

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

3D Lung Nodule Classification in Computed Tomography Images

Ana Rita Felgueiras Carvalho



Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Supervisor: Aurélio Joaquim de Castro Campilho

Supervisor: António Manuel Cunha

October 25, 2019

3D Lung Nodule Classification in Computed Tomography Images

Ana Rita Felgueiras Carvalho

Mestrado Integrado em Engenharia Electrotécnica e de Computadores

October 25, 2019

Resumo

O cancro do pulmão é a maior causa de morte por cancro em todo o mundo. O diagnóstico é muitas vezes realizado pela leitura de tomografias computadorizadas, TC's - uma tarefa propensa a erros. Existe bastante subjetividade na avaliação do diagnóstico do nódulo. Esta, entre outras causas, podem levar a um diagnóstico inicial incorrecto o que pode ter um grande impacto na vida do paciente [1]. O uso do sistema Diagnóstico Assistido por Computador, CADx, pode ser uma ajuda para este problema, fornecendo uma segunda opinião aos médicos quanto ao diagnóstico da malignidade dos nódulos pulmonares. No entanto, os médicos acabam por evitar o uso destes sistemas [2], pois não conseguem perceber o “como e porquê” dos resultados que apresentam. De forma a aumentar a confiança dos médicos nos sistemas CADx, Shen *et al.* [3] propuseram uma rede neuronal convolucional semântica hierárquica profunda, HSCNN, que para além de ter como resultado um diagnóstico de malignidade, prevê também características semânticas relacionadas com o mesmo, demonstrando assim evidências visuais do diagnóstico. Como entrada para a HSCNN, Shen *et al.* usaram a LIDC-IDRI, Lung Image Database Consortium and Image Database Resource Initiative, uma base de dados pública que contém 2,668 nódulos anotados por radiologistas relativamente a algumas das características semânticas que apresentam [4].

Esta base de dados contém bastante variação entre as avaliações dadas pelos médicos, para o mesmo nódulo, na mesma característica, principalmente nas características lobulação e espiculação, as duas da base de dados com mais relevância para a malignidade. Por essa razão, Shen *et al.* não previram na sua rede essas duas características. Motivada por estas razões, uma questão sugerida por Shen *et al.* foi levantada: com a diminuição da variabilidade nas avaliações das características semânticas, será que a previsão da malignidade na HSCNN irá melhorar?

Para responder a esta pergunta foi apresentada como uma possível solução a obtenção de de um dataset LIDC-IDRI anotado de forma mais coerente. Para tal, foi realizada inicialmente uma análise à base de dados e uma revisão da literatura, que permitiu a criação de critérios de avaliação das características semânticas mais exatos do que os existentes. Depois destes serem criados e validados por radiologistas do Centro Hospitalar do São João, foi reanotada a base de dados LIDC-IDRI, seguindo uma metodologia proposta com base nos critérios concebidos. Finalmente, a base de dados reclassificada foi usada como entrada para a rede HSCNN, de forma a avaliar o impacto que esta tinha na previsão da malignidade. Foram utilizados 1,386 nódulos e todas as anotações foram binarizadas. O treino da rede foi realizado usando 4-fold cross-validation.

Os resultados desta solução proposta apenas foram melhores, comparativamente ao trabalho de Shen *et al.*, nos valores de sensibilidade, existindo uma melhor previsão de quando os nódulos são realmente malignos. A previsão da malignidade teve uma exatidão de 0.78, uma área sob a curva da característica de operação do receptor, AUC da ROC, de 0.74, uma sensibilidade de 0.83, e uma especificidade de 0.89. No trabalho de Shen *et al.* foram obtidas uma exatidão de 0.84, uma AUC da ROC de 0.86, uma sensibilidade de 0.71 e uma especificidade de 0.89. Uma análise às previsões incorretas revelou que a maior parte destas eram compostas por nódulos pequenos, com bastante matéria à volta, com calcificação sólida ou juxta-pleurais. Os nódulos corretamente

previstos eram na sua maior parte solitários, ou com tecido de pleura adjacente. No entanto, esta comparação com o trabalho de Shen *et al.* não é linear, pois o número de dados usado foi diferente (1,386 nódulos foram usados neste trabalho ao invés de 4,252 no trabalho de Shen *et al.*) e a conversão das avaliações para binário foi também realizada de outra forma, o que pode ter tido influência nos resultados finais.

Com base nos resultados obtidos ao utilizar a metodologia proposta para a diminuição da variabilidade nas avaliações das características semânticas da LIDC-IDRI, foi verificado que não existe uma melhoria na previsão do diagnóstico da malignidade na HSCNN, comparativamente aos resultados de Shen *et al.*.

Abstract

Lung cancer is the leading cause of cancer death worldwide. Usually, the diagnosis is made by physicians reading computed tomographies, CTs - a task prone to errors. There is a lot of subjectivity in the nodules' diagnose assessment. That, among other causes, can lead to an incorrect initial diagnosis, delaying the correct treatment, which can have a high impact on the patient's life [1]. The use of Computer-Aided Diagnosis, CADx, can be a great help for this problem by assisting doctors assessing lung nodule malignancy, giving a second opinion. However, physicians end up avoiding the use of CADx [2], because can not understand the "how and why" of its diagnostic results. To increase the physicians' confidence in the CADx system, Shen *et al.* [3] proposed a deep hierarchical semantic convolutional neural network, HSCNN, that besides having as result the malignancy' diagnosis, also predicts the features related to it, giving this way visual evidence of the diagnosis. As input for HSCNN Shen *et al.* used LIDC-IDRI, Lung Image Database Consortium and Image Database Resource Initiative, a widely used public dataset, that has 2,668 abnormalities annotated by radiologists regarding the semantic features that they present [4].

This dataset has a high variance between the assessments given by physicians, for the same nodule, in the same semantic feature, especially in lobulation and spiculation, the two LIDC-IDRI features most relevant to malignancy. For this reason, Shen *et al.* did not predict in their work these two characteristics. With that motivation, a question suggested by Shen *et al.* was raised: will the decrease of variability in LIDC-IDRI semantic features evaluations improve the HSCNN malignancy prediction?

To answer this question was presented as a possible solution to obtain a more coherently annotated LIDC-IDRI dataset. For this purpose, a dataset analysis and literature review were initially realised, which allowed the creation of more accurate semantic characteristics evaluation criteria than the existing ones. After these were created and validated by radiologists at the São João Hospital Center, the LIDC-IDRI database was revised, following a proposed methodology based on the devised criteria. Finally, the reclassified dataset was used as input to the HSCNN to evaluate its impact on malignancy prediction. 1,386 nodules were used and all annotations were binarised. The network training was performed using 4-fold cross-validation.

The results of this proposed solution improved only, compared to the work of Shen *et al.*, in the sensitivity, with a better prediction of when the nodules are truly malignant. The malignancy prediction obtained an accuracy of 0.78, an area under the receiver operating characteristic curve, AUC of ROC, of 0.74, a sensitivity of 0.83, and a specificity of 0.89. In the Shen *et al.* work were obtained as accuracy 0.84, as AUC of ROC 0.86, as sensitivity 0.71 and as specificity 0.89. The analysis of the network incorrect predictions revealed that most of them were small nodules, with a lot of matter around, with solid calcification, or juxta-pleural. The correctly predicted nodules were solitary, or with pleural-tail. However, the comparison with Shen *et al.* work is not linear since the quantity of data used was different (1,386 nodules instead of the 4,252 used in Shen *et al.* work) and the conversion to binary was also performed differently, which may have influenced

the final results.

Based on the results obtained, using the proposed methodology to the decrease of LIDC-IDRI semantic features evaluations variability, it is verified that there was no increase in malignancy prediction in HSCNN, comparing to Shen *et al.* results.

Acknowledgments

Ao co-orientador António Cunha, pela constante preocupação e orientação na duração desta dissertação e por me ter proporcionado a oportunidade de trabalhar nesta área. Ao grupo C-BER, pela sua companhia e ajuda.

Ao Diogo Moreira, e Daniela Afonso por me terem acompanhado fielmente nestes últimos 6 anos, por termos crescido tanto juntos. À Mariana Caixeiro, por toda a amizade e conversas em que as palavras não são necessárias.

À Manuela Rodrigues e Sara Alves, por me fazerem acreditar que existem amizades que podem durar para sempre. Por serem uma parte da minha alma, por me darem um pouco da sua força, e acreditarem em mim sempre, mesmo quando eu não acreditei. Nunca vão existir palavras suficientes. À Daniela Afonso, mais uma vez, por me mostrar o que é a amizade na sua mais pura forma de ser. Por me ter acompanhado nas fases mais, e menos, belas da minha vida, e por sempre nos termos compreendido na essência do que somos. Obrigada às três por, conhecendo todas as minhas versões, me ajudarem sempre a ser a melhor.

Ao Diogo Pernes, por toda a paciência, apoio e companhia, imprescindíveis nesta etapa da minha vida. Por tudo o que me ensinou e por ter tornado isto possível. Obrigada por presentear os meus dias com o teu brilho.

À minha família, a quem devo grande parte do que sou. Ao meu Pai, por me dar o privilégio de sentir tudo. Ao meu Irmão, Miguel, por trazer tanto amor no seu coração e transbordá-lo para o meu. À minha Mãe, por representar a beleza mais pura do meu mundo. À minha Avó Lina, por me ensinar que o amor é o maior ensinamento da vida.

Ana Rita Felgueiras Carvalho

*“Procura em ti
O que não encontras
Em lado nenhum”*

Mariano Alejandro Ribeiro

*“É naquilo que faço
que encontro
o que procuro”*

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Objective	3
1.3	Contributions	3
1.4	Document Structure	3
2	Problem Contextualization	5
2.1	Lung Cancer	5
2.2	Lung Cancer Diagnose	7
2.3	Computer-Aided Diagnosis System	10
2.4	Summary	10
3	Literature Review	11
3.1	Deep Learning Fundamentals	11
3.1.1	Neural Networks	12
3.1.2	Convolutional Neural Networks	16
3.1.3	Learning	18
3.2	LIDC-IDRI Dataset	21
3.3	Lung Nodule Classification Methods	23
3.4	Discussion	28
3.5	Summary	30
4	Methodology	31
4.1	LIDC-IDRI Analysis	32
4.1.1	Dataset Review Tool	32
4.1.2	LIDC-IDRI Semantic Features Bibliographic Review	34
4.2	Dataset Reassessment	42
4.3	Impact Evaluation	44
4.4	Summary	47
5	Results and Discussion	49
5.1	Evaluation Norms	49
5.2	Dataset Reassessment	54
5.3	Impact Evaluation	57
5.3.1	Dataset Preparation	57
5.3.2	Training	60
5.3.3	Results	61
5.3.4	Discussion	65

5.4 Summary	70
6 Conclusions and Future Work	71
A Worksheet File	73
B Norms	77
References	85

List of Figures

1.1	Age Standardized incidence Rates vs. Mortality Rates, in 2018 [5].	1
2.1	Lungs anatomy and structure [6].	6
2.2	Original CT slice with lung tissues localization and intensity values are given in HU [7].	7
2.3	Characterization by nodule’s location from [8].	8
2.4	Characterization by nodule’s intensity from [8].	8
2.5	Lung nodules from [4], representing solid, non-solid, and part-solid densities, in axial view.	9
3.1	Artificial neuron.	12
3.2	Identity, logistic sigmoid, hyperbolic tangent and ReLu functions graphics. . . .	14
3.3	Multilayer perceptron, with one hidden layer.	14
3.4	Comparison between a typical MLP and a CNN, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (red, green, blue) channels.) [9].	16
3.5	Example of a convolution operation, with two filters of size 3x3, applied with a stride (the length that the kernel is moving) of 2 [9].	17
3.6	Max pooling layer operation with kernel size 2x2 and stride length 2.	18
3.7	CNN example, with an output of 5 classes [9].	18
3.8	Multi-crop architecture [10]. The numbers along each side of the cuboid indicate the dimensions of the feature maps.	23
3.9	Multi-crop pooling [10].	24
3.10	CT examples of the three types of ternary classification (from left to right, benign, primary malignant and metastatic malignant) with, in each one, the different 7 different views areas, from left to right 20x20, 30x30, 40x40, 50x50, 60x60, 70x70 and 80x80, taken from [11].	25
3.11	Multi-view’s strategies: one-view-one-network(top) and multi-view-one-network (bottom).	26
3.12	The architecture of 3.12a was taken from [11] and architectures from figures 3.12b and 3.12c were taken from [12].	26
3.13	HSCNN architecture taken from [3].	27
4.1	The methodology used for LIDC-IDRI review and impact evaluation in the malignancy prediction.	31
4.2	Nodule images from [4], without and with binary mask in axial, coronal, and sagittal planes.	33
4.3	Popcorn, laminated, solid, non-central, central and absent calcification example nodules from LIDC-IDRI dataset [4].	36

4.4 Malign and benign types of calcification by [13]. 37

4.5 Internal structure’s examples. 4.5a is from LIDC-IDRI dataset [4], and 4.5b and 4.5c are from [14]. 38

4.6 Non lobulated and lobulated nodules examples from LIDC-IDRI dataset [4]. 38

4.7 Poorly defined and sharp margins nodules from LIDC-IDRI dataset [4]. 39

4.8 Linear and round nodules examples from LIDC-IDRI dataset [4]. 40

4.9 No spiculated and spiculated nodules examples from LIDC-IDRI dataset [4]. 41

4.10 Extremely subtle and obvious nodules examples from LIDC-IDRI dataset [4]. 42

4.11 Method used to do the dataset reassessment. 43

5.1 Nodule with a non-central calcification that has an LIDC-IDRI’s doctors averages evaluation of central calcification from [4]. 50

5.2 Implemented 4 fold cross-validation. 61

5.3 Confusion matrices for each semantic feature, using fold 2 as test set. 63

5.4 ROC curve for each semantic feature, using fold 2 as test set. 64

5.5 Malignancy ROC curve for fold 2, with an AUC = 0.74. 65

5.6 Nodules images with their predictions and truth labels. 66

A.1 Nodule information, and evaluations’ average given by the doctors, in the worksheet file. 73

A.2 Classifications given by a doctor in evaluation 1, in the worksheet file. 74

A.3 Nodule’s images without and with the average binary mask of the doctors, in the worksheet file. 75

