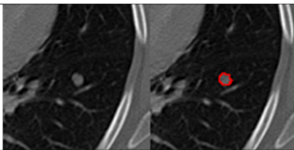
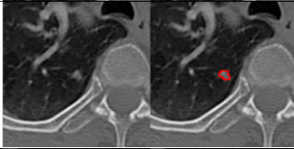
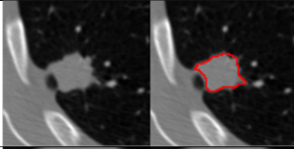
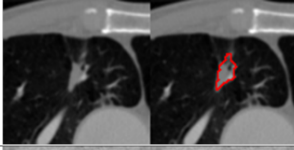
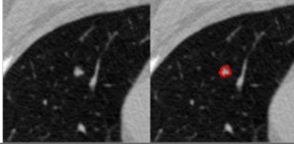
True Label	Benign			Malignant	
Case		Confidence (%)	Case		Confidence (%)
			271		82,8
			272		88,8
			470		87,7
			470		76,9
			580		79,5

Figure 5.8: Classification results for the Diagnostic dataset. Classification results for the Diagnostic dataset. The columns are the true labels of the nodules, where the left column presents the benign nodules and right column the malignant. The red contours represent nodules classified as malignant and green contours nodules classified as benign.

5.3.3 Inter-datasets validation

This section presents the classification results using both datasets for cross-validation. Here the Radiologists' data and Diagnostic data are first used as training and testing, respectively, as for now referred as *test 1*, and secondly as testing and training, respectively, as for now referred as *test 2*. Because the origin of the GTs of both datasets is very different, we are expecting poor results, due to the Radiologists' data being . Either way, *test 1* is mainly done to assess if there is any classification capability on using the GT of the radiologists and the corresponding selected features to distinguish benign from malignant nodules of the Diagnostic data. *test 2* is done to assess if the Diagnostic data can be used as a training model that would give similar classification performance as the radiologists. Also, for *test 1*, only Ground Truth 1 is used to train the classifier, as it is the one who presents the highest level of agreement between radiologists. In *test 2*, however, all three

Ground Truths are used for testing. The block diagrams of the tests are presented in figure 5.9.

Giving the classification results obtained from the Radiologists' dataset and Diagnostic dataset, it was concluded that the better option was to use the 1-SVM with an exponential Kernel classifier and a CFS feature selector in both tests, as the combination gave the best overall results for both datasets. However, the number of features in the subsets was different. For *test 1* only 5 features were used and for *test 2* 12 features were used.

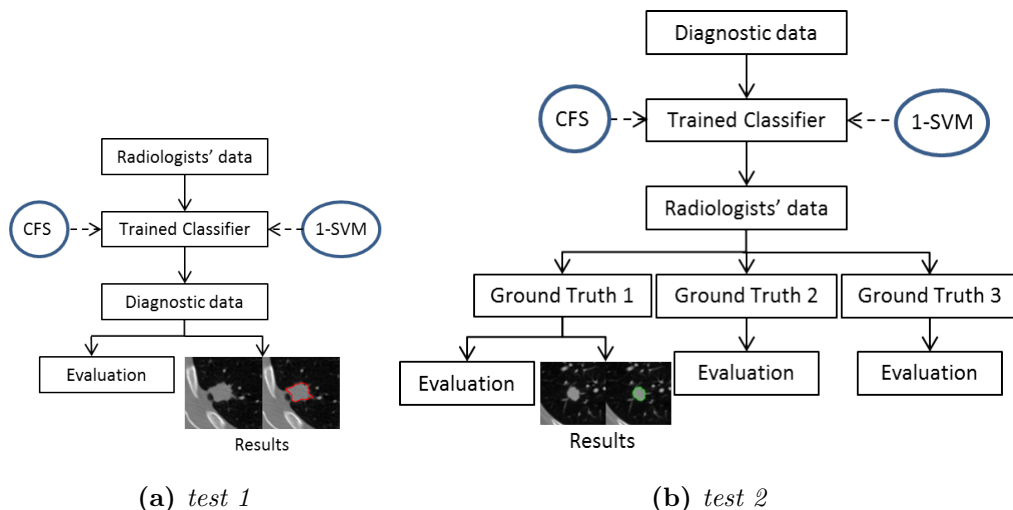


Figure 5.9: Block diagrams of *test 1* and *test 2*.

The results are presented in form of confusion matrices for better understanding. A confusion matrix simply gathers in one table the number of nodules that were correctly classified, meaning that the *Estimated Label* is equal to the True Label (TP and TN), and the number of nodules that were incorrectly classified, meaning that the *Estimated Label* is different from the True Label (FP and FN). Table 5.9 is the confusion matrix for *test 1* and tables 5.10, 5.11 and 5.12 are the confusion matrices for *test 2* using Ground Truth 1, 2 and 3, respectively. The sensitivity and specificity for both tests are also presented in table 5.13. Here, the sensitivity is presented as correctly classified malignant nodules and specificity as correctly classified benign nodules.

Table 5.9: Confusion matrix of *test 1*.

		<i>Estimated Labels</i>		Totals
		Malignant	Benign	
<i>True Labels</i>	Malignant	11	10	21
	Benign	4	9	13
	Totals	15	19	34

Table 5.10: Confusion matrix of *test 2* using Ground Truth 1.

		<i>Estimated Labels</i>		
		Malignant	Benign	Totals
<i>True Labels</i>	Malignant	106	71	177
	Benign	65	56	121
	Totals	171	127	298

Table 5.11: Confusion matrix of *test 2* using Ground Truth 2.

		<i>Estimated Labels</i>		
		Malignant	Benign	Totals
<i>True Labels</i>	Malignant	183	19	202
	Benign	76	129	205
	Totals	259	148	407

The overall results show a low performance on both tests. The highest result for *test 2* is achieved testing in Ground Truth 2, but as seen in table 5.13, it stays only at 51.7% for specificity and 59.4% for sensitivity, which is not significant. Interestingly, the lowest results come from Ground Truth 1 with a specificity of 46.3% and a sensitivity of 59.9%, though sensitivity is higher than any other and it is more important to correctly classify malignant nodules than benign. It was expected that due to the greater agreement between radiologists for Ground Truth 1, the results would present, at least, better results when comparing to the others, but this do not happens. For *test 1*, the results are also not very good, with specificity staying at 69.2% and sensitivity at 52.2%.

Classification results for some nodules with the corresponding confidence of the classifier are shown in figures 5.10, 5.11, 5.12 and 5.13. The columns present the true labels, where the left column is presents the benign nodules and right column the malignant, whereas the contours give the classification results, where the red contours represent nodules classified as malignant and green contours nodules classified as benign.

The confidence of the classifier presents valuable and clear information of what happens in the system and visually confirms what was stated previously. For *test 1*, it is visible in figures 5.10 and 5.11 that round, small nodules are labelled as benign with a high confidence, even if they are labelled incorrectly, and the same happens to big, spiculated/lobulated nodules. In figure 5.11, the confidence of the benign incorrect instances tend to be lower as nodules present rounder shapes, lower intensity and smaller sizes.

Similarly to *test 1*, it is visible in figure 5.12 that some round, small nodules are labelled as benign with a high confidence and the same happens to big, spiculated/lobulated nodules. However, due to the fact that the Diagnostic data as more variation in what concerns types of malignant and benign nodules, as seen in figure 5.11, some more standard nodules that would be classified as benign if

Table 5.12: Confusion matrix of *test 2* using Ground Truth 3.

		<i>Estimated Labels</i>		
		Malignant	Benign	Totals
<i>True Labels</i>	Malignant	225	50	275
	Benign	76	129	205
	Totals	301	179	480

Table 5.13: Performance of the CAD system in form o Sensitivity and Specificity for *test 1* and *test 2*.

	<i>Performance (%)</i>			
	Radiologists' data			Diagnostic data
	Ground Truth 1	Ground Truth 2	Ground Truth 3	
Sensitivity	46.3	51.7	51.7	69.2
Specificity	59.9	59.4	57.8	52.2

presenting small size, high intensity and round shape, are in fact classified as malignant. This is clearly visible in figure 5.13. The same thing happens to big, spiculated nodules.