B.1 Calcification and its classes definition. 77

B.2 Very small nodules for each class in calcification. 77

B.3 Small and big nodules for each class in calcification. 78

B.4 Lobulation and its classes definition. 78

B.5 Very small for each class in lobulation. 78

B.6 Small and big nodules for each class in lobulation. 79

B.7 Margin and its classes definition. 79

B.8 Very small for each class in margin. 79

B.9 Small and big nodules for each class in margin. 80

B.10 Spiculation and its classes definition. 80

B.11 Very small for each class in spiculation. 80

B.12 Small and big nodules for each class in spiculation. 81

B.13 Subtlety and its classes definition. 81

B.14 Very small for each class in subtlety. 81

B.15 Small and big nodules for each class in subtlety. 82

B.16 Texture and its classes definition. 82

B.17 Very small for each class in texture. 82

B.18 Small and big nodules for each class in texture. 83

List of Tables

3.1	Explanation of various cases with results of Disease vs Test.	20
3.2	LIDC nodule characteristics, definitions and ratings proposed in [15].	22
3.3	Binary labels proposed by Shen <i>et al.</i> [3] from LIDC-IDRI rating scales for nodule characteristics.	28
3.4	Results comparison between 3D CNN and HSCNN.	29
3.5	Results for semantic features predictions	29
3.6	Final best results for the different relevant CNN's architectures presented	30
4.1	Characteristics and their classes in LIDC-IDRI dataset proposed by Opulencia <i>et al.</i> [15]	35
5.1	Ratings division obtained from the analysis.	53
5.2	Comparison between the nodules number in the original and reclassified datasets	56
5.3	Characteristics conversion from ternary to binary classification.	58
5.4	Ratings division obtained from the binarisation.	58
5.5	Total number of nodules for each characteristic	59
5.6	Total number of nodules for each characteristic and malignancy, after selecting only the abnormalities that had 3 or more evaluations in LIDC-IDRI	60
5.7	Average test results for each semantic feature.	62
5.8	Best test results for each semantic feature, using fold 2.	62
5.9	Test results for each fold, for lung nodule malignancy prediction.	62
5.10	Confusion matrix for malignancy, using fold 2 as test set	63

Abbreviations and Symbols

2D	Two Dimensional
3D	Three Dimensional
2.5D	Two-and-a-half Dimensional
ANN	Artificial Neural Network
AUC	Area Under the Curve
CADx	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
GGO	Ground Glass Opacity
HSCNN	Hierarchical Semantic Convolution Neural Network
IRB	Institutional Review Board
k-NN	K-Nearest Neighbors
LDCT	Low Dose Computed Tomography
LIDC-IDRI	Lung Image Database Consortium image collection
LNDetector	Lung Detector
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
MVCNN	Multi-view Convolutional Neural Network
MC-CNN	Multi-croop Convolutional Neural Network
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SVM	Support-Vector Machine

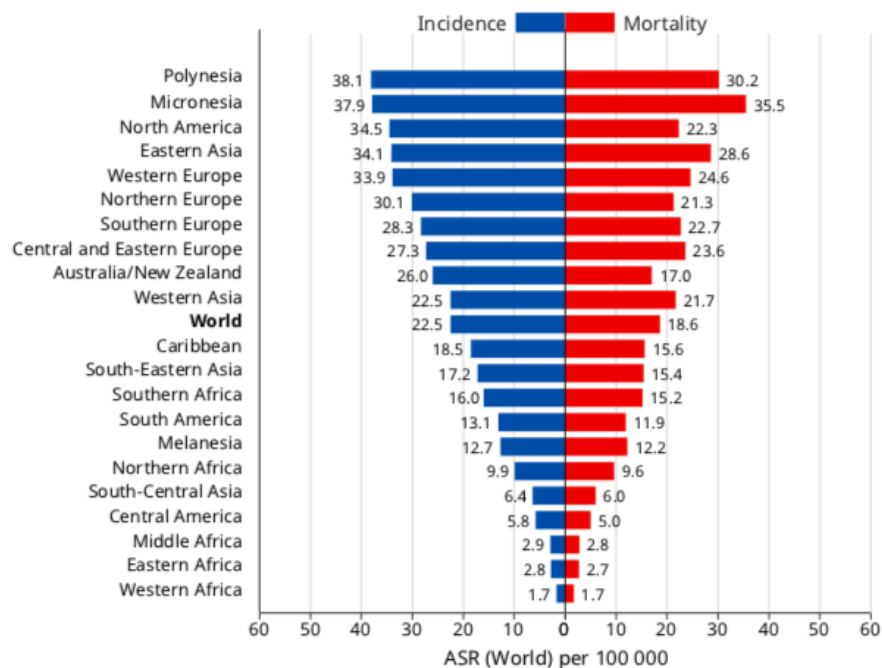
Chapter 1

Introduction

Lung cancer is the pathology that has more mortality globally, accounting for more deaths than cancers of the breast, prostate, colon, and pancreas combined [16]. Worldwide, it was estimated to be the cause of death of 1,761,007 people, from both sexes, on a total of 2,093,876 incidences (giving a death rate of 84%, when is known to have cancer), in 2018 [5].

In Portugal, the International Agency for Research on Cancer [17] estimated that in 2018 one-quarter of the population is in risk of developing cancer until 75 years, expecting 5,284 incidences, having a standardized rate per 100,000 of 22.6%, from both sexes, and all ages. To a better perception of the problem, the incidence and mortality age-standardized rate per 100,000 population, for other world areas could be seen in Figure 1.1.

Figure 1.1: Age Standardized incidence Rates vs. Mortality Rates, in 2018 [5].



The survival rate for lung cancer (five-year) in the United States of America is approximately

18.1% [18], much lower than other cancer types as patients only begin to show clinical symptoms once lung cancer has advanced enough. However, early-stage lung cancer has a survival rate (five-year) of 60-75% [18]. With this, it is concluded that early lung cancer diagnosis and treatment are associated with a higher survival rate than when there is a late diagnosis [1, 19, 20].

In 1972, Godfrey Hounsfield and Allan Cormack invented computed tomography (CT), a method that permits to produce cross-sectional images (slices) that, together, make possible to analyze the internal organs and tissue of the human body [21]. Currently, CT is a widely used method for the diagnosis of lung cancer.

Patients with suspicions of this type of cancer must be examined each year by making chest CTs so that radiologists can check for the presence of new nodules or following-up detected nodules in previous screening sessions. This task is error-prone due to experts fatigue because CT has hundreds of images, and lung nodules are often difficult to distinguish from lung anatomical structures. The distinction between malign and benign lung nodules is also complicated, and there is often a high variance between experts opinions, concerning the classification of the same nodule [22, 23]. When there is cancer, and the initial diagnosis is incorrectly made, it can lead to a late diagnosis which has a high impact on the patient's life, delaying the treatment can lead to the patient death.

1.1 Motivation

The use of CADx, Computer-Aided Diagnosis, systems, invented in 1961, can be a great help for this problem, assisting specialists in lung nodules detection and classification. CADx aims at providing a second opinion for the detection of the nodules and establishes their probability of malignancy, helping to reduce cases of uncertainty and, consequently, reduce the number of unnecessary biopsies, thoracotomies, and surgeries [24]. Although the efficiency of CADx systems has already been proven [25], many experts refer to such systems as a "black box" they can not rely on as a "fair third-party reference" because they can not understand the "how and why" of the resulting forecasts, ultimately avoiding the use of these systems [2].

A possible solution to increase the expert's confidence in the predictions, is the presentation of evidence with which they can understand the CADx results in benign or malign classification: a prediction of lung nodules semantic features correlated with malignancy.

Most traditional methods for automated nodules classification includes feature extraction, classifier training, and testing. They do not work in an end-to-end manner: the feature selection is extracted by hand or with predefined filters. Alternatively, deep learning methods learn from the data through a general learning process [26] without needing to extract features as the traditional way.

Recently, convolutional neural networks, an end-to-end deep learning framework, have shown to outperform the state of the art in medical images analysis [27]. As the lung nodule semantic features can be characterized numerically [28], the suggestion is to use CNN techniques that

take advantage of the visible features of nodules, used by specialists in the diagnostic decision as calcification, spiculation, lobulation, and others suggested by Shen *et al.* [3].

Shen *et al.* proposed an architecture, HSCNN [3], that predicts semantic features, and using the information of each feature, predicts malignancy. They use as input the LIDC-IDRI, a known public dataset, very used for lung nodule CADx systems, that presents characteristics evaluations by doctors, in each nodule [4]. They did not use all the information that LIDC-IDRI has, not using the most malignancy related characteristics evaluated in the dataset: lobulation, and spiculation [29]. That occurred because of the high disparity that those characteristics present in the ratings given by doctors. Since there is not a norm well defined to evaluate characteristics, the evaluation is very subjective, especially in lobulation and spiculation, which induces much variety in the physicians' opinions.

One of the questions suggested by Shen *et al.*, that remains to be solved is: will the decrease of variability in LIDC-IDRI semantic features evaluations improve the HSCNN malignancy prediction?

1.2 Objective

The principal objective of this dissertation is to decrease LIDC-IDRI semantic annotation variability and evaluate its impact on the malignancy prediction. The solution proposed to decrease the variability is to have a more coherent annotated dataset. To do so, first, was made an LIDC-IDRI analysis. With the information obtained by the analysis, were created evaluation criteria, validated by physicians, which served to reassess the dataset in a more standardized form. Finally, to evaluate the impact on the malignancy prediction, HSCNN architecture was trained with the adjusted dataset, and the results reported and discussed.

1.3 Contributions

This dissertation presents the following contributions:

- Evaluate norms for reassessing the LIDC-IDRI semantic annotation dataset;
- LIDC-IDRI dataset reassessed;
- Analysis and discussion of the prediction of HSCNN nodule semantic features and malignancy with the reassessed dataset.

1.4 Document Structure

The MSc dissertation is organized in six chapters, as follows. The present chapter introduces the importance of a lung cancer early diagnosis and a CAD system. Is identified the motivation behind the decrease of the variability on LIDC-IDRI dataset evaluations. Moreover, the dissertation's objectives and expected contributions are also presented. Chapter 2 gives a problem contextualization, presenting the lung anatomy and cancer, how to detect it, and CADx systems. In chapter 3, a

literature review is made about the classification module of CADx system. A contextualization of deep learning fundamentals is made and the dataset used to the system is presented. Lung nodules classification methods using that dataset, are shown, as its results and a discussion about them. Chapter 4 presents the methods used for the LIDC-IDRI analysis, for the dataset reassessment and for the impact evaluation. Chapter 5, shows the results of each one of the methodology modules: the evaluation norms, the dataset reassessed, and the impact evaluation. Is held a discussion about them. And lastly, in chapter 6 some insight final is done writing recommendations about the work done and future working that can be explored.

Chapter 2

Problem Contextualization

In this chapter will be presented the background of this dissertation subject. In section 2.1 is presented the lung anatomy and a contextualization on lung cancer. In section 2.2 is given a brief clarification about what a CT is, and how can a lung cancer diagnosis be performed with it. As there are many false-negative diagnoses in lung cancer, and CAD systems can be an aid to this problem, section 2.3 gives a concise explanation of it.

2.1 Lung Cancer

The lungs function in the respiratory system is to draw oxygen from the atmosphere, transfer it into the bloodstream, and release carbon dioxide from the bloodstream into the atmosphere, resulting in a gas exchange process.

The right lung has three lobes while the left lung is split into two lobes and a small structure named lingula, which is equivalent to the central lobule of the right side of the lungs.

The major airways that enter the lungs are the bronchi that arise from the trachea, which is outside of the lungs. The bronchi branch into airways progressively smaller, called bronchioles, ending in small sacs known as alveoli. It is in these bags that gas exchange occurs. The lungs and thoracic wall are covered by a thin layer of tissue called pleura [30]. Lungs anatomy can be seen in Figure 2.1.

Normally the body maintains a growth system of verification and balance of the cells so that they divide to produce new cells when these are necessary. When there is an interruption in this cell growth balance, it results in an uncontrollable cell division and proliferation of cells, which originates nodules. If a nodule appears in the lungs, and is greater than 3 centimeters in diameter, it is called lung mass and is more likely to represent cancer.

A tumor can be benign or malign. Generally, a benign can be removed and does not spread to other parts of the body. The equivalent does not happen with a malignant that normally has an aggressively fast growth in the area where they originate and can also enter the bloodstream and lymphatic system, spreading to other areas of the body. Due to its fast growth, as soon as it is formed, it is a tumor with a high degree of life-threatening and the most difficult to treat.

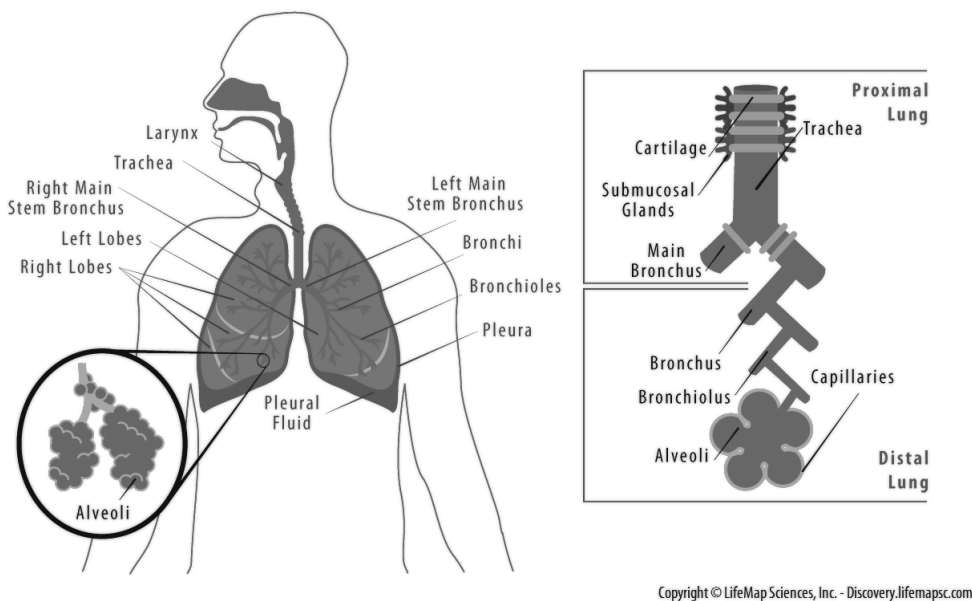


Figure 2.1: Lungs anatomy and structure [6].

There are several possible causes or risk factors that increase an individual's chance of developing lung cancer. These include smoking (correlated with most cases), exposure to tobacco smoke, inhalation of chemical agents, and cancer history in the patient or family [31].

As for the symptoms they vary depending on where and how much the cancer is spread. These are not always present or are easy to identify and may not cause pain or other types of symptoms in some cases, which can cause a delay in the diagnosis. Faced with a set of symptoms and after a deeper understanding of the patient's state of health, an examination is requested by a doctor for a diagnosis. Of all the tests, the X-ray is the first one often to be asked for diagnosis when there are already symptoms that alert to the disease. X-ray can only reveal suspicious areas in the lungs, not being able to determine if they are carcinogenic or not. Another exam frequently asked is the CT (Computed Tomography) scan, for detecting both acute and chronic changes in the lung parenchyma, that is, the internals of the lungs. It is particularly relevant because normally, two-dimensional X-rays do not reveal such defects.

When there is a suspicion, the only way to confirm the disease is to obtain some cells in the zone where exists the uncertainty, for microscopic post-observation. This acquisition is called a biopsy, which covers several techniques such as bronchofibroscopy, transthoracic aspiration biopsy, pleural fluid examination, aspiration biopsy of enlarged lymph nodes or when none of these can produce a diagnosis, surgical biopsy.

Treatment for this type of cancer involves surgery for removal of cancer, chemotherapy, radiation therapy or a combination of these treatments. The overall prognosis of lung cancer, referring to the likelihood of cure or prolongation of life, is very low compared to other types of cancers. Survival rate (five-year) in the United States of America is approximately 18.1% , against 64.5% of colon cancer, 89.6% of breast cancer and against 98.2% of prostate cancer, for example [18, 32].

2.2 Lung Cancer Diagnose

Lung cancer manifests itself through the appearance of nodules, which can be seen in CT's. A Computed Tomography, invented in 1972 by Godfrey Hounsfield and Allan Cormack, is widely used to detect lung nodules. It is a non-invasive computerized X-ray imaging method in which an among of X-rays, produced by an X-ray machine with the shape of a ring, is aimed at a patient and rotated around the body. During the rotation, are taken several snapshots by the narrow beam of an X-ray source, and after each full rotation, the machine's computer uses techniques to combine all the images and construct a 2D cross-sectional images. These cross-sectional images, or slices, only have a few millimeters each (usually ranges from 1-10 millimeters) and contain more detailed information than conventional X-rays demonstrating, for example, smaller nodules. A CT slice could be seen in Figure 2.2. After collecting some slices, when a whole slice is completed, with the help from a machine's computer, the slices are digitally combined to generate a three-dimensional image. A lung CT scan usually consists of 400 slices of 512x512 pixel images of the whole lungs.

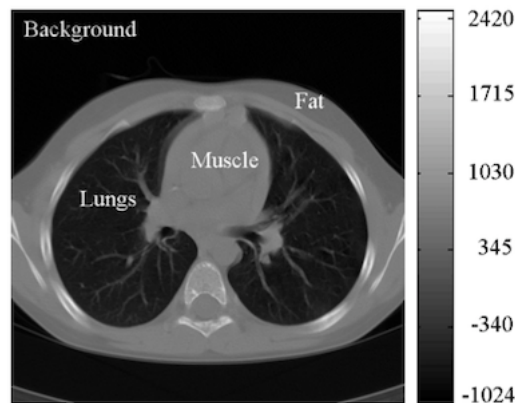


Figure 2.2: Original CT slice with lung tissues localization and intensity values are given in HU [7].

Motivated by the fact that, compared to the chest X-rays, there is a 20% lung cancer mortality reduction by screening CT in the United States of America (demonstrated by National Lung Screening Trial) [33], CT scans are now being even more recommended in current smokers, in people who, having given up smoking, have smoked continuously in the last 30 years (aged 55-80) or smokers that had quit in the past 15 years. Other advantages for using CT's are the ability to rotate the three-dimensional image in space or to view slices in succession, making it easier to identify the basic structures and the exact locations of a tumor or some abnormality.

When reading CT's, physicians frequently do a nodule characterization according to its location or the intensity level (density) present. According to location and connection with surrounding pulmonary structures, lung nodules can be classified into four main categories [34] represented in Figure 2.3 such as:

- Well-circumscribed nodule - with attachment to their neighboring vessels and other anatomical structures;

- Juxta-vascular nodule - with a strong attachment to their nearby vessels;
- Nodule with pleural tail - having a tail which belongs to the nodule itself, show minute attachments to nearby pleural wall;
- Juxta-pleural nodule - with connection to the nearby pleural surface.

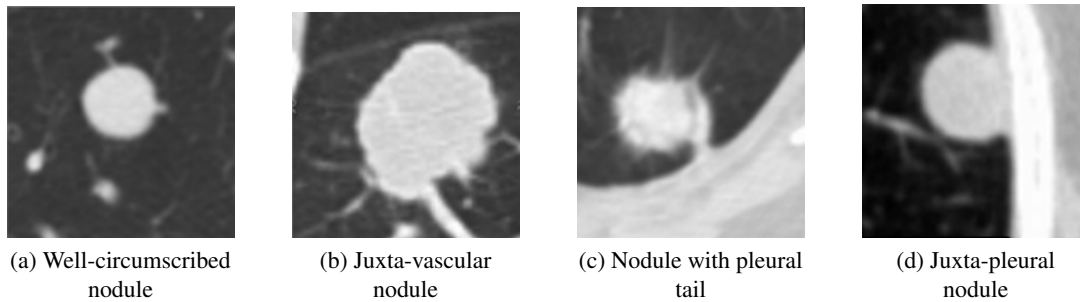


Figure 2.3: Characterization by nodule's location from [8].

As for the nodules intensity (the level of grayscale intensity), these can vary between two classes, represented in Figure 2.4: solid and ground-glass opacity nodules (GGO). Whereas solid nodules have uniform intensity distribution, GGO nodules have a hazy increase in lung attenuation, on CT studies that do not obscure the underlying pulmonary parenchymal architecture.

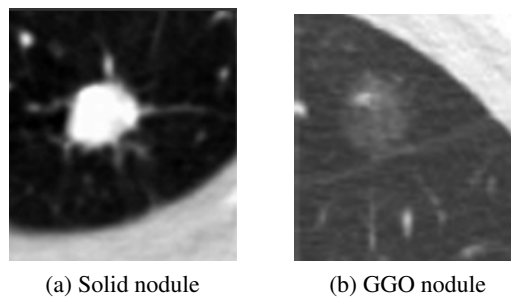


Figure 2.4: Characterization by nodule's intensity from [8].

A GGO nodule can be:

- Part solid – according to Ferretti *et al.* [35] part-solid nodules, also called mixed ground-glass nodules, refer to a ground-glass nodule with a nonsolid component containing tissue density components located centrally, peripherally or forming several islets.
- Non-solid – according to Ferretti *et al.* [35] non-solid nodules, known also for pure ground-glass nodules, are nodules with less than 3 cm in diameter of lower density, than the vascular pattern within them (which is not modified as there are not completely obscured areas).

These three types, solid, non-solid, and part-solid, could be seen in Figure 2.5.

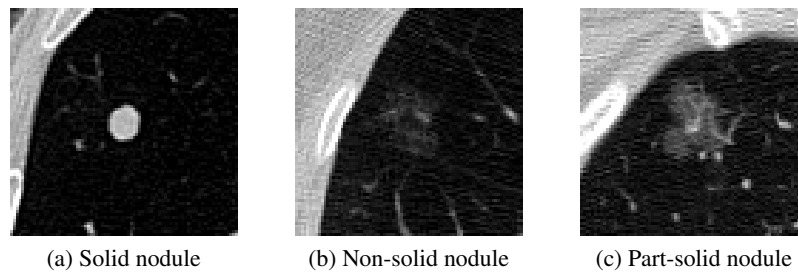


Figure 2.5: Lung nodules from [4], representing solid, non-solid, and part-solid densities, in axial view.

Besides the lung nodules characterization according to intensity, there are some additional characteristics seen in CT's, that represent an essential role in the correct diagnosis [36], being correlated with the malignancy or benignity. Example of those characteristics are size, growth rate, the existence of calcification, the shape, and spiculated, lobulated, well-defined, or ill-defined margins. The size, is a characteristic that is very relevant to diagnosis. Usually, in a solitary nodule, a lesion that has more than 3 cm has more probability to be malignant, being called a mass while a smaller lesion can be either benign or malign, being more likely to be benign if has less than 2 cm.

Although these characteristics are relevant to the diagnosis, the evaluation of many of these is very subjective and may lead to errors in diagnosis. Allied to this, the detection of lung tumors, especially the smaller ones, on a CT scan is still a difficult assignment for physicians. Less experienced specialists have highly variable detection rates as interpretation heavily relies on past experience [22]. Also, as it could be seen in Figure 2.3, normal structures of the lungs are hardly distinguishable from tumors, because of lymph nodes, nerves, muscles and blood vessels look very much like the tumor growth in CT images. Even having a huge dependency on the experience of each specialist, the most experienced eye has trouble finding small tumors in a CT scan.

All of this can lead to an initial diagnosis incorrectly made. If the radiologist evaluation made is a false negative (when there is cancer but the radiologist says there is not) it causes a high impact on the patient's life, as delaying the right treatment can lead to patient death. However, CT scans results also in an increase in false positive screens (when there is no cancer but the radiologist says there is) being the rate upwards of 20%. The juxta-vascular, nodule with pleural tail and juxta-pleural nodules generate also mostly false positive results, due to the structures where they are. This is correlated to unnecessary costs (medical, economic and psychological) [3].

CAD (Computer-Aided Detection and Diagnosis) systems have been very explored in the literature [10, 11, 12], assisting physicians in reducing these diagnose errors, reducing the false positives [25].

2.3 Computer-Aided Diagnosis System

A CAD system is a class of computer systems that aim to assist in the detection and/or diagnosis of diseases through a “second opinion”. Invented by American Sutherland in 1961, the goal of CAD systems is to improve the accuracy of radiologists. These systems are classified into two groups: Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) systems. CADe are systems geared toward the location of lesions in medical images. Moreover, CADx systems, intended to be used in this dissertation, perform the characterization of the lesions, for example, benign and malignant tumors [37].

In the literature, a lung nodule CADx system typically has three modules: automatic detection, segmentation/characterization, and classification. The input for that system will be a thorax CT scan that provides the following outputs:

1. a lung nodule detection, automatically detecting the nodule locations;
2. lung nodule segmentation, showing the detected nodule boundary and characterization with feature measurement - computed features such as texture, morphology and gray-level statistics;
3. lung nodule classification map, highlighting the degree of nodule malignancy.

Although these are the typical modules, sometimes authors do not use the segmentation module and do classification directly from detection.

The part of CADx that will be carried out in this dissertation will be the classification module, predicting nodules’ malignancy and semantic features. The classification is a type of supervised learning, in machine learning.

2.4 Summary

Lung cancer is the most lethal type of cancer. CT scanning is used for lung nodule diagnosis, where physicians can observe its location and its semantic features. However, the CT revision is error-prone, leading sometimes to an initial diagnosis incorrectly made. The overall diagnosis performances can be affected by the subjectivity present when assessing the diagnosis, by the difficulty of detecting the nodules from the structures, by the experience of each physician, or by the fatigue caused by reading hundreds of CT’s. CAD systems can help to improve the nodule diagnosis rate, reducing the false-positive cases. CADx systems give a second opinion about the diagnosis to physicians. CADx’ classification module will be the one employed in this dissertation.

Chapter 3

Literature Review

Broadly, a lung nodule CADx system is typically composed of 3 modules: lung nodule's detection, segmentation, and classification. In the present chapter is given a contextualization, in the module employed in this dissertation, the classification. In this third module, the nodule candidates are categorized by classifiers based on the extracted features from the second module. Some examples of traditional methods that could be used as classifiers are SVM, Support-Vector Machine and k-NN, K-Nearest Neighbors. A proposed work by Gonçalves *et al.* in 2016, using both of these classifiers could be seen in [38]. However, the majority portion of these second module extracted features, such as volume and shape, is sensitive to the lung nodule delineation, becoming inaccurate features that are subsequently used as inputs into the classifier [10]. Another disadvantage of computing features is the difficulty associated with the definition of the optimal subset of features to best capture quantitative lung nodules characteristics [10, 39]. To overcome these issues, deep learning methods have proven to outperform results in lung nodule classification, without needing to extract features as the traditional way.

In section 3.1 are explained the fundamentals of deep learning: neural networks, and the associated learning process in section 3.1.1, convolutional neural networks in 3.1.2, and some training techniques used in it, and how to evaluate its results in section 3.1.3. In section 3.2 is presented the dataset employed in this dissertation, as some existent information about the semantic features evaluated in it. In section 3.3 a research on the state-of-the-art of lung nodule classification methods is made, its results are shown and discussed.

3.1 Deep Learning Fundamentals

Most deep learning methods use Neural Networks, known also for Artificial Neural Networks. First proposed by McCulloch and Pitts [40] in 1943, neural network is a feature learning method that has the advantage of not needing to extract the features traditionally because it learns from the data through a general learning process. The original data is transformed into a higher level,

more abstract by some simple, non-linear model transformations and then the extracted features are trained and tested in the classifier [41].

3.1.1 Neural Networks

Architecturally, a neural network is modeled using layers of artificial neurons able to receive input. An artificial neuron can be seen in Figure 3.1.

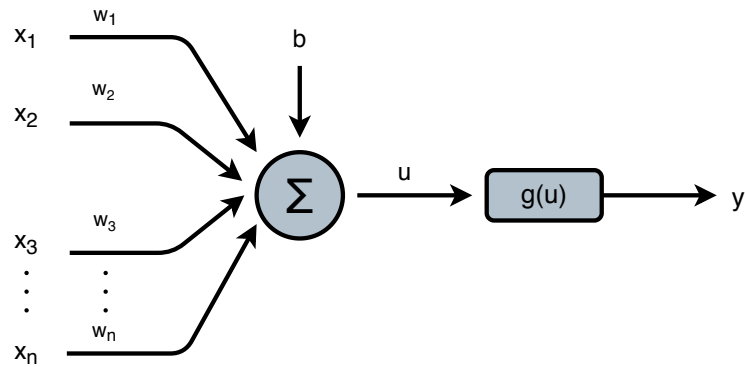


Figure 3.1: Artificial neuron.