5.4 Evaluation and discussion of results

5.4.1 Evaluation of the radiologists' dataset classification

Before any evaluation on the performance of the classification and discussion of the results, we must first look at the features selected by both CFS and Relief F. The majority of features chosen by both methods were texture features, though CFS also selected two shape features and Relief F three intensity features. In a particular analysis, the CFS selected a great number of GLCM and Laws features. The inclusion of Volume and Compactness1 by CFS is coherent, as radiologists tend to consider small or round nodules as benign and big or spiculated as malignant. The lack of intensity features can be give two indications, one is that CFS finds intensity information similar for both malignant and benign nodules, and the other is that the intensity features can be simply redundant to the problem if there are already corresponding high correlated features in the set. Relief F, however, selected features that focus on the center calcification of the nodules, which are also good at predict the malignancy of the nodules. It also includes a lot of GLCM features (8 in 12 features), implying that the GLCM has great discrimination capacity.

The AUC values for the Radiologists' dataset were already presented in section 5.3. To aid the evaluation, the ROC curves for the subset of the CFS are presented in figure 5.14a and in figure 5.14b the ROC curves for the subset of the Relief F.

The AUC values for all classifiers and subsets for this dataset were already concluded to be high, though there is no considerable difference between them.

Correctly Classified					
Benign		Confidence (%)	Malignant		Confidence (%)
		99,3			62,9
		96,5			93,8
		93,8			90,8
		80,4			89,5

Figure 5.10: Examples of correctly classified nodules from *test 1*.

The lowest value is achieved using the 13-KNN and CFS with an AUC of 93.23 %. The 3-SVM is the one who presents the best results having an AUC of 96.43%. The SVM based classifiers outperform all the KNN classifiers using either of the subsets, though the results are slightly higher for the CFS subset.

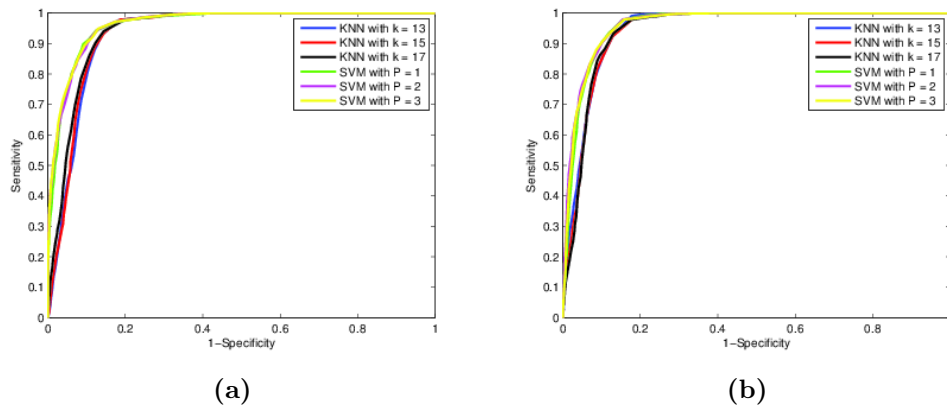


Figure 5.14: ROC curves of the classification performances using the Radiologists' data for six classifiers. a) Results presented for 12 features selected by the CFS algorithm. b) Results presented for 12 features selected by the Relief-F algorithm.

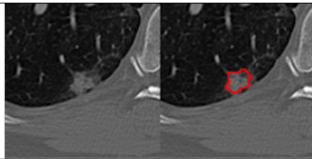
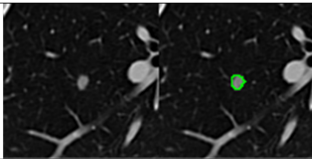
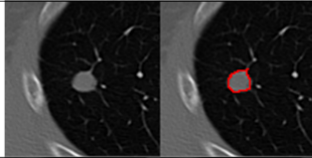
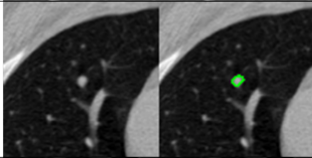
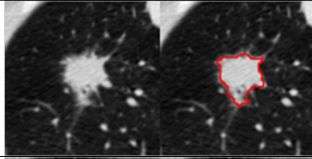
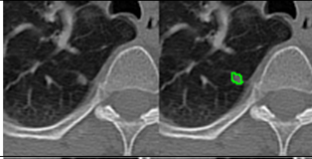
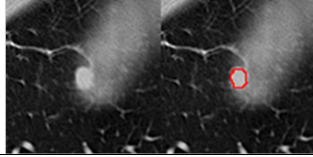
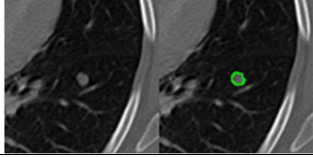
Incorrectly Classified			
Benign	Confidence (%)	Malignant	Confidence (%)
	72,1		97,7
	54,4		99,39
	72,1		97,7
	54,4		99,39

Figure 5.11: Examples of incorrectly classified nodules from *test 1*.

The ROC curves in figures 5.14 support the notion that the SVMs are better than KNN classifiers and that their performance is very similar. Although this is true, the subset from Relief F improves the results of the KNN and decreases the performance of the SVMs.

The overall results show that it is possible to build a lung nodule classification system, similar to the radiologists assessment with high performance. The results can be improved by including a wrapper based algorithm coupled with a classifier to select the features more efficiently. Additionally, an independent database must be used to increase the validation of the system.

5.4.2 Evaluation of the diagnostic dataset classification

Similarly to the Radiologists' dataset, the majority of features chosen by both methods were texture features. Both CFS and Relief F also selected one geometric feature and one intensity feature. The intensity features give information about the central intensity of the nodules, which is a good indicator of the presence of calcification in that region. Both *Compactness2* and *Sphericity_ratio3* give information about the roundness of the nodules, so is natural that they are selected based on the fact that many round nodules are benign.

Again, and similarly to the classification results using the Radiologists' data, the SVM classifiers provide better results using the subset obtained from the CFS

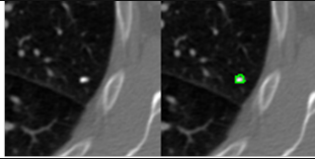
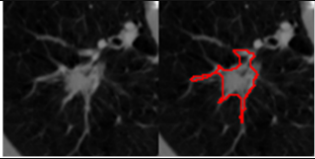
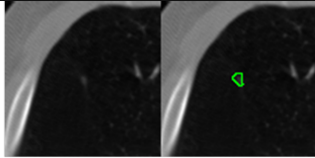
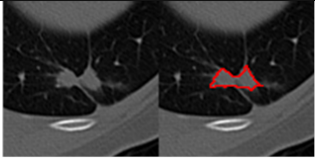
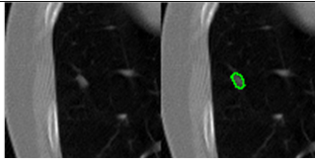
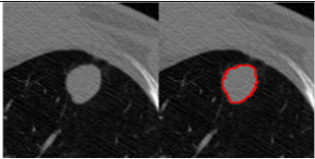
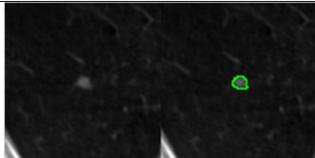
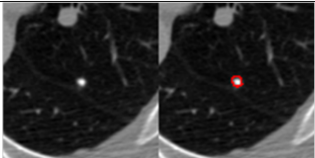
Correctly Classified					
Benign		Confidence (%)	Malignant		Confidence (%)
		81,1			81,5
		94,9			91,0
		92,4			90,8
		55,6			51,6

Figure 5.12: Examples of correctly classified nodules from *test 2*, Ground Truth 1.

than from the subset obtained from the Relief F, and the inverse happens to the KNN classifiers. The ROC curves in figures 5.15a,b support this notion as the one in the left shows that the SVM classifiers using CFS are clearly better than the KNN classifiers, and on the right, for Relief F, the curves are close to each other and further, the 3-KNN outperforms all the other classifiers when sensitivity is lower than 60%.

The best result is given by the 1-SVM, presenting an AUC value of 90.5%. This result is better than most of the results provided in literature that use the AUC value as performance measure, being only inferior to the one presented by Haifeng *et al.* [23], which achieved an AUC value of 91%.

Incorrectly Classified			
Benign	Confidence (%)	Malignant	Confidence (%)
	97,9		54,6
	63,5		53,1
	94,3		72,9
	79,1		64,6

Figure 5.13: Examples of incorrectly classified nodules from *test 2*, Ground Truth 1.

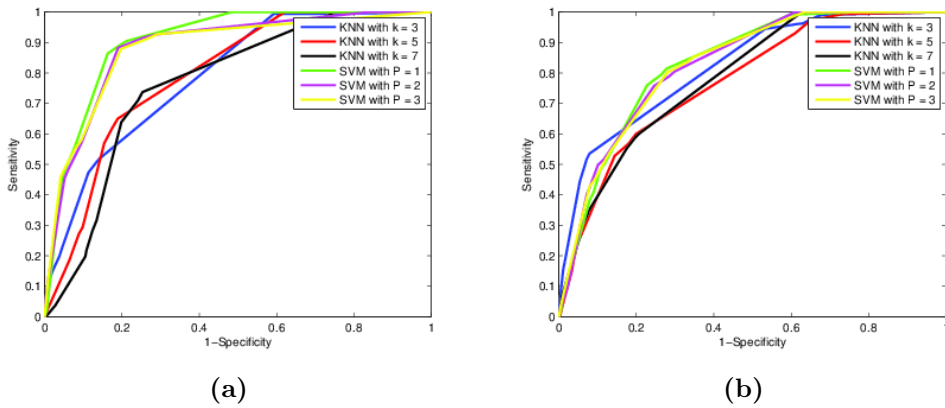


Figure 5.15: ROC curves of the classification performances using the Diagnostic data for six classifiers. a) Results presented for 5 features selected by the CFS algorithm. b) Results presented for 5 features selected by the Relief-F algorithm.

The overall results show that it is possible to build a lung nodule classification system with performances close to a biopsy or a surgical procedure. However, the database is small presenting a low diversity of nodules. To better validate the system, a bigger, independent database, with a large variety of nodules, must be used. Additionally, the results can be improved by including a wrapper based

algorithm coupled with a classifier to select the features more efficiently. Other classifiers can also be implemented such as the ANN, as they are being frequently used in the last years.

5.4.3 Evaluation of the inter-datasets classification

The results from the inter-dataset classifications were not good, but, as mentioned before, this was expected. To understand why, it is important to look at the performance of the radiologists in correctly classifying the Diagnostic data. In table 5.14 are presented the radiologists' performance versus the CAD system performance for only 19 nodules (8 benign and 11 malignant), as 15 of the 34 nodules are labelled as *Indeterminate* by radiologists. The results show that the radiologists managed a performance of 60.0% for specificity and 55.6% for sensitivity, which is a very low predicament, and this is not counting with the *Indeterminate* nodules. If the performance of the radiologists is low, than it is normal that training a classifier using a GT that is far from a true diagnose will not give any good discriminant capacity when classifying the Diagnostic data. As the system tends to classify the nodules as seen by radiologists, some small, round, but malignant nodules are classified as benign by the system and the inverse happens to big, somehow spiculated, benign nodules. Also, the *Indeterminate* nodules are used in both tests 1 and 2, presenting an even further obstacle to a good classification, because they present both common benign and malignant features. If only non-Indeterminate nodules are considered for performance analysis, the CAD system increases its performance achieving 87.5% for specificity and 55.0% for sensitivity for these nodules and using the Ground Truth 1 for training, versus a specificity of 69.2% and sensitivity of 52.2% for all nodules and using Ground Truth 1 for training. This allows two conclusions. First, many of the benign nodules using Ground Truth 1 are similar to the 8 benign nodules found in the Diagnostic data. Secondly, training the classifier with different opinions and examples gave a better performance when comparing to the majority opinion of four radiologists.

Table 5.14: Performance of the CAD system versus the performance of the radiologists. Only 19 nodules nodules were used for comparison as they were the only ones to be labelled as non-Indeterminate (1 and 5 labels) by radiologists.

		<i>Performance %</i>		
		Diagnostic data		
		Specificity	Sensitivity	Accuracy
Radiologists		60.0	55.6	57.9
CAD system		87.5	55.0	68.4

Both *test 1* and *test 2* are affected by this poor capacity of the radiologists in correctly classifying the nodules, or at least in classifying the nodules of the Diagnostic data, which is a small sample. But if we would calculate their Accuracy (the correctly classified instances) for all 34 nodules, the performance would be

26.47% (1 in 4 nodules are correctly classified), whereas the CAD system would achieve 58.82%. Considering only 19 nodules, the radiologists' Accuracy increases to 57.9% and the CAD system's Accuracy is 68.4%. This means that, either case, the CAD system still performs better than radiologists.

Chapter 6

Conclusions and Future Work

This dissertation presents an automatic CAD system for lung nodule classification in CT images. This system determines the malignancy of a nodule using information retrieved solely from the ROI of the nodules. It was designed following two different learning strategies:

- A system that provides a lung nodule classification learned from radiologists/opinions.
- A system that provides a lung nodule classification learned from real biopsy, surgery exam or follow up during several years.

For this purpose, several studies were performed to optimize the system, namely, an analysis of the LIDC-IDRI database in what concerns the agreement of the radiologists' segmentations and intensity ranges of the images. It was based on the development of a nodule segmentation algorithm by Novo *et al.* [75]. Additionally, the CAD system was built using a set of optimal features and an exponential SVM classifier of order 1.

Concerning the analysis of the LIDC-IDRI database, the Jaccard index was used to determine the agreement between the segmentations of the radiologists and assess if the segmentations could be used as ground truth. This index calculates the rate of voxels that are common to two areas and the agreement is high if the rate is close to 1. For our purpose, the rate of agreement between every radiologists' segmentations was obtained and a mean value for every radiologist calculated. The results showed a low inter-agreement between radiologists for both solid and GGO nodules. In the solid type, the lowest agreement is verified for small nodules. The highest agreement is seen in big nodules achieving only 71%.

Due to the low accuracy verified in the radiologists' segmentations, a novel method for lung nodule segmentation was developed. The work from Novo *et al.* [75] was adapted here to perform nodule segmentation using two methods to generate two different nodule masks. Additionally, an analysis on the segmentations was performed by assessing the agreement between them, and the radiologists' segmentations using Jaccard index. The Bland-Altman method was used

with the Jaccard Index to compare the agreement between the methods and the radiologists' segmentations. The results showed that the combination using both methods gives higher agreement than when used separately or when compared with the agreement among radiologists. Nevertheless, a study on the optimal threshold values for both methods must be made so better segmentations are obtained. New improvements are needed for justa-vascular nodules. Also, the method failed in some justa-pleural nodules and nodules crossed by airways, so that must be addressed as well.

Just like a good nodule segmentation is vital to perform a reliable feature measurement, the intensity ranges of the images can in fact be a problem if there is great variability. In fact, the database is composed by images obtained from different hospitals and CT scanners so a study was conducted to see what were the implications of that variability, particularly in the feature measurement stage. Five representative images from the database with different intensity ranges were chosen for evaluation and the histograms of the unprocessed images and the histograms of the normalized images were obtained for comparison purposes. The results showed different intensity ranges from scan to scan, but that was not the case for the nodule's ROI. Ultimately, the conclusion was that feature measurement should be done in the raw, unprocessed images.