In mathematical terms, an output from an artificial neuron can be described by:

$$y = g(u) \quad (3.1)$$

$$u = \sum_{i=1}^n x_i * w_i + b \quad (3.2)$$

$$y = g\left(\sum_{i=1}^n x_i * w_i + b\right) \quad (3.3)$$

where x_1, x_2, \dots, x_n represent the input signals from the application, w_1, w_2, \dots, w_n are the synaptic weights, b is the term bias, g is the activation function chosen, u the activation value of a neuron before applying g , and y the output obtained with $g(u)$.

Activation Function Layer For every neural network layer it is common to apply activation functions, $g(u)$, to find out which node should be fired. They convert large outputs from the units into a smaller value depending on the function. It is important for them to be differentiable, to perform the optimization algorithms in the training process.

It is important to select the function properly in order to create an effective network. There are two properties types that activation function layers could introduce to the network: linear and non-linear. The identity function, the only linear activation function, could be expressed as Equation 3.4:

$$g(u) = u \quad (3.4)$$

As this is a linear function, the outputs will not be confined between any range. The graphic from this function could be seen in Figure 3.2a.

The non-linear activation functions are the most used activation functions because, unlike identity function, it eases the model to generalize or adapt to a variety of data and to differentiate between the output. Some examples are the logistic sigmoid, hyperbolic tangent or ReLU function, and, for multi-class problems, the softmax function.

The logistic sigmoid function, is expressed as in the Equation 3.5:

$$g(u) = \frac{1}{(1 + e^{-u})} \quad (3.5)$$

Sigmoid function values exists between 0 to 1. In particular, large negative numbers become 0 and large positive numbers become 1. The result from firing the node could from not firing at all (0) to fully-saturated firing (1). The graphic from this function could be seen in Figure 3.2b.

The disadvantage of the logistic sigmoid function is that it can cause a neural network to get stuck at the training time, because of the gradient value that is equal to zero in the majority of the function. However, when it is desired a probability as output, it is the proper function.

The hyperbolic tangent function, see Equation 3.6, it is similar to the logistic sigmoid function but with a range from -1 to 1. The advantage from this to the last one is that, here negative inputs will be mapped to -1, instead of being mapped to zero. The graphic from this function could be seen in Figure 3.2c.

$$g(u) = \frac{(e^{2u} - 1)}{(e^{2u} + 1)} \quad (3.6)$$

Hinton *et al.* in 2012 [42], proposed the Rectified Linear Units (ReLU). The ReLU function is described in Equation 3.7. The graphic from this function could be seen in Figure 3.2d.

$$g(u) = \max(0, u) \quad (3.7)$$

This type of function is easy to compute and differentiate (for backpropagation). Its range is from zero to infinite: if $u < 0$, $g(u) = 0$ and if $u \geq 0$, $g(u) = u$.

ReLU has become quite popular lately, mainly in deep learning models, bringing improvements relatively to sigmoidal activation functions, which have smooth derivatives but suffer from gradient saturation problems and slower computation.

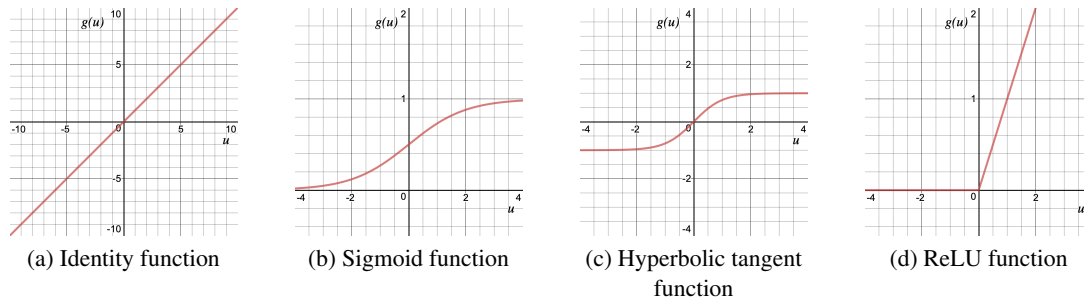


Figure 3.2: Identity, logistic sigmoid, hyperbolic tangent and ReLU functions graphics.

The softmax activation function, see Equation 3.8, is used in the output layer of the network, for multi-class classification problems:

$$y_j = \frac{e^{y_j}}{\sum_{i=1}^n e^{y_i}} \quad (3.8)$$

It converts each element to a value between zero and one by dividing them with the sum of the elements. Thus, the vector is normalized into positive values that sum to one and will correspond to probabilities of the inputs belonging to each class.

Multilayer Perceptron Although there are many ANN architectures, the Multilayer Perceptron is the most utilized model in neural network applications. MLP's are typically composed of three types of layers:

1. input layer - accepts inputs in different forms;
2. hidden layer - transform inputs and do several calculations and feature extractions;
3. output layer - produces the desired output, representing the class scores.

Each node in the network consists of certain random weights, and each layers has a bias attached to it that influence the output of every node. Certain activation functions are also applied to each layer to decide which nodes to fire.

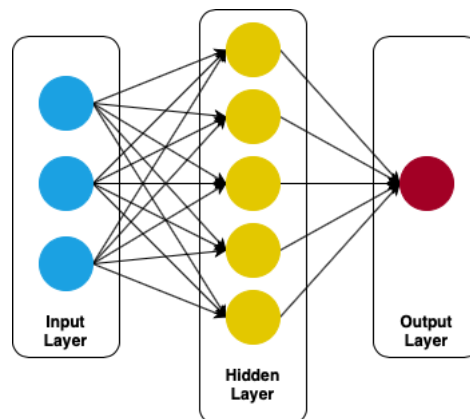


Figure 3.3: Multilayer perceptron, with one hidden layer.

Resuming and as could be seen in Figure 3.3, is received an input (a single vector) that is transformed through a series of hidden layers (in Figure 3.3 only exists one hidden layer). Each hidden layer is made up of a set of neurons, where each neuron in a single layer function completely independently and don't share any connections. However, each neuron in each hidden layer is fully connected to all neurons in the previous layers. In the end, the last fully-connected layer, the "output layer", represents the class scores [41].

Recently MLP fell out of favor because it has weights associated with every input, having with these reason dense connections that didn't allow them to scale easily for various problems, due to complexity and overfitting. A network architecture that does not have this problem is Convolutional Neural Network, being recently much more used [3].

Learning Process The ability to learn is the key concept of neural networks. Neural networks could learn in supervised and/or unsupervised manners. For this dissertation will be used supervised learning. Supervised learning is when it is taught to a computer how to do a thing, in other words, when is given to the machine not only a set of inputs but also the expected outputs to them. There are two types of supervised learning: classification and regression. The classification will be the type used in this dissertation.

LeCun *et al.* [26] presented, in 2015, a simple way to understand how classification works, by following a few steps. It is assumed that there is a set of images containing houses/cars/people and that it is intended to classify each image in each of these three categories. According to LeCun *et al.* [26], the first step to be taken is to collect a large number of images with the different selected categories, being each image labeled with its respective category. The weights need to be initialized to some values. There are different approaches, being random initialization the most common. Then, the training process starts:

1. Forward propagating each image to the machine, is produced an output in the form of a vector of scores, one for each category;
2. It is computed a cost function, that measures the error (or the distance) between the output scores and the desired scores (labeled data);
3. The internal parameters (weights) are adjusted to reduce the error of the computed cost function since the objective is to minimize this function.

There are various optimization algorithms to find the global minima of a cost function (the goal of step 3 in the last items list). Generally, in neural networks is used backpropagation algorithm, with gradient descent algorithm. The **backpropagation** algorithm begins with computing the cost function gradient (it starts with the final layer of weights), which can also be called the error value. This error value is then propagated back through the network, calculating layer by layer up to the first layer of weights, the error value to get the total error value.

Then, to adjust the weight vector values, the **gradient descent** algorithm is commonly used. As the gradient vector is also the derivative of the error value with the gradient vector computed,

for each weight, will be known the direction that is needed to move towards, to reach the local minimum (the minimum error). Founded that direction it is needed to nudge the weight, with a constant called learning rate, and it is needed to find how much to nudge the weight, to reach the local minima. With the gradient value and the learning rate step chosen, the current values of each parameters will be updated. This will happen until the weights converge.

After these steps, the training process (from 1 to 3) is repeated accordingly to the specified training number of epochs (the number of full training cycles). After that, the desired image to classify is feed in the network and it is predicted a category.

Lastly, the test step starts, evaluating the network with a different dataset (validation dataset) from the one used to train the network. The use of different network inputs helps to know how accurate its predictions are, assessing the level of generalization of the system.

Other optimization algorithm used, instead of the gradient descent is the Adaptive Moment Estimation, Adam, proposed by Kingma *et al.* [43]. It is an adaptive learning rate method, computing individual learning rates for different parameters.

3.1.2 Convolutional Neural Networks

Convolution Neural Network (ConvNet) first proposed by Fukushima in 1988 [44], only become more popular in 1998 when LeCun *et al.* applied a gradient-based learning algorithm to CNNs and obtained successful results for the handwritten digit classification problem [45]. This network type is essentially a Multilayer Perceptron that has been extended across space using shared weights. CNN's are designed to recognize images; they take advantage of the fact that the inputs consists of images and take them as arrays of pixel values, restricting the architecture in a more sensible way. They have neurons arranged in 3 dimensions: width, height, depth. (depth stands to the third dimension of an activation volume, see Figure 3.4). For example, if the input is 32x32x3 the image has a width of 32, height of 32 and three color channels (R,G,B). For each of these numbers is given a value from 0 to 255 which describes the pixel intensity at that point.

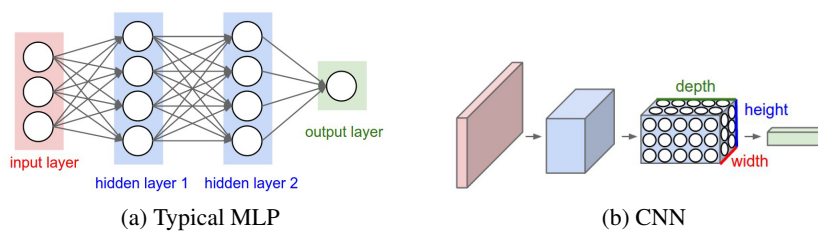


Figure 3.4: Comparison between a typical MLP and a CNN, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (red, green, blue) channels.) [9].

The hidden layer consists of several different layers which carry out feature extraction. These layers can be: convolution, nonlinear, pooling, normalization or fully-connected layers. For concluding, the output layer, that is also a final fully connected layer, recognizes the objects in the image.

Resuming, the way CNN's works is: given an image this image pass it through a series of convolutional, nonlinear, pooling and fully connected layers, and get an output. As said, the output can be a single class or a probability of classes that best describes the image. Following will be presented the different hidden layers from CNNs.

The **convolutional layer** uses a filter matrix (kernel) over the pixels array of image and performs a convolution operation to obtain a convolved feature map. An example of that process could be seen in Figure 3.5.

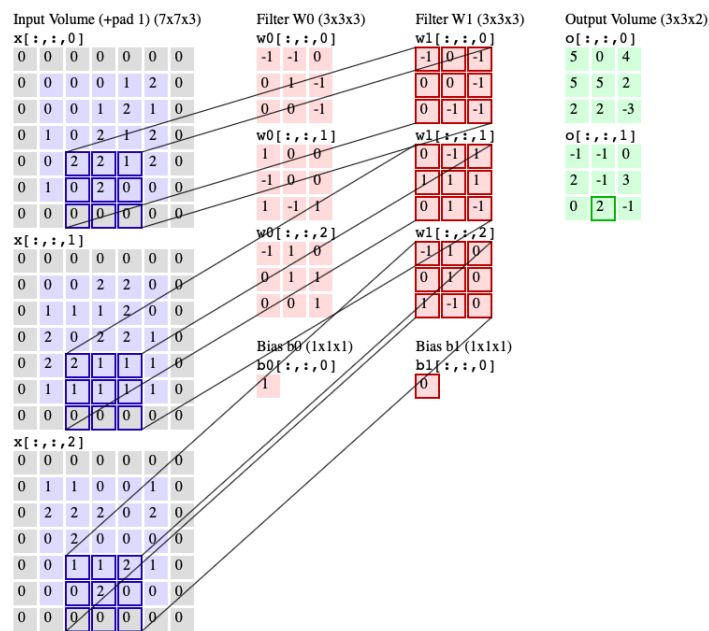


Figure 3.5: Example of a convolution operation, with two filters of size 3x3, applied with a stride (the length that the kernel is moving) of 2 [9].

The **activation functions** that are used are the same from section 3.1.1. In CNN its applied mostly ReLU activation function, first used in AlexNet, which was a deep CNN proposed in 2012 by Hinton *et al.* [42]. This activation function replaces all the negative values to 0 and remains same with the positive values. Mathematically can be expressed as in Equation 3.7 and the graphic from this Equation is shown in Figure 3.2d.

From all the **pooling layers** types, **max pooling** first proposed by Hinton *et al.* [42], is the most used since has been shown to work better in practice. Usually, after the convolutional and activation function, a max pooling layer is added that simplifies the feature map by reducing the number of parameters. In max pooling a kernel is used to extract the largest value within the receptive field of the feature map. For example, if the stride length is set to two and the kernel has the dimension 2x2, then the feature map will be reduced by a factor of two in both width and height (see Figure 3.6).

Generally, at the end of the network, there are **fully-connected layers**. They take an input volume (whatever the output is from the convolution, ReLU or pool layer preceding it) and output

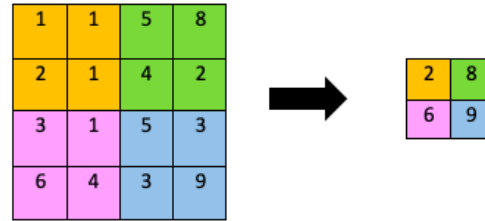


Figure 3.6: Max pooling layer operation with kernel size 2x2 and stride length 2.

the probability of belonging each class for the initial input.