Regarding the system for lung nodule classification, 293 shape, intensity and texture features were computed using both the nodules masks and the ROI of the nodules. The features were defined in order to match and characterize the common radiologic features described in literature. Two different datasets were used to find the best combination of feature selection method and classifier. The first dataset is the Ground Truth 1 of the Radiologists' data and the other the Diagnostic data, obtained from either biopsy, surgery or follow up. To eliminate redundant and irrelevant information, two feature selection methods were used, the CFS and Relief F. Six different classifiers were evaluated for both datasets.

The best performance was achieved using a first order SVM with an exponential kernel for a subset of 12 features using the Radiologists' data, Ground Truth 1, and 5 features for the Diagnostic data, both obtained from the CFS, respectively. Most of the selected features were texture based, being in agreement with what is found in literature. Using the Radiologists' data, the system achieved an AUC value of 96.2 ± 0.5 %, which is a good performance. Using the Diagnostic data, an AUC value of 90.5 ± 4.0 % was obtained being the second best result found in literature, though the sample is very small and additional validation is necessary. In future work, wrapper based feature selection methods should be tested, as better results can be achieved by guiding the selection using classifiers. Also, an ANN should be implemented for classification comparison as they have been used in the last years with good results.

An additional classification was performed to see if one dataset had some sort of ability on predicting the nodule's malignancy of the other. Two tests were designed, where a cross-validation procedure was implemented. Here the Radiolo-

gists' data and Diagnostic data were first used as training and testing, referred as *test 1*, and secondly as testing and training, respectively, referred as *test 2*. The evaluation was performed by calculating the sensitivity and specificity measures. *test 1* achieved a specificity of 69.2% and sensitivity of 52.2% and *test 2* achieved the best results for Ground Truth 2 with a specificity of 51.7% and a sensitivity of 59.4%. The results indicate that the nodules presenting common radiologic characteristics of malignancy or benignity were labelled as such by the system. The global performance was low as many nodules in the Diagnostic data have many mixed characteristics and some that resemble as benign are in fact malignant and vice versa. However, the performance of the CAD system surpassed the one from the radiologists. The same low performance was observed in *test 2* and the same reasons appointed to *test 1* are applied here. Because some of the nodules used for training have uncommon visual characteristics for benign or malignant nodules, the system incorrectly labels the nodules as seen from the radiologists.

For future work, it is of the most interest to cooperate with radiologists in an active learning procedure to increase the performance of the system. This means that the results from classification are reviewed by radiologists, which in turn make their assessment, so new and improved characteristics are measured and more accurate classifiers are implemented. Additionally, the system can be improved to further classify partially solid and sub-solid nodules.

Appendices

Appendix A

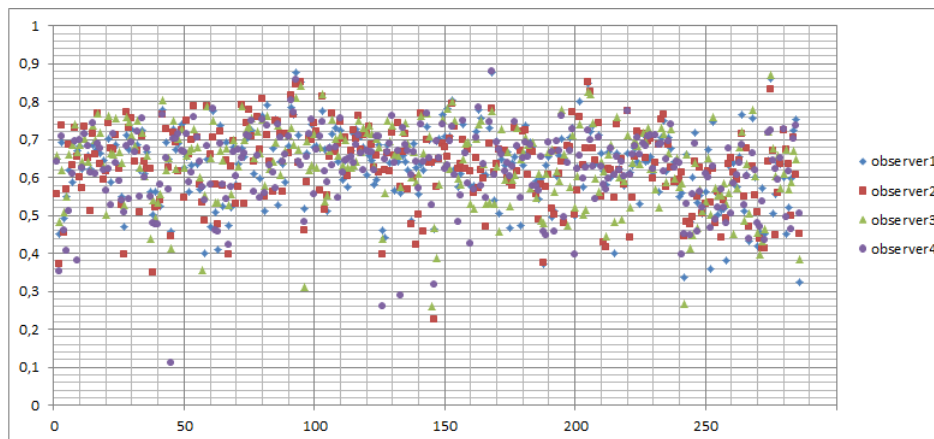


Figure A.1: It is presented, for each case, the average Jaccard inter-observer agreement for the segmentation of small nodules between one and the other radiologists. In order to better distinguish the agreement between segmentations, the radiologists are observers 1, 2, 3 and 4, but that labelling does not represent the same radiologist in every case.

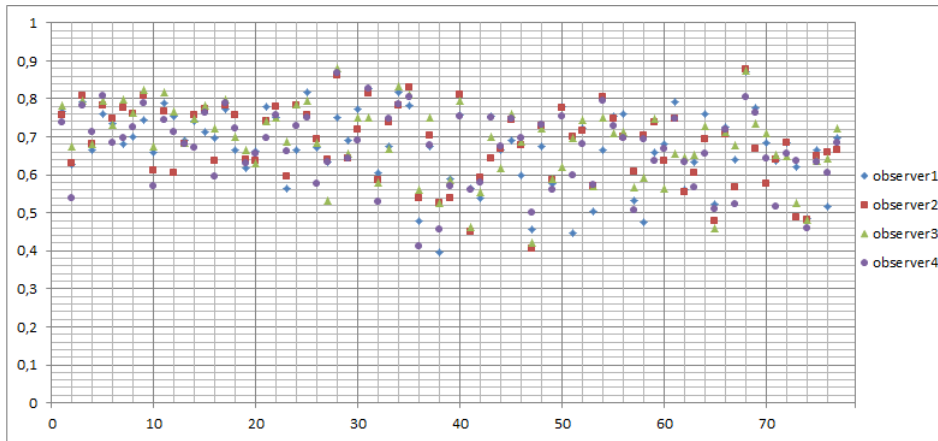


Figure A.2: It is presented, for each case, the average Jaccard inter-observer agreement for the segmentation of medium sized nodules between one and the other radiologists. In order to better distinguish the agreement between segmentations, the radiologists are observers 1, 2, 3 and 4, but that labelling does not represent the same radiologist in every case.

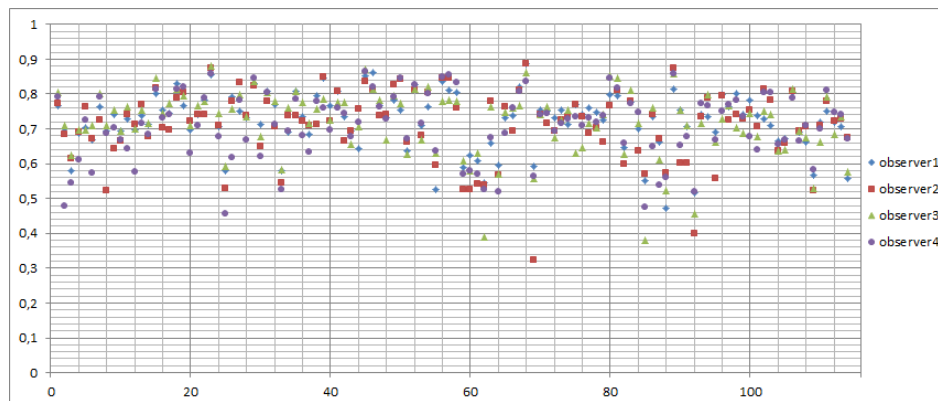


Figure A.3: It is presented, for each case, the average Jaccard inter-observer agreement for the segmentation of big nodules between one and the other radiologists. In order to better distinguish the agreement between segmentations, the radiologists are observers 1, 2, 3 and 4, but that labelling does not represent the same radiologist in every case.

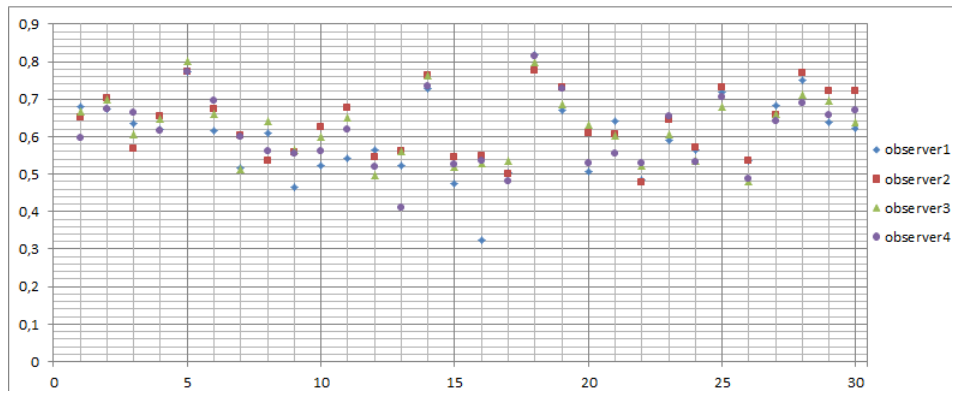


Figure A.4: It is presented, for each case, the average Jaccard inter-observer agreement for the segmentation of sub-solid nodules between one and the other radiologists. In order to better distinguish the agreement between segmentations, the radiologists are observers 1, 2, 3 and 4, but that labelling does not represent the same radiologist in every case.

Appendix B

		Radiologists					
		1	2	3	4	5	Total
GT1	1	50	4	0	0	0	54
	2	24	27	12	4	0	67
	3	5	89	116	33	2	245
	4	1	16	24	47	37	125
	5	0	2	3	11	38	54
	Total	80	138	155	95	77	

Figure B.1: Number of nodules for each label of Ground Truth 1 versus the labels from radiologists.

		Radiologists					
		1	2	3	4	5	Total
GT2	1	55	9	3	0	2	69
	2	24	81	23	8	0	136
	3	0	24	96	16	2	138
	4	1	22	30	60	37	150
	5	0	2	3	11	38	54
	Total	80	138	155	95	77	

Figure B.2: Number of nodules for each label of Ground Truth 2 versus the labels from radiologists.

		Radiologists					
		1	2	3	4	5	Total
GT3	1	55	9	6	0	2	72
	2	24	93	58	8	0	183
	3	0	0	13	0	0	13
	4	1	33	71	76	37	218
	5	0	3	7	11	38	59
	Total	80	138	155	95	77	

Figure B.3: Number of nodules for each label of Ground Truth 3 versus the labels from radiologists.

Appendix C

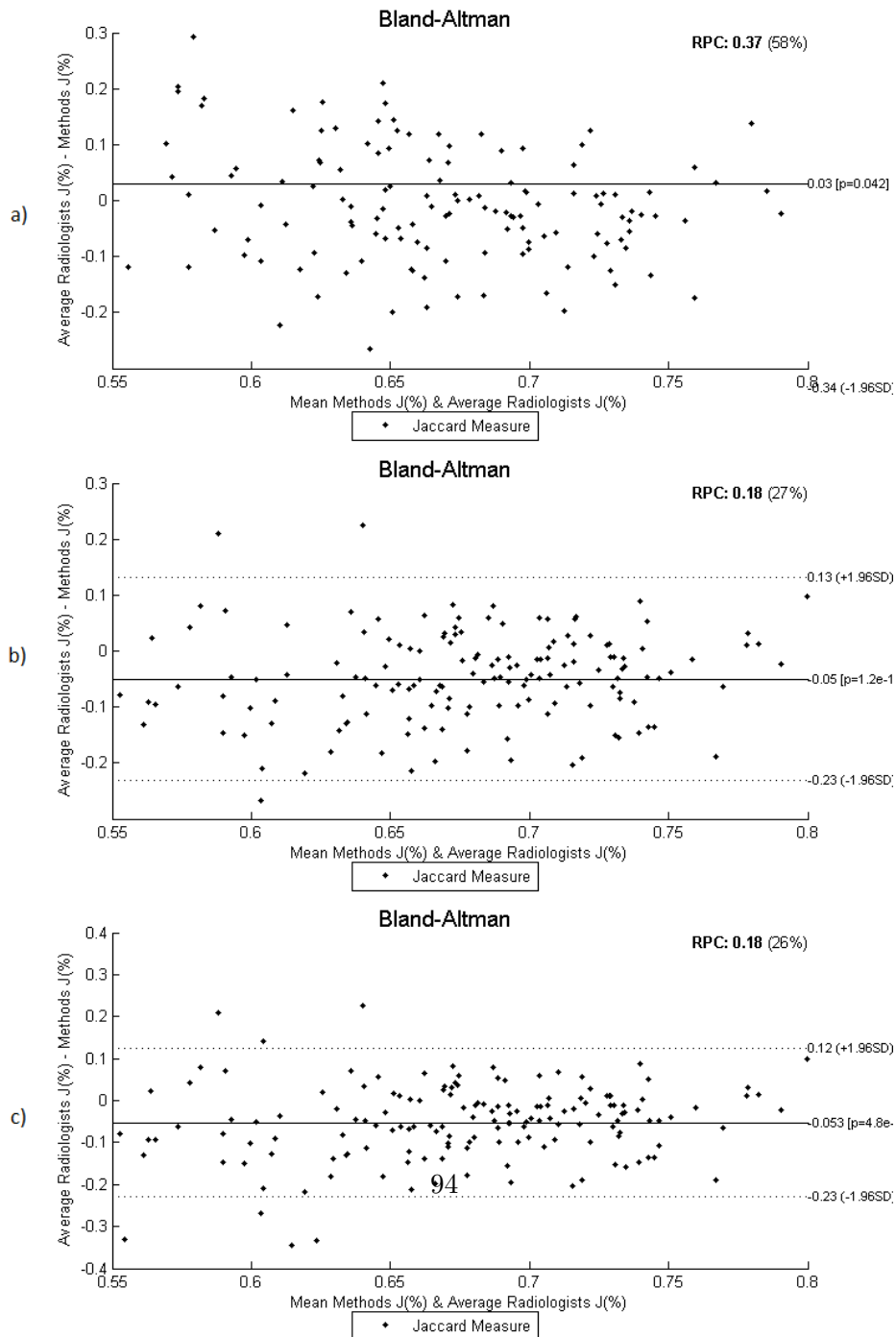


Figure C.1: Bland-Altman results for small nodules. a) Murphy's method. b) Krissian's method. c) Combination of both.