An example of a CNN that outputs the probability of each class for each input can be seen in Figure 3.7.

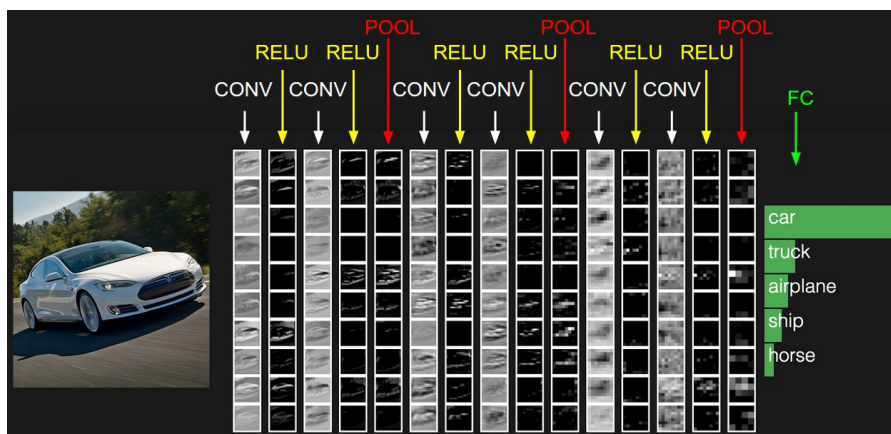


Figure 3.7: CNN example, with an output of 5 classes [9].

3.1.3 Learning

3.1.3.1 Training Techniques

There are different advanced techniques to apply to train the deep learning model efficiently like changing the activation function or learning rate, batch normalization, and network regularization approaches.

There are various **activation function** types (see section 3.1.1). The different activation functions can be switched according to its advantages or disadvantages to have a more effective network.

According to LeCun *et al.* [26] if the **learning rate** value is well chosen, it can make the training process faster. However, if chosen a small value for learning rate, it can take more time as the expected to the network converge. Instead, if it is chosen a high value for learning rate, the

network instead of converging may start diverging. A solution is to reduce the learning rate during training.

Sergey Ioffe and Christian Szegedy [46] introduced, in 2015, a method called **batch normalization**. Batch normalization aims to normalize the output distribution of every node in a layer. By doing so, it allows the network to be more stable.

One of the major problems of CNN's is the overfitting. Overfitting is the case where the weights of a neural network converge for the training dataset. Meaning that the network performs very good for the training dataset, but with any other data does not work very well. Some solutions to prevent overfitting are regularization methods like dropout, data augmentation or using k-fold cross validation.

Common regularization methods add a new term to the cost function, which influences the weights in certain ways. There are two widely used methods, among others, dropout and L2. The key idea of **dropout**, proposed first by Hinton *et al.* in 2012 [47], is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of different "thinned" networks. **L2** is to multiply to the term added to the cost function a penalty equal to the square of the magnitude of weights. It shrinks all the weights by the same factor, tending to end up with many small weights, but eliminates none. These methods significantly reduces overfitting.

Another way to prevent overfitting is doing **data augmentation**. Some popular augmentations are gray scales, horizontal flips, vertical flips, random crops, color jitters, translations or rotations. By applying just a couple of these transformations to training data, it can easily doubling or tripling the number of training examples.

K-fold cross-validation is a method used to identify overfitting. With a early clear description in [48], is a statistical method used to evaluate models on a new data sample, showing how well the model is generalizing. The general procedure starts with shuffling the input dataset randomly, and split the dataset into k groups. Then, for each fold, take a group as a test set and the others as training, and/or validation set. The model is trained with the training set, the hyperparameters are tuned with the validation set, and is evaluated on the test set. Then, the evaluation score is retained and the model discard. The summarize of the model skill is shown, using the sample of the evaluation scores of the k-models.

Once seen the context of the classification module on CADx systems, will be seen how to evaluate them.

3.1.3.2 Evaluation

Usually the CADx systems performance for diagnosis of lung cancer is evaluated through the ROC (Receiver Operating Characteristic) curve, more precisely through the area under the ROC curve. They are frequently used to show in a graphical way the connection/trade-off between clinical Sensitivity and "1-Specificity" for every possible cut-off for a test or a combination of

tests. Accuracy is also a evaluation method that is vital in all scientific endeavors. These evaluation methods are pretended to be used in this dissertation.

To better understand these concepts, when a test is evaluated there are four types of results. In general:

- Positive (P) = identified;
- Negative (N) = rejected;
- True positive (TP) = correctly identified;
- False positive (FP) = incorrectly identified;
- True negative (TN) = correctly rejected;
- False negative (FN) = incorrectly rejected.

Seeing Table 3.1, a better understanding of all components is obtained.

Table 3.1: Explanation of various cases with results of Disease vs Test.

		Test		
		-	+	
Disease	-	True Negative (TN)	False Positive (FP)	All without disease = FP + TN
	+	False Negative (FN)	True Positive (TP)	All with disease = TP + FN
		Negative	Positive	

Accuracy describes how closely a measured value approximates its true value. See Equation 3.9.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.9)$$

Sensitivity, also called the True Positive Rate (TPR) or recall, measures the proportion of actual positives, that are correctly identified as such. Is given by the Equation 3.10:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (3.10)$$

Specificity, also called the True Negative Rate (TNR), measures the proportion of actual negatives that are correctly identified as such. The specificity equation (see Equation 3.11) is given by:

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (3.11)$$

"1-Specificity", used in the trade-off in ROC curves, is given by the False Positive Rate (FPR), see Equation 3.13:

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR \quad (3.12)$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR = 1 - Specificity \quad (3.13)$$

Concluding, the trade in ROC curves is between all the cases that have True Positive Rate and False Positive Rate. The area under the ROC curve gives an idea about the benefit of using the test(s) in question, where the best cut-off has the highest True Positive Rate together with the lowest False Positive Rate (a greater area means a more useful test).

3.2 LIDC-IDRI Dataset

This section will approach the dataset that will be analyzed to use as CADx input in this dissertation.

The LIDC-IDRI, public Lung Image Database Consortium image collection, has proven to be very useful in validating system performance [4]. Its size, variability in nodule type, and great discrimination allied with the fact of being available to the public, gives scientists a splendid resource and helps to improve their area of research. This database consists of totally heterogeneous set of 1,018 chest CT image scans with a variable number (between 102 and 310) of 512x512 slices, from 1,010 patients (with a total of 243,858 thoracic CT).

The data was collected under the Institutional Review Board (IRB) obtaining approvals at seven academic institutions in the United States [49]. All of CT scans were annotated by 4 of 12 experienced thoracic radiologists, from the clinical institutions, under a rigorous two-phase image reading process (blinded and unblinded). In a first blind read phase, the radiologists annotate the scans independently. In the subsequent phase, each radiologist independently reviews their annotations, taking into consideration the marks of the other three radiologists. The annotations by radiologists were divided into three categories, regarding to diameter and to nodules and non-nodules:

- nodule's diameter greater or equal to 3 mm but smaller than 30 mm - it is assigned an identification number and given the rating scores between 1 to 5 with respect to the characteristics such as subtlety, internal structure, sphericity, margin, spiculation, texture, and malignancy and ratings between 1 to 6 with respect to calcification;
- nodule's diameter smaller than 3 mm - it is registered only the nodule's approximate centroid, and if the opacity is clearly benign, there is no marking. There are not given any rating scores for characteristics;
- non-nodule's diameter greater than 3 mm - it is indicated the nodule approximate centroids and there are not given subjective assessment of characteristics.

The database has 7,371 lesions marked as nodule by at least one radiologist, from which 2,669 of these have a diameter greater or equal to 3 mm.

Characteristics definitions A problem present in the LIDC-IDRI dataset is the ambiguity that exists in defining the characteristics evaluated, being the evaluation very subjective. It results in a disparity in the ratings given in the evaluations since physicians have different opinions when evaluating the same nodule. The only attempt made to define the characteristics and each class was made by Opulencia *et al.* [15]. The proposed definitions could be seen in Table 3.2.

Table 3.2: LIDC nodule characteristics, definitions and ratings proposed in [15].

Characteristic	Description	Ratings
Calcification	Calcification appearance in the nodule - the smaller the nodule, the more likely it must contain calcium in order to be visualized. Benignity is highly associated with central, non-central, laminated and popcorn calcification	Popcorn Laminated Solid Non-central Central Absent
Internal Structure	Expected internal composition of the nodule	Soft tissue - Fluid Fat Air
Lobulation	Whether a lobular shape is apparent from the margin or not-lobulated margin is an indication of benignity	Marked - - - None
Malignancy	Likelihood of malignancy of the nodule - malignancy is associated with large nodule size while small nodules are more likely to be benign. Most malignant nodules are noncalcified and have speculated margins.	Highly unlikely Moderately unlikely Indeterminate Moderately suspicious Highly suspicious
Margin	How well defined the margins of the nodule are	Poorly defined - - - Sharp
Sphericity	Dimensional shape of nodule in terms of its roundness	Linear - Ovoid - Round
Spiculation	Degree to which the nodule exhibits spicules, spike-like structures, along its border - spiculated margin is an indication of malignancy	Marked - - - None
Subtlety	Difficulty in detection - refers to the contrast between the lung and its surroundings	Extremely subtle Moderately subtle Fairly subtle Moderately obvious Obvious
Texture	Internal density of the nodule - texture plays an important role when attempting to segment a nodule, since part-solid and nonsolid texture can increase the difficulty of defining the nodule boundary	Nonsolid - Part-solid/mixed - Solid

However, this proposed definitions has some errors, e.g. the order of the ratings suggested to spiculation or lobulation. Also, although less, the ambiguity present in how to classify some characteristics remains. Thus, there is still a large gap in how to define the classes of some characteristics. Is expected that having more objective definitions for each class and each characteristic, the physicians' assessments are more consistent, producing fewer errors in the diagnosis.

Once obtained the essential information to implement the CADx classification module, and to understand the input dataset that will be used, a research was made in the state-of-the-art of lung nodule classification methods.

3.3 Lung Nodule Classification Methods

To develop an architecture to lung nodule classification in malignancy and characteristics, research was made. This research was carried out with a search limitation of articles dated by 2017 and 2018, using the LIDC-IDRI dataset, and based on deep learning, more precisely CNN's (since it has shown promising results without requiring prior segmentation of the nodule or handcrafted feature designs) that presented different architectures for extracting noticeable information from lung nodules.

Shen *et al.* [10] proposed a **multi-crop CNN**, that automatically extract nodule salient information. This proposed architecture has as main goals to discover a set of discriminative features and to capture the essence of suspiciousness-specific nodule information. This will be achieved underscoring the knowledge of the feature space, being the computation specified in the intermediate convolutional features rather than different scales of raw input signals, image space. This simplifies the conventional classification method, removing the segmentation and hand-craft features from the nodule, bridging the gap between raw nodule image and the malignancy suspiciousness. To achieve the malignancy suspiciousness by extracting highly compact features, they proposed a Conv + pool design, and a multi-crop pooling strategy, which is a specialized pooling strategy for producing multi-scale features, to surrogate the conventional max pooling. This strategy crops different regions from convolutional feature maps and then applies max-pooling different times. The proposed architecture could be seen in Figure 3.8 and the multi-crop pooling could be seen in Figure 3.9.

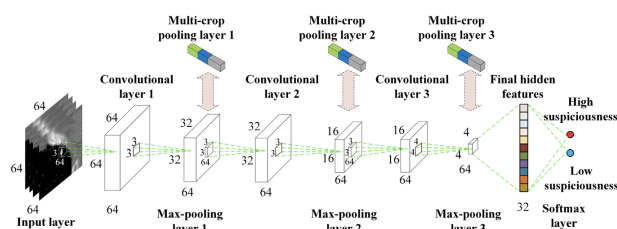


Figure 3.8: Multi-crop architecture [10]. The numbers along each side of the cuboid indicate the dimensions of the feature maps.

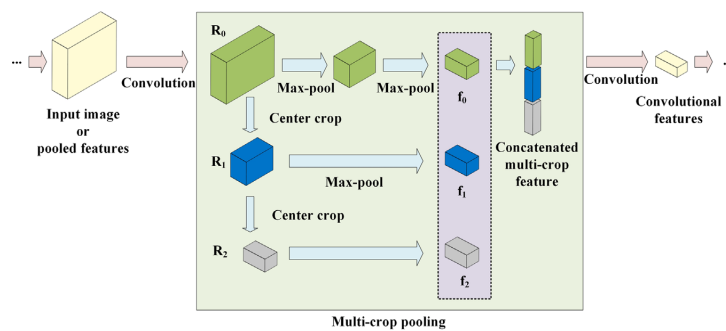


Figure 3.9: Multi-crop pooling [10].

Besides achieving nodule suspiciousness, this proposal characterizes nodule semantic attributes (subtlety and margin) and nodule diameter, than can be a useful predictor of malignancy. For predicting those outputs, they model them as binary classification. For the diameter estimation, the architecture was modified to be a regression model. To binarise the malignancy, subtlety and margin, they used all the nodules that had an average evaluation in LIDC-IDRI lower than 3, labelling them as 0, and higher than 3, evaluating those with 1. That resulted in a total nodules number of 756 for subtlety, 658 for margin, and 1375 for malignancy. The nodules that had an average evaluation equal to 3 were not used, except for malignancy. For malignancy prediction experiments were done, with and without the nodules that had an average of 3, to see how it was achieved better results. To training the model, it was employed five-fold cross-validation using in each experiment, three folds as training set, one as validation set, and another as test set. The results were obtained by averaging the accuracy and AUC scores across 30 times tests. These were compared with autoencoder [50], and massive-feat [51].

Liu *et al.* [11] used a **multi-view CNN** that takes multiple-views of each entered nodule as input channels. They took advantage from taking different information from the different views of the same nodule, as patches with small view areas can provide the details of the nodules while large views areas can provide information surrounding the tumor tissue.

The multi-crop view method, described before, used by Shen *et al.* [10] is similar to this one. The difference noted is that Shen *et al.* mainly improve the max-pooling operation and, in contrast, this method makes an adjustment to the input layers. The multi-view method proposed consists of:

1. Find the geometrical center of each nodule;
2. Centered on these points, crop patches in different sizes which give seven different view areas - 20x20 for the lesions without redundant space and for all the rest 30x30, 40x40, 50x50, 60x60, 70x70, and 80x80;
3. Resize them into the same size using spline interpolation before feeding them into different channel.

CT examples from the seven different views areas used in the three different types of ternary classification could be seen in Figure 3.10.

Another difference is that in the work of Shen *et al.* is only used the case of binary (benign

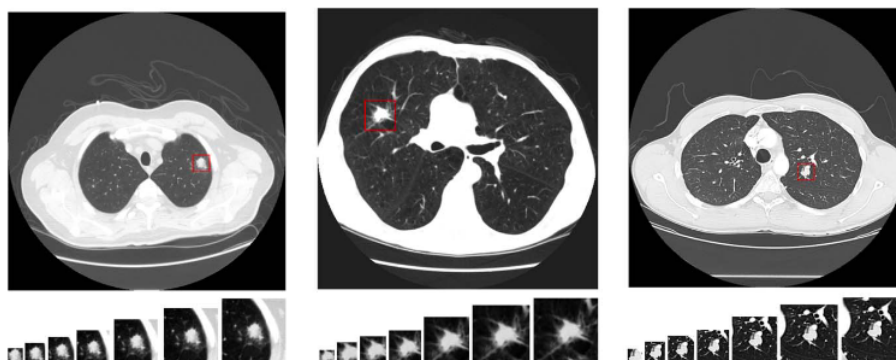


Figure 3.10: CT examples of the three types of ternary classification (from left to right, benign, primary malignant and metastatic malignant) with, in each one, the different 7 different views areas, from left to right 20x20, 30x30, 40x40, 50x50, 60x60, 70x70 and 80x80, taken from [11].

and malign) classification. In this work, they had two types of classification: binary and ternary (benign, primary malignant and metastatic malignant). The data preprocessing was different in their work. They choose to use the actual diagnosis of malignancy, from LIDC-IDRI, that was available for 157 patients. Extracting the information of the patients, it was only possible to obtain the diagnosis data for 96. The final diagnostic result was obtained using two levels, the nodule level, with each individual evaluation, and patient level, using the actual diagnosis obtained from the 96 patients. At each level, the nodules were marked as 0 - unknown, 1 - benign, 2 - malignant, primary lung cancer, or 3 - malignant metastatic.

For the binary classification, they labelled lung nodules as benign if the diagnosis result was 1 at the nodule level, and malignant if the diagnosis result was 2 or 3 at the nodule level. If the diagnosis result at the nodule level was 0, then it was labelled in accordance with the diagnosis result at the patient level. As each evaluation was counted as a different nodule it resulted in 764 benign lesions and 3530 malign. For the ternary classification, the same was done with the difference that when the diagnosis result was 3, it was labelled as metastatic malignant. Were obtained a total of 764 benign, 1431 primary malignant, and 2099 metastatic malignant lesions.

To train they used a 10-fold cross validation, using 9 samples to training and 1 for validating. This was done five times, and the results were averaged obtaining thus the final results.

The architecture of the multi-view strategy could be seen in Figure 3.12a.

Kang *et al.* [12] proposed a **3D multi-view CNN**. The advantage of a multi-view strategy has been shown to be useful for improving the performance in classifying lung nodules. In this work, its used the same multi-view strategy that in the work proposed by Liu *et al.*[11] but, its taken the advantage of using a 3D multi-view CNN, that is able to make full use of the spatial 3D context information of lung nodules, being the channel of the hidden layer a 3D feature vector. The same data preprocessing was made, and the same k-fold cross validation was employed, compared to [11]. They compared the use of 3D multi-view CNN with chain architecture with directed acyclic graph architecture, including the Inception1 (designed by Google that made use of different kernel sizes in the same convolutional layer [52]). For another test, they compared the use of 3D multi-

view CNN with a 2D multi-view CNN (proposed in [11]), to prove if the use of spatial 3D context information was really useful. Lastly, to see which multi-view strategy could achieve really better results, they compared the multi-view-one-network strategy with a one-view-one-network strategy used in the work proposed by Dou *et al.* [53]. The difference between the two strategies could be seen in Figure 3.11. In this article, they used also two classification types: binary and ternary, with the same subtypes that are used in [11].

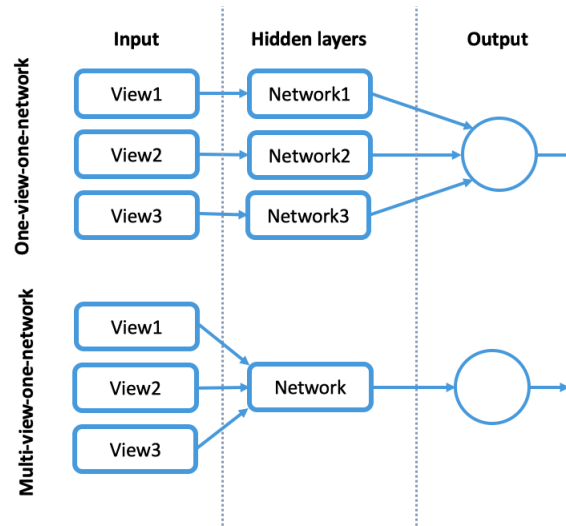


Figure 3.11: Multi-view's strategies: one-view-one-network(top) and multi-view-one-network (bottom).

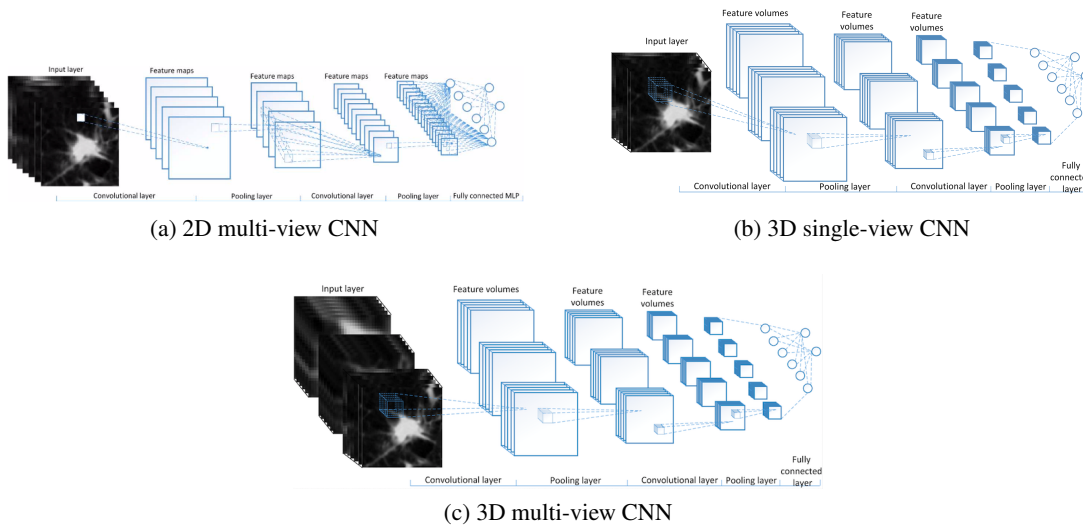


Figure 3.12: The architecture of 3.12a was taken from [11] and architectures from figures 3.12b and 3.12c were taken from [12].

To improve the specialist's confidence for CADx, Shen *et al.* [3] proposed the HSCNN, a novel interpretable deep **hierarchical semantic convolution neural network**. This work makes an improvement on the interpretability of the specialists for this type of system, giving not only

a binary malign/benign classification but also a binary characteristic classification. This network has two levels of outputs: low-level radiologist semantic features and a high-level malignancy prediction score. Therefore, this hierarchical design integrates both semantic features and deep features to predict malignancy. The information learned for each specific low-level semantic feature is then fed into the final high-level malignancy prediction task. The architecture could be seen in Figure 3.13. They used five features to predict: malignancy, sphericity, margin, subtlety, texture and calcification. They binarised these nodules characteristics, by averaging the scores for each nodule and assigning a threshold at 4 to distinguish scores of 1-3, Label 0, from scores 4-5, Label 1. Label 0 is characterized by being a benign nodule, poorly defined margin, lesser roundness, poor conspicuity between nodule and surroundings and a non-solid (ground-glass like) consistency; Label 1 being a malignant nodule, sharp margins, higher sphericity, high conspicuity between nodule and surroundings and solid consistency. Calcification was used differently, being 1 to 5 present of calcification and 6 absent of calcification. This division could be seen in Table 3.3.

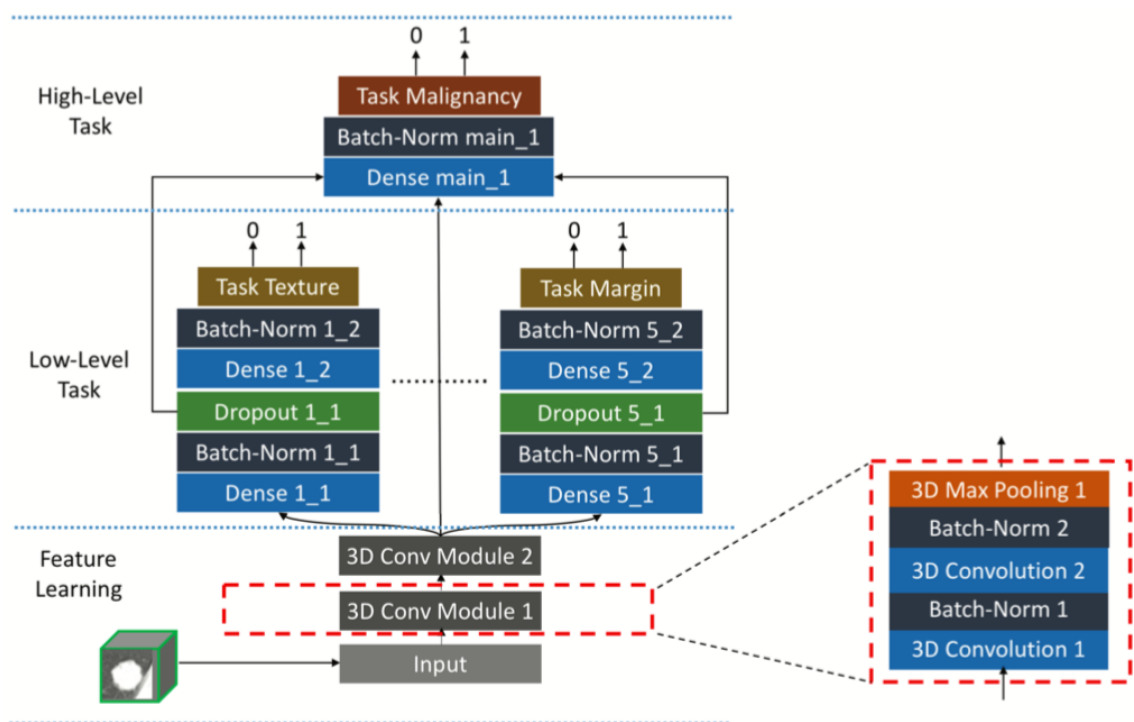


Figure 3.13: HSCNN architecture taken from [3].

To predict malignancy, they used, as a comparison, a 3D CNN. The architecture has the same structure as the HSCNN but without the low-level semantic task component. For features prediction, only HSCNN has been used.

In relation to the dataset usage, they only included nodules identified by at least three radiologists in four, and with a slice thickness smaller than 3 mm. Each one of the individual evaluations was counted independently as a different nodule, resulting in 4,252 nodule annotations. The only

Table 3.3: Binary labels proposed by Shen *et al.* [3] from LIDC-IDRI rating scales for nodule characteristics.

Nodules Characteristics	Label 0	Label 1
Malignancy	Scale 1-3 Benign	Scale 4-5 Malignant
Sphericity	Scale 1-3 Lesser roundness	Scale 4-5 High degree of roundness
Margin	Scale 1-3 Poorly defined margin	Scale 4-5 Sharp margin
Subtlety	Scale 1-3 Poor contrast between nodule and surroundings	Scale 4-5 High contrast between nodule and surroundings
Texture	Scale 1-3 Non-solid internal density	Scale 4-5 Solid internal density
Calcification	Scale 1-5 Present of calcification	Scale 6 Absent of calcification

semantic feature that this does not happened was calcification, were they averaged the ratings for each nodule. The CTs pixel values were normalized. Using the DICOM series header information, they transformed the values to Housenfield scales, and converted these from (-1000,500)HU to (0,1). For each candidate, they extracted a 3D cube of 40x40x40mm, centered around the candidate. Then they rescaled each cube to a fixed size of pixels, 52x52x52 voxels. For training the model they used a 4-fold cross validation, splitting the cases into four subsets with a similar nodules number, 2 for training, one for validation, and another for holdout testing, reporting the true model performance without information leakage. All subsets varied over cross-validation, never having the same as test set. They had also applied 3D data augmentation to the training and validation set, translating the position of the nodule within 4mm or flipping the nodule along one of the three axes. These were randomly picked in each training epoch.

3.4 Discussion

In this section, is presented a discussion about the researched work's conclusions.

The multi-crop convolutional neural network, proposed by Shen *et al.*, [10] has proved to achieve better results than the compared works, autoencoder, first mentioned in 1986 [50] and massive-feat (employing the strategy used in [51]). Comparing the network performance with and without the nodules with an average rating of 3 were obtained better results without them since the uncertain values introduced variation within each category. The results for classification accuracy of subtlety was 0.7432 and of margin 0.7699 (from 0 to 1).

Comparing the 2D multi-view CNN proposed by Liu *et al.* [11] with the 3D multi-view CNN proposed by Kang *et al.* [12], the last proved to achieve better results. From all the comparisons made in the Kang *et al.* work, the most relevant results were:

- the 3D multi-view CNN proved to achieve better results than 2D multi-view, for both binary and ternary classification;
- the error rate of the multi-view network is generally lower than that of the single-view network;
- the multi-view-one-network strategy achieve a lower error rate than one-view-one-network strategy.

For the proposed HSCNN and the used architecture for comparison, 3D CNN, Shen *et al.*, the achieved mean values for malignancy prediction could be seen in Table 3.4. After analyzing the results in the table, it can be concluded that the HSCNN proved to have better results than 3D CNN.