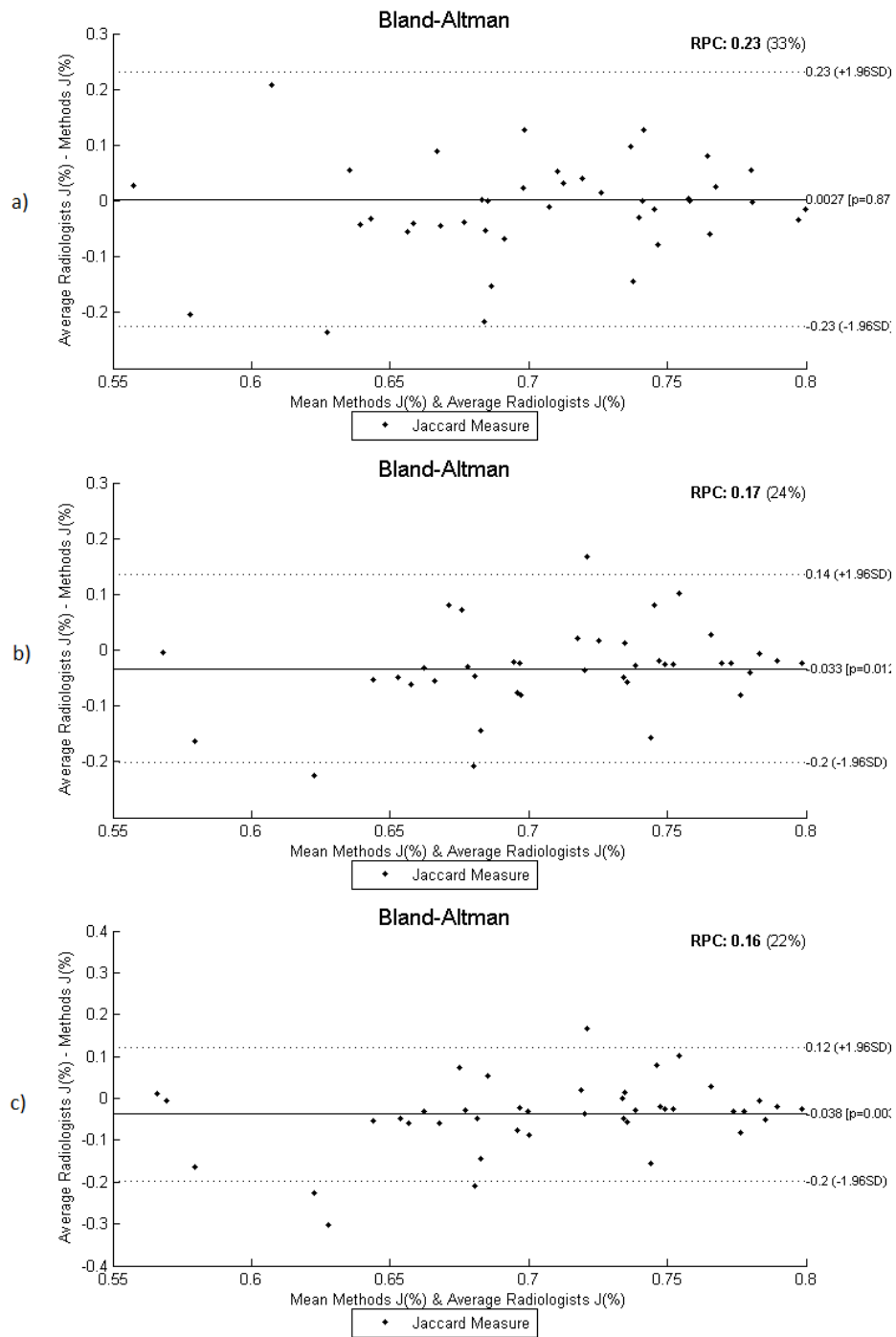


Figure C.2: Bland-Altman results for medium sized nodules. a) Murphy's method. b) Krissian's method. c) Combination of both.

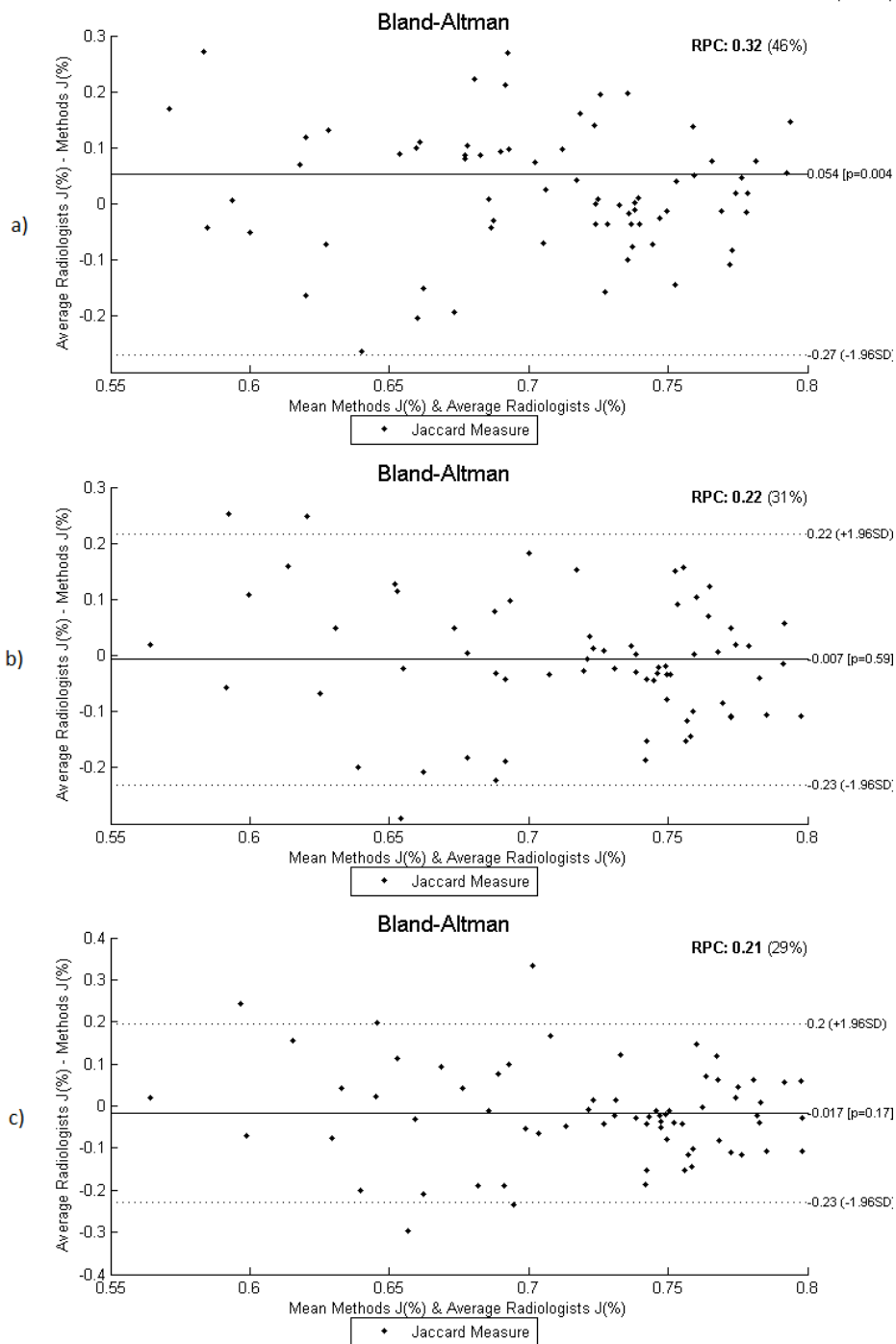


Figure C.3: Bland-Altman results for large nodules. a) Murphy's method. b) Krissian's method. c) Combination of both.

Bibliography

- [1] C. J. Breads Moore and N. J. Screamon, "Classification, staging and prognosis of lung cancer," *European Journal of Radiology*, vol. 45, pp. 8–17, 2003.
- [2] I. A. for Research on Cancer (IARC), "Latest world cancer statistics," World Health Organization, Press Release 223, December 2013.
- [3] J. Erasmus, J. Connolly, P. McAdams, and V. Roggli, "Solitary pulmonary nodules: Part i. morphologic evaluation for differentiation of benign and malignant lesions," *RadioGraphics*, vol. 58, pp. 20–43, 2000.
- [4] *Harrison's Principles of Internal Medicine, Eighteenth Edition*. McGraw-Hill Professional, 2011, ch. Neoplasms of the Lung, pp. 1420–1421.
- [5] B. van Ginneken, *Computer-Aided Diagnosis in Thoracic Computed Tomography*, 2008, vol. 12, pp. 11–22•.
- [6] T. W. Way, "Computer-aided diagnosis of pulmonary nodules in thoracic computed tomography," Ph.D. dissertation, The University of Michigan, 2008.
- [7] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, 2005.
- [8] T. Way, H. Chan, L. Hadjiiski, B. Sahiner, A. Chughtai, T. Song, C. Poopat, J. Stojanovska, L. Frank, A. Attili, N. Bogot, P. Cascade, and E. Kazerooni, "Computer-aided diagnosis of lung nodules on ct scans: Roc study of its effect on radiologists' performance," *Academic Radiology*, vol. 17, no. 3, pp. 323–332, 2010.
- [9] K. N. V. Philips. (2014) Visualization software. [Online]. Available: <http://www.healthcare.philips.com>
- [10] G. S. Healthcare, "Syngo lungcare ct and syngo lung cad," 2009. [Online]. Available: <http://www.healthcare.siemens.com>
- [11] G. E. Company. (2014) Lung vcar. [Online]. Available: <http://www.gehealthcare.com>

- [12] K. Murphy, B. van Ginneken, A. Schilham, B. de Hoop, H. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification," *Elsevier-Medical Image Analysis*, vol. 13, pp. 757–770, 2009.
- [13] K. Krissian, G. Malandain, N. Ayache, R. Vaillant, and Y. Troussel, "Model-based detection of tubular structures in 3d images," *Computer vision and image understanding*, vol. 80, no. 2, pp. 130–171, 2000.
- [14] X. Ye, X. Lin, J. Dehmeshkia, G. Slabaugh, and G. Beddoe, "Shape-based computer-aided detection of lung nodules in thoracic ct images," *IEEE-Transactions on Biomedical Engineering*, vol. 56, no. 10, 2009.
- [15] U. D. of Health and H. Services. Lidc-idri. [Online]. Available: <http://imaging.cancer.gov/informatics/lidcidri>
- [16] F. Li, S. Sone, H. Abe, H. MacMahon, and K. Doi, "Malignant versus benign nodules at ct screening for lung cancer: Comparison of thin-section ct findings," *Radiology*, vol. 233, pp. 793–798, 2004.
- [17] D. Xu, H. van der Zaag-Loonen, M. Oudkerk, Y. Wang, R. Vliegenthart, E. Scholten, J. Verschakelen, M. Prokop, H. de Koning, and R. van Klaveren, "Smooth or attached solid indeterminate nodules detected at baseline ct screening in the nelson study: Cancer risk during 1 year of follow-up," *Radiology*, vol. 250, no. 1, pp. 264–272, 2009.
- [18] S. Brandman and J. Ko, "Pulmonary nodule detection, characterization, and management with multidetector computed tomography," *Journal of Thoracic Imaging*, vol. 26, no. 2, pp. 90–105, 2011.
- [19] A. Farag, S. Elhabian, J. Graham, A. Farag, and R. Falk, "Toward precise pulmonary nodule descriptors for nodule type classification," *Springer-Verlag Berlin Heidelberg 2010*, p. 626–633, 2010.
- [20] S. Diciotti, G. Picozzi, M. Falchini, M. Mascalchi, N. Villari, and G. Valli, "3d segmentation algorithm of small lung nodules in spiral ct images," *IEEE-Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, 2008.
- [21] F. Zhang, W. Cai, Y. Song, M.-Z. Lee, S. Shan, and D. D. Feng, "Overlapping node discovery for improving classification of lung nodules," *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 5461–5464, 2013.
- [22] A. El-Baz, G. M. Beache, G. Gimel'farb, K. Suzuki, K. Okada, A. Elnakib, A. Soliman, and B. Abdollahi, "Computer-aided diagnosis systems for lung cancer: challenges and methodologies," *International journal of biomedical imaging*, vol. 2013, 2013.

- [23] H. Wu, T. Sun, J. Wang, X. Li, W. Wang, D. Huo, P. Lv, W. He, K. Wang, and X. Guo, "Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography," *Society for Imaging Informatics in Medicine*, vol. 26, p. 797–802, 2013.
- [24] H. Wang, X. Guo, Z. Ji, H. Li, Z. Liang, K. Li, and Q. He, "Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of ct image," *European Journal of Radiology*, vol. 74, p. 124–129, 2010.
- [25] S. Iwano, T. Nakamurab, Y. Kamiokac, M. Ikeda, and T. Ishigaki, "Computer-aided differentiation of malignant from benign solitary pulmonary nodules imaged by high-resolution ct," *Computerized Medical Imaging and Graphics*, vol. 32, pp. 416–422, 2008.
- [26] T. Messay, R. C. Hardie, and T. R. Tuinstra, "Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset," *Medical Image Analysis*, vol. 22, no. 1, pp. 48 – 62, 2015.
- [27] Y. Zhu, Y. Tan, Y. Hua, M. Wang, G. Zhang, and J. Zhang, "Feature selection and performance evaluation of support vector machine (svm)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography," *Journal of Digital Imaging*, vol. 23, no. 1, pp. 51–65, 2010.
- [28] M. Lee, L. Boroczky, K. Stasik, A. Cann, Alain, Borczuk, S. Kawut, and C. Powell, "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction," *Artificial Intelligence in Medicine*, vol. 50, pp. 43–53, 2010.
- [29] E. Silva, A. Silva, A. Paiva, and R. Nunes, "Diagnosis of lung nodule using moran's index and geary's coefficient in computerized tomography images," *Pattern Analysis and Applications*, vol. 11, pp. 89–99, 2008.
- [30] T. Way, L. Hadjiiski, B. Sahiner, H.-P. Chan, P. Cascade, E. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided diagnosis of pulmonary nodules on ct scans: Segmentation and classification using 3d active contours," *Medical Physics*, vol. 37, no. 7, p. 2323–2337, 2006.
- [31] F. Zhang, Y. Song, W. Cai, M.-Z. Lee, Y. Zhou, H. Huang, S. Shan, M. J. Fulham, and D. Feng, "Lung nodule classification with multilevel patch-based context analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, 2014.

- [32] H. Chen, Y. Xu, Y. Ma, and B. Ma, “Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on ct images,” *Academic Radiology*, vol. 17, pp. 595–602, 2010.
- [33] T. Way, B. Sahiner, H. Chan, L. Hadjiiski, P. Cascade, A. Chughtai, N. Bogot, and E. Kazerooni, “Computer-aided diagnosis of pulmonary nodules on ct scans: Improvement of classification performance with nodule surface features,” *Medical Physics*, 2009.
- [34] C. Yeh, C. Lin, M. Wub, C. Yen, Jen, and Wang, “A neural network-based diagnostic method for solitary pulmonary nodules,” *Neurocomputing*, vol. 72, pp. 612–624, 2008.
- [35] S. A. III, M. Altman, J. Wilkie, S. Sone, F. Li, K. Doi, and A. Roy, “Automated lung nodule classification following automated nodule detection on ct: A serial approach,” *Medical Physics*, vol. 30, 2003.
- [36] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [37] F. Zhang, Y. Song, W. Cai, Y. Zhouy, S. Shanz, and D. Feng, “Context curves for classification of lung nodule images,” *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference*, pp. 1–7, 2013.
- [38] A. Vedaldi and S. Soatto, “Quick shift and kernel methods for mode seeking,” in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 705–718.
- [39] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, p. 2507–2517, 2007.
- [40] I. Guyon, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [41] A. Jain and D. Zongker, “Feature selection: Evaluation, application and small sample performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, 1997.
- [42] H. Krewer, B. Geiger, L. Hall, D. Goldgof, Y. Gu, M. Tockman, and R. Gillies, “Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography,” in *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 2013, pp. 3887–3891.
- [43] M. A. Hall, “Correlation-based feature selection for machine learning,” Ph.D. dissertation, The University of Waikato, 1999.

- [44] I. Kononenko, “Estimating attributes: analysis and extensions of relief,” in *Machine Learning: ECML-94*. Springer, 1994, pp. 171–182.
- [45] S. B. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.
- [46] S. Sun, Y. Guo, Y. Guan, H. Ren, L. Fan, and Y. Kang, “Juxta-vascular nodule segmentation based on flow entropy and geodesic distance,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1355–1362, 2014.
- [47] M. A. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, “Support vector machines,” *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.
- [48] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
- [49] S. A. Dudani, “The distance-weighted k-nearest-neighbor rule,” *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 325–327, 1976.
- [50] H. Chen, Y. Xu, Y. Ma, and Binrong, “Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on ct images,” *Academic Radiology*, vol. 17, no. 5, pp. 248–257, 2010.
- [51] M. Sonka and J. M. Fitzpatrick, *Handbook of Medical Imaging, Volume 2 - Medical Image Processing and Analysis*. SPIE, 2009.
- [52] S. Worthy, “High resolution computed tomography of the lungs,” *BMJ*, vol. 310, no. 6980, p. 616, 1995.
- [53] *Insight into Images: Principles and Practice for Segmentation, Registration, and Image Analysis*. AK Peters Ltd, 2004, ch. Introduction - Medical Imaging Technology, pp. 7–10.
- [54] *Dail and Hammar’s Pulmonary Pathology*. Springer New York, 2008, ch. Anatomy and Histology of the Lung, pp. 20–48.
- [55] J. Novo, J. Rouco, A. Mendonça, and A. Campilho, “Reliable lung segmentation methodology by including juxtapleural nodules,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8815, pp. 227–235, 2014.
- [56] P. Badura and E. Pietka, “Soft computing approach to 3d lung nodule segmentation in ct,” *Computers in Biology and Medicine*, vol. 53, pp. 230–243, 2014.