Table 3.4: Results comparison between 3D CNN and HSCNN.

Model	Accuracy	AUC	Sensitivity	Specificity
3D CNN	0.834	0.847	0.668	0.889
HSCNN	0.842	0.856	0.705	0.889

For semantic feature the average values obtained using HSCNN can be seen in Table 3.5.

Table 3.5: Results for semantic features predictions

Semantic Features	Accuracy	AUC	Sensitivity	Specificity
Calcification	0.908	0.930	0.930	0.763
Margin	0.725	0.776	0.758	0.632
Subtlety	0.719	0.803	0.673	0.796
Texture	0.834	0.850	0.855	0.636
Sphericity	0.552	0.568	0.552	0.554

HSCNN achieved results that suggest that it is able to learn feature representations while having high performance in malignancy prediction. Some limitations were found on this work by Shen *et al.* [3]. First, there was an exclusion of features that have a high association with malignancy, such as nodule size, spiculation, lobulation, and anatomical location [54, 55] due to known existent errors in LIDC-IDRI. Second, is the great variability existent by radiologists in reading semantic labels. One proposal by Shen *et al.* [3] to solve that is to limit semantic labels to those that have a high reader agreement rate. As a third restraint, the malignancy labels present in the LIDC-IDRI dataset are not from the pathological diagnosis, but from the radiologist's interpretation. Lastly, because the semantic features are staggered between 1 to 5 or 6, their binarisation could cause the loss of some important information.

For future work it is suggested by Shen *et al.* [3], the inclusion of discriminating semantic features, and investigate model variability using different semantic labeling schemes.

Concluding, the relevant results obtained for each CNN architecture presented could be seen in Table 3.6:

Table 3.6: Final best results for the different relevant CNN's architectures presented

Model	Accuracy	AUC	Sensitivity	Specificity
MC-CNN	0.87	0.93	0.77	0.93
3D MV-CNN(Binary Classification)	0.95	0.99	0.96	0.95
HSCNN	0.84	0.86	0.71	0.89

3.5 Summary

Deep learning have proven to outperform the results in lung nodule classification comparing to traditional classification methods. Most deep learning models are based in Neural Networks. MLP is the most used model in Neural Networks applications. Neural Networks are typically composed by an input and an output layer, with hidden layers augmented to it based on the complexity of the problem. Once the data is passed through these layers, the neurons learn and identify patterns. Once this model is trained, is asked to the network to find the predictions based on the test data. CNN, on the other hand, is a special type of neural network which works exceptionally well on images, a MLP that has been extended across space, using shared weights.

There are various techniques to improve the networks learning, and to prevent the overfitting problem. To evaluate the networks performance, are known various metrics such as accuracy, sensitivity, specificity, and area under the ROC curve. LIDC-IDRI will be the dataset used as basis for the network input. This database has a lot of known errors and variability in its semantic features assessments, essentially due to the lack of objective definitions for its evaluation.

After a restricted search for deep learning works in the literature, four that fit the restrictions were analysed and discussed: a multi-crop CNN, 2D and 3D multi-view CNN, and HSCNN.

In the multi-crop CNN was employed a multi-crop pooling strategy, instead of the conventional max pooling. This network predicted as outputs lung nodule subtlety, margin, diameter, and malignancy. The 2D and 3D multi-view CNN take advantage of a multi-view strategy to predict as output the lung nodule malignancy, obtaining the 3D multi-view CNN promising results, compared to the 2D. The last architecture presented, HSCNN, take advantage of LIDC-IDRI semantic features information, predicting as outputs the lung nodules' semantic features and malignancy. This way was given confidence to doctors when using CADx systems. From the presented, although it did not had the best results, HSCNN is the architecture that best fits the problem. Therefore, it will be the one used in this work.

Chapter 4

Methodology

The LIDC-IDRI has a known high variability regarding the physicians' assessments of lung nodules semantic features, especially in the most malignancy relevant features, lobulation and spiculation. Many possible reasons contribute to this high variability, e.g., the fatigue caused by reading CTs, the difficulty of distinguishing in the CT the abnormality from the lung structures, and the fact that criteria for assessing the semantic features are not exact, being the evaluation subjective.

The proposed solution in this dissertation to lower the variability in lung nodule dataset assessments is to have a more coherently annotated dataset. To accomplish this and to evaluate the impact it has on the malignancy prediction, a methodology was devised and is presented in Figure 4.1.



Figure 4.1: The methodology used for LIDC-IDRI review and impact evaluation in the malignancy prediction.

Firstly was made an LIDC-IDRI analysis, to get the information needed to define evaluations norms. Second, with the evaluation norms created and validated by the C-BER radiologist partners of the Centro Hospitalar de São João, the dataset was reclassified with it. Third, and finally, having the reassessed dataset, it was measured the impact of this on malignancy prediction, using HSCNN.

This chapter is organized by the three steps determined. In section 4.1 is shown how the LIDC-IDRI analysis was done, in section 4.2 is presented the approach used for the dataset reassessment, and in section 4.3 is explained the method used to evaluate the impact on the malignancy prediction.

4.1 LIDC-IDRI Analysis

In this section is firstly presented a tool developed to help in LIDC-IDRI analysis, section 4.1.1. With this tool, it was possible to analyse and understand how doctors evaluated LIDC-IDRI. It has been seen that the LIDC-IDRI documentation does not have specific information about each characteristic and how to evaluate it. In section 4.1.2, it is done a bibliographic review about the definitions for the characteristics, its classes, and is explained the relationship that they have with lung nodule malignancy. Thus, it was possible to come up with evaluation norms, to, after being validated by doctors, more consistently evaluate the data set.

The only characteristic present in the LIDC-IDRI evaluations that was not used to review was malignancy. Malignancy, being diagnosed by several factors, has no specific definition. For this reason, its evaluation must be done with the knowledge of a specialist, with the help of characteristics results. Considering this was decided to not review it. The characteristics reviewed were calcification, internal structure, lobulation, margin, sphericity, spiculation, subtlety, and texture.

4.1.1 Dataset Review Tool

In the LIDC-IDRI dataset, when the specialists, considered what was in the scans as a nodule, and with a diameter ≥ 3 mm but < 30 mm, they identified it with a number and gave rating scores from 1 to 5 for the characteristics subtlety, internal structure, sphericity, margin, spiculation, texture, and malignancy, and ratings between 1 to 6 with respect to calcification. This resulted in 2668 abnormalities evaluated at least one time out of four.

The analysis realized in the LIDC-IDRI dataset was done creating a worksheet file, with all the information considered relevant provided by the dataset. In this created file, each line represented each one of the 2668 considered nodules, and each column contained relevant information about it. The information present in the columns were, by order, as follows:

- Nodule information, with the columns: "nodule id", "patient number", "number of total evaluations given to the nodule";
- Evaluations' average given by the doctors, with the columns: "texture", "calcification", "margin", "sphericity", "lobulation", "subtlety", "spiculation", "internal structure";
- Classifications given by a doctor in evaluation 1, with the columns: "texture", "calcification", "margin", "lobulation", "subtlety", "spiculation", "internal structure";
- Classifications given by a doctor in evaluation 2, with the columns: "texture", "calcification", "margin", "sphericity", "lobulation", "subtlety", "spiculation", "internal structure";
- Classifications given by a doctor in evaluation 3, with the columns: "texture", "calcification", "margin", "sphericity", "lobulation", "subtlety", "spiculation", "internal structure";
- Classifications given by a doctor in evaluation 4, with the columns: "texture", "calcification", "margin", "sphericity", "lobulation", "subtlety", "spiculation", "internal structure";
- 2.5D nodule's images, with the columns: "image name in the axial plane", "image in the axial plane", "image name in the coronal plane", "image in the coronal plane", "image name in the sagittal plane", "image in the sagittal plane";

- 2.5D nodule's images with the average binary mask of the doctors, with the columns: "image name in the axial plane", "image in the axial plane", "image name in the coronal plane", "image in the coronal plane", "image name in the sagittal plane", "image in the sagittal plane".

All this information was considered essential to the LIDC-IDRI analysis. The columns of "nodule information" were substantial to identify the nodule, being the "patient number" determinant to, when there was doubt analysing the 2.5D images, consult the respective CT. The evaluations average, made possible to select and observe the nodules by the intervals of the doctors' average evaluation, becoming easier to understand how each characteristic, and each class was assessed by the physicians. The individual assessments were also crucial to analyze the consistency of the evaluations given for every nodule.

The 2.5D nodules images, with and without the binary mask, were also imperative to have a better perception of each nodule, in the three anatomical planes: axial, coronal, and sagittal. The inclusion of the binary mask images was important because many nodules were very small or had plenty matter around them, being difficult to identify. An example of its relevance could be seen in Figure 4.2.

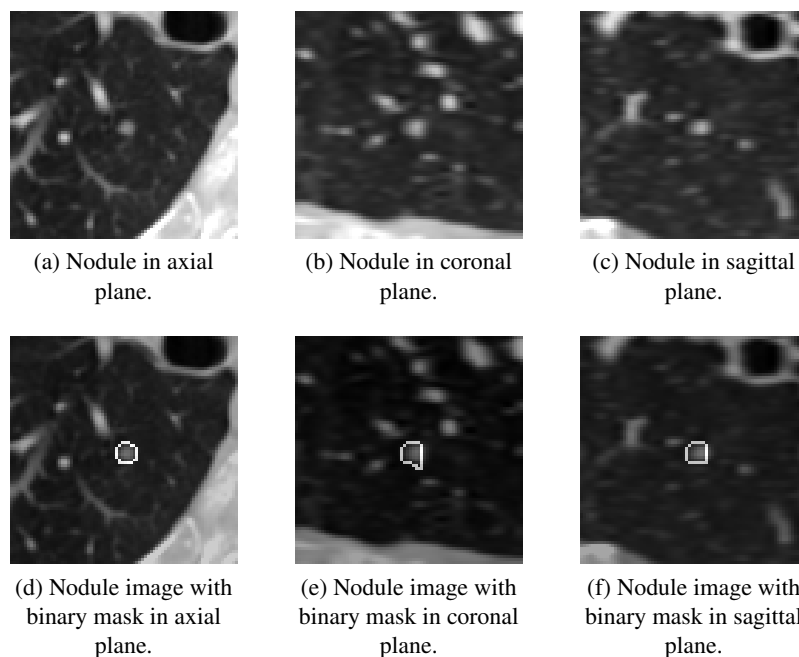


Figure 4.2: Nodule images from [4], without and with binary mask in axial, coronal, and sagittal planes.

To obtain these images for each nodule was made a data preprocessing. Each nodule had a correspondent .pkl file, provided by C-BER group. Each file was obtained after a processing of LIDC-IDRI data, having a dimension of $80 \times 80 \times 80$, with a space between the neighbouring voxel of 0.7. To visualize the nodules from the .pkl files had to be done a preprocessing, making a conversion from HU (-1000,400) to a binary scale (0,1). This HU values were previously used

in the literature to best see the nodules [56]. For each file, were extracted three planes with the dimension of 31.5x31.5 mm. This dimension was chosen because, being the nodules concerned centralized and with a diameter of less than 30 mm, it is thus possible to fully view them. With this dimension, was also not given much importance to the other structures around, as it was not needed. The process for having the nodule images with the binary mask was the same, but initially, for each nodule .pkl file, it was multiplied the respectively mask.

The image names for the three planes in the worksheet file were essential for automatically inserting the nodule images into each plane. As the images were a lot, was created a code that to each image name inserted the image existent in that directory, in the column after the image name. Appendix A shows a part of the worksheet file. By having the data organized in this way, a review of the LIDC-IDRI nodules and their characteristics have become simpler.

4.1.2 LIDC-IDRI Semantic Features Bibliographic Review

LIDC-IDRI database has no precise information about what is each semantic feature and how to specifically classify it, to our knowledge. To reevaluate the dataset in a more standardized form, adding to the review made in the worksheet file, was done bibliographic research.

To our knowledge, the only designations for the classes of each characteristic, based on LIDC-IDRI evaluations, found were the ones proposed by Opulencia *et al.* [15]. These classes designations could be seen in Table 4.1. They also provided in their work possible definitions for the evaluated characteristics in LIDC-IDRI, that could be seen in Table 3.2.

Nevertheless, this does not present exact information on how to evaluate the semantic features, having only definitions for each and designations for some of its classes. To find more about each characteristic was sought information about its definition, by what different types it could appear, and what influence it and its types could have on malignancy diagnosis.

4.1.2.1 Calcification

Definitions The characteristic calcification is defined as the appearance of calcification in the nodule by [15]. The LIDC-IDRI dataset has 6 rating terms, with proposed definitions by Opulencia *et al.* [15] as it could be seen in Table 4.1. They are popcorn, laminated, solid, non-central, central, and absent, by order.

Was challenging to find a definition for popcorn calcification in the literature. The closest definition found was in BI-RADS: coarse (popcorn-like) calcification, usually dense [57]. After a deeper analysis, seeing popcorn calcification images, an approximate definition was inferred: a popcorn shape calcification in a nodule. In Figure 4.3a could be seen an LIDC-IDRI example.

A definition of laminated calcification is lacking in the literature. From the observation of the nodule images present in the literature, and in the worksheet file, the definition that came closest is the presence of calcification in a laminated form in a nodule. An example of this calcification type could be seen in Figure 4.3b.

Table 4.1: Characteristics and their classes in LIDC-IDRI dataset proposed by Opulencia *et al.* [15]

Characteristic	Rating	Rating Definition
Calcification	1	Popcorn
	2	Laminated
	3	Solid
	4	Non-central
	5	Central
	6	Absent
Internal Structure	1	Soft tissue
	2	-
	3	Fluid
	4	Fat
	5	Air
Lobulation	1	Marked
	2	-
	3	-
	4	-
	5	None
Malignancy	1	Highly unlikely
	2	Moderately unlikely
	3	Indeterminate
	4	Moderately suspicious
	5	Highly suspicious
Margin	1	Poorly defined
	2	-
	3	-
	4	-
	5	Sharp
Sphericity	1	Linear
	2	-
	3	Ovoid
	4	-
	5	Round
Spiculation	1	Marked
	2	-
	3	-
	4	-
	5	None
Subtlety	1	Extremely subtle
	2	Moderately subtle
	3	Fairly subtle
	4	Moderately obvious
	5	Obvious
Texture	1	Non-solid
	2	-
	3	Part-solid/mixed
	4	-
	5	Solid

After analyzing the literature and the worksheet file, an approximate definition of solid calcification was concluded: the presence of calcification in all or almost the whole part of the nodule. This calcification's type is in a way related to the solid texture type, see section 4.1.2.8, since a nodule with solid calcification is regularly a solid nodule at texture level. A Figure of a solid calcification nodule could be seen in 4.3c.