- [57] F. Heckel, H. Meine, J. Moltz, J.-M. Kuhnigk, J. Heverhagen, A. Kießling, B. Buerke, and H. Hahn, “Segmentation-based partial volume correction for volume estimation of solid lesions in ct,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 462–480, 2014.
- [58] S. Diciotti, S. Lombardo, M. Falchini, G. Picozzi, and M. Mascalchi, “Automated segmentation refinement of small lung nodules in ct scans by local shape analysis,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12 PART 1, pp. 3418–3428, 2011.
- [59] C. Jacobs, E. van Rikxoort, T. Twellmann, E. Scholten, P. de Jong, J.-M. Kuhnigk, M. Oudkerk, H. de Koning, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, “Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images,” *Medical Image Analysis*, vol. 18, no. 2, pp. 374–384, 2014.
- [60] S. Saien, A. Hamid Pilevar, and H. Abrishami Moghaddam, “Refinement of lung nodule candidates based on local geometric shape analysis and laplacian of gaussian kernels,” *Computers in Biology and Medicine*, vol. 54, pp. 188–198, 2015.
- [61] H. Han, L. Li, H. Wang, H. Zhang, W. Moore, and Z. Liang, “A novel computer-aided detection system for pulmonary nodule identification in ct images,” in *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 9035, 2014.
- [62] P. Aggarwal, R. Vig, and H. Sardana, “Patient-wise versus nodule-wise classification of annotated pulmonary nodules using pathologically confirmed cases,” *Journal of Computers (Finland)*, vol. 8, no. 9, pp. 2245–2255, 2013.
- [63] P. Aggarwal, R. Vig, and K. Sardana, “Largest versus smallest nodules marked by different radiologists in chest ct scans for lung cancer detection,” in *Lecture Notes in Engineering and Computer Science*, vol. 1, 2013, pp. 462–466.
- [64] F. Han, H. Wang, B. Song, G. Zhang, H. Lu, W. Moore, H. Zhao, and Z. Liang, “A new 3d texture feature based computer-aided diagnosis approach to differentiate pulmonary nodules,” in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8670, 2013.
- [65] X. He, B. Sahiner, B. Gallas, W. Chen, and N. Petrick, “Computerized characterization of lung nodule subtlety using thoracic ct images,” *Physics in Medicine and Biology*, no. 4, pp. 897–910, 2014.
- [66] S. Armato, G. McLennan, L. Bidaut, F. McNitt-Gray, R. Meyer, P. Reeves, B. Zhao, R. Aberle, I. Henschke, A. Hoffman, A. Kazerooni, H. MacMahon, R. van Beek, and D. Y. 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*, vol. 38, pp. 915–931, 2011.
- [67] U. D. of Health and H. Services. (NA) Lidc-idri. National Cancer Institute. [Online]. Available: <http://imaging.cancer.gov/informatics/lidcidri>
- [68] S. Silva, J. Madeira, B. S. Santos, and C. Ferreira, “Inter-observer variability assessment of a left ventricle segmentation tool applied to 4d mdct images of the heart,” in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 3411–3414.
- [69] B. Lassen, C. Jacobs, J. Kuhnigk, B. van Ginneken, and E. van Rikxoort, “Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans,” *Physics in medicine and biology*, vol. 60, no. 3, pp. 1307–1323, 2015.
- [70] D. A. Auger, X. Zhong, F. H. Epstein, E. M. Meintjes, and B. S. Spottiswoode, “Semi-automated left ventricular segmentation based on a guide point model approach for 3d cine dense cardiovascular magnetic resonance,” *Journal of Cardiovascular Magnetic Resonance*, vol. 16, p. 8, 2014.
- [71] S. Balocco, C. Gatta, F. Ciompi, A. Wahle, P. Radeva, S. Carlier, G. Unal, E. Sanidas, J. Mauri, X. Carillo *et al.*, “Standardized evaluation methodology and reference database for evaluating ivus image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 38, no. 2, pp. 70–90, 2014.
- [72] L. J. Stapleford, J. D. Lawson, C. Perkins, S. Edelman, L. Davis, M. W. McDonald, A. Waller, E. Schreiber, and T. Fox, “Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 77, no. 3, pp. 959–966, 2010.
- [73] H.-H. Chang, A. H. Zhuang, D. J. Valentino, and W.-C. Chu, “Performance measure characterization for evaluating neuroimage segmentation algorithms,” *Neuroimage*, vol. 47, no. 1, pp. 122–135, 2009.
- [74] N. B. Smith and A. Webb, *Introduction to medical imaging: physics, engineering and clinical applications*. Cambridge university press, 2010.
- [75] J. Novo, L. Gonçalves, A. M. Mendonça, and A. Campilho, “3d lung nodule candidates detection in multiple scales.” Japan: MVA 2015 – IAPR International Conference on Machine Vision Applications, 2015, pp. 5–8.
- [76] M. Bevk and I. Kononenko, “A statistical approach to texture description of medical images: a preliminary study,” in *Computer-Based Medical Systems, 2002.(CBMS 2002). Proceedings of the 15th IEEE Symposium on*. IEEE, 2002, pp. 239–244.

- [77] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.
- [78] R. W. Conners and C. A. Harlow, “A theoretical comparison of texture algorithms,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 3, pp. 204–222, 1980.
- [79] L.-K. Soh and C. Tsatsoulis, “Texture analysis of sar sea ice imagery using gray level co-occurrence matrices,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 2, pp. 780–795, 1999.
- [80] D. A. Clausi, “An analysis of co-occurrence texture statistics as a function of grey level quantization,” *Canadian Journal of remote sensing*, vol. 28, no. 1, pp. 45–62, 2002.
- [81] F. Albrechtsen *et al.*, “Statistical texture measures computed from gray level coocurrence matrices,” *Image processing laboratory, department of informatics, university of oslo*, pp. 1–14, 2008.
- [82] S. E. Grigorescu, N. Petkov, and P. Kruizinga, “Comparison of texture features based on gabor filters,” *Image Processing, IEEE Transactions on*, vol. 11, no. 10, pp. 1160–1167, 2002.
- [83] J. Yang, L. Liu, T. Jiang, and Y. Fan, “A modified gabor filter design method for fingerprint image enhancement,” *Pattern Recognition Letters*, vol. 24, no. 12, pp. 1805–1817, 2003.
- [84] M. Haghghat, S. Zonouz, and M. Abdel-Mottaleb, “Identification using encrypted biometrics,” in *Computer Analysis of Images and Patterns*. Springer, 2013, pp. 440–448.
- [85] K. I. Laws, “Textured image segmentation.” DTIC Document, Tech. Rep., 1980.
- [86] Y. Liu and Y. F. Zheng, “Fs_sfs: A novel feature selection method for support vector machines,” *Pattern recognition*, vol. 39, no. 7, pp. 1333–1345, 2006.
- [87] P. S. Bradley and O. L. Mangasarian, “Feature selection via concave minimization and support vector machines.” in *ICML*, vol. 98, 1998, pp. 82–90.
- [88] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Piaggio, and V. Vapnik, “Feature selection for svms, iadvances in neural information processing systems”, 13,” 2001.
- [89] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

- [90] K. Mao, “Feature subset selection for support vector machines through discriminative function pruning analysis,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 1, pp. 60–67, 2004.
- [91] M. G. Genton, “Classes of kernels for machine learning: a statistics perspective,” *The Journal of Machine Learning Research*, vol. 2, pp. 299–312, 2002.
- [92] J. M. Bland and D. G. Altman, “Measuring agreement in method comparison studies,” *Statistical methods in medical research*, vol. 8, no. 2, pp. 135–160, 1999.