For non-central calcification, not much information has been found, only mentioning in [15] that this class has a vague definition. After a review it was found that it is more popular known by eccentric. Analysing the worksheet file and the information found in the literature, a definition was reached: a calcification type with a regular shape and with a considerable size, located peripherally in the nodule. An example from LIDC-IDRI dataset can be seen in Figure 4.3d.

A center calcification definition is missing. After seeing nodules examples in the literature and LIDC-IDRI dataset, a definition for this calcification type was established as calcification in the nodule's center. An example of this type could be seen in Figure 4.3e.

Lastly, for absent, the definition found in [15] was the absence of calcification in a nodule. LIDC-IDRI example could be seen in 4.3f.

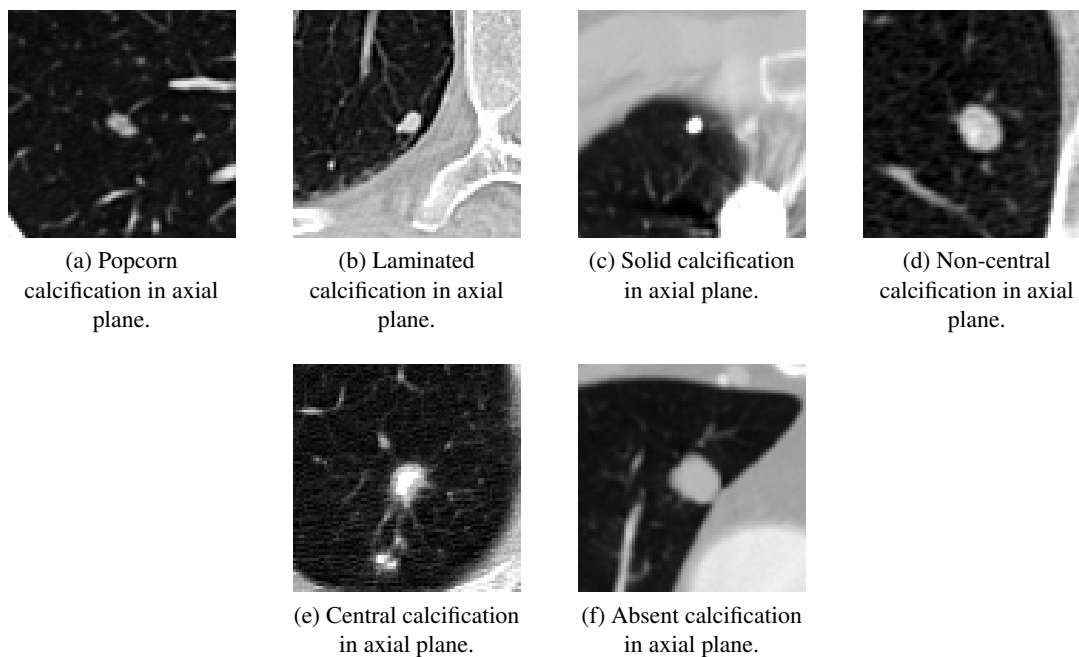


Figure 4.3: Popcorn, laminated, solid, non-central, central and absent calcification example nodules from LIDC-IDRI dataset [4].

Malignancy relation From the eight characteristics evaluated in LIDC-IDRI (excluding malignancy), this characteristic is the fourth with more relevance for lung nodule malignancy classification [29].

Calcification in a nodule is generally associated with benign conditions. However, after research, there was found that there are some calcification types related to benignity and some

associated with malignancy. In the interpretability done by [15], the types central, non-central, laminated, and popcorn are related to the benignity, and all other patterns should not be considered a sign of benignity. However, in [58, 59] was found that the types that are correlated with benignity are the diffuse/solid, central, laminated/concentric or popcorn, being the malignancy-related types non-central/eccentric and stippled/ground-glass. Stippled calcification is represented by a nodule with numerous small dots or specks of calcification. This last malignancy relation was also confirmed by Hospital Sao Joao physicians. Absent calcification does not represent any indication related to malignancy. In Figure 4.4 could be seen this malignancy division.

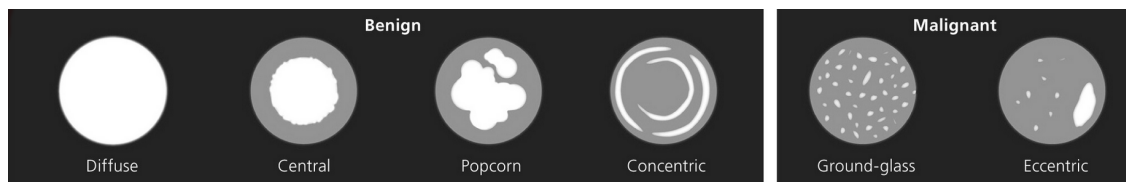


Figure 4.4: Malign and benign types of calcification by [13].

Concluding, dense and uniform calcifications are frequently encountered and strongly related with benignity [19]. In malignant nodules these are more diffuse, amorphous or punctate, few in number and more eccentric in location [60, 61].

4.1.2.2 Internal structure

Definitions Internal structure is defined by Opulencia *et al.* [15] as the internal composition of the nodule. In LIDC-IDRI dataset, it is evaluated in 5 classes. The classes defined are: 1 - soft tissue, 3 - fluid, 4 - fat and 5 - air.

A nodule with internal structure as soft tissue is composed entirely by a soft tissue. An example could be seen in Figure 4.5a. Concerning fluid' internal structure, the only information found in the literature that related to it, was found in [15], stating that fluid is a continuous amorphous substance, that tends to flow and to conform to the outline of its container, i.e. a liquid or a gas. Although this description could be related to this characteristic, in the literature was not found a definition about what is this in a nodule. As there is not a nodule in LIDC-IDRI dataset evaluated by 3 or more doctors as fluid, was not possible to find an example in the dataset. Fat is a nodule with fat presence [62]. A nodule with fat internal structure could be seen in Figure 4.5b.

An air internal structure is a nodule composed with a mixture of gases [15]. In Figure 4.5c could be seen an example.

Malignancy relation The fat' presence in a nodule is in up to 50% of benign nodules [63]. However, although very rare, sometimes malign nodules can present fat-containing lesions [64, 65]. A nodule with an air' internal structure is usually seen in benign cases [64]. Soft tissue nodule is dependent on other characteristics to obtain an assertion about its correlation with malignancy.

In Shen *et al.* work [3] is perceived that this feature in LIDC-IDRI has not shown to have a high disparity in the evaluations and, being almost always classified as soft tissue (being the average

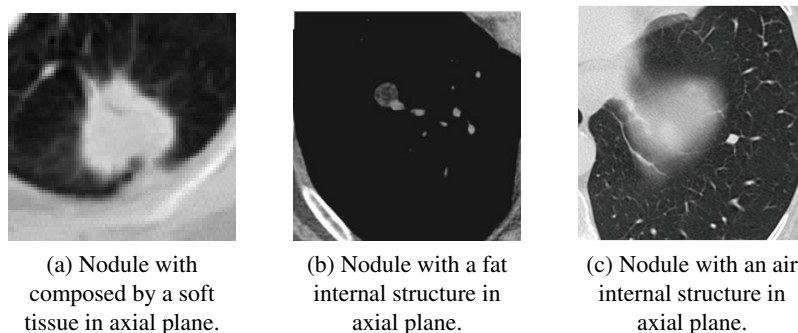


Figure 4.5: Internal structure's examples. 4.5a is from LIDC-IDRI dataset [4], and 4.5b and 4.5c are from [14].

evaluation given in 2641 of 2668 nodules) , does not have a high impact on the determination of a nodule's malignancy. It is also shown by Hancock *et al.* [29], that the addition of this feature is not very relevant for lung nodule malignancy classification, being the second less relevant. For that reason, was not created an evaluation norm to this characteristic, thus not being reassessed, and not used as an input for HSCNN.

4.1.2.3 Lobulation

Definitions Lobulation has been defined in the literature as the presence in the margin of lobules or of something that resembles a lobule [15]. The classification shown in Table 3.2, or in Table 4.1, is categorized inversely and it has been reported that there are 399 known cases in the LIDC dataset for which a subset of 100 may have been annotated using the inconsistent rating system for spiculation and lobulation [29]. Lobulation is evaluated in LIDC-IDRI dataset from class 1 to 5, where, correcting previous errors, 1 is "no lobulation present", and 5 is "lobulation present". Nodule examples from classes 1 and 5, from LIDC-IDRI dataset, could be seen respectively in 4.6a, and 4.6b.

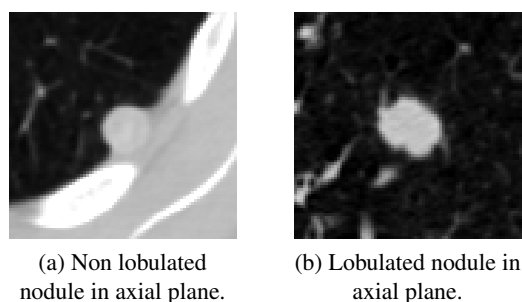


Figure 4.6: Non lobulated and lobulated nodules examples from LIDC-IDRI dataset [4].

There is not a definition in the literature for what is each degree of lobulation and how to define it. For that reason, rating a nodule in LIDC-IDRI in the range from 2 to 4 (including) is very subjective. That results in a high disparity between the doctors' ratings regarding the same

nodule. To reach to the different classes definitions it will be indispensable the use of the dataset review tool.

Malignancy relation In Hancock *et al.* work [29] is proved that lobulation is the second most relevant feature for lung nodule malignancy classification, from eight evaluated characteristics in LIDC-IDRI (excluding malignancy).

In Opuencia *et al.* work [15], see Table 3.2, this semantic feature is incorrectly pointed as an indication of benignity. Lobulation in a nodule is attributed to different or uneven growth rates in the margin, a finding that is highly associated with malignancy [58]. In [59, 64] appears as a margin sign with intermediate probability of malignancy. The reason it appears as intermediate probability is because although the lobulated contours of a lesion are more frequently associated with malignant pathology, they are also found in benign nodules [66].

However, a lobulated margin has an increased likelihood for lung cancer compared to smooth, round, or polygonal shape [65, 67, 68]. In conclusion, a nodule with a lobulated margin is more relevant for a malignancy diagnosis than a nodule that is not lobulated.

4.1.2.4 Margin

Definitions The characteristic margin as its defined in [15] measures how well defined the margins are, see Table 3.2, being the margin the boundary line or the area immediately inside the boundary. In the LIDC-IDRI dataset it is classified from 1 to 5, with only the extremes defined in [15] with class 1 as poorly defined, and class 5 as sharp. "Poorly defined" is known also as a ill-defined nodule, and "sharp" , also known as a smooth margin, is defined as a continuous regular outline [67], having well-marginated borders. Figures from poorly defined and sharp margins nodules could be seen respectively in Figure 4.7a, and Figure 4.7b.

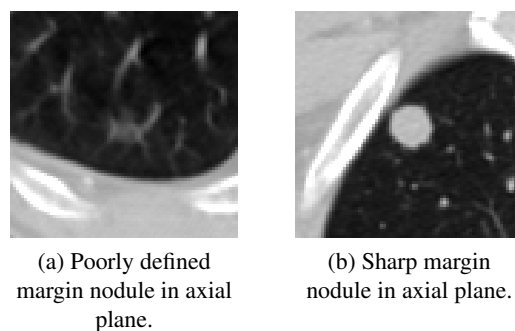


Figure 4.7: Poorly defined and sharp margins nodules from LIDC-IDRI dataset [4].

Malignancy relation From the eight characteristics evaluated in LIDC-IDRI (excluding malignancy), this characteristic is the fourth with less relevance for lung nodule malignancy classification [29].

Smooth margin nodule, with well-marginated borders, are more likely to be benign lesions. On the other hand malignant nodules usually have ill-defined, irregular contours [64, 65]. However, although it is more common smooth margins in benign solitary nodules, it does not exclude malignancy. About 21% up to one third of malignant solitary nodules have smooth margins [58, 69, 70]. However, in conclusion, a nodule with a poorly-defined margin has more probability to be malignant than a nodule with a well-defined margin.

4.1.2.5 Sphericity

Definitions Sphericity is defined in [15] as nodule's shape according to his sphericity. In LIDC-IDRI is classified from 1 to 5. The classes proposed by Opulencia *et al.* [15] for LIDC-IDRI are classes 1 as linear, 3 as ovoid, and 5 as round, see Table 4.1. Examples from linear and round nodules could be seen respectively in Figures 4.8a, and 4.8b.

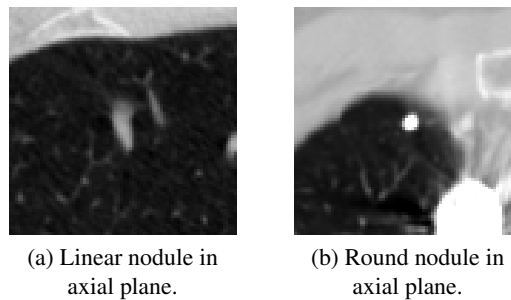


Figure 4.8: Linear and round nodules examples from LIDC-IDRI dataset [4].

Malignancy relation The typical shape of a solitary nodule is round or oval. Irregular or complex shapes are known to have a correlation with malignancy as round and oval are correlated with benignity. As this ratings only measures the nodule's sphericity, from linear to round, it does not add much information to the malignancy classification. Hancock *et al.* [29] shown also that the addition of this feature is not very relevant for lung nodule malignancy classification. For that reason, was not created an evaluation norm to this characteristic, thus not being reassessed, and not used as an input for HSCNN.

4.1.2.6 Spiculation

Definitions The most approximate definition found for spiculation is the presence of strands extending from the lung margin into the lung parenchyma [67]. Another definition found to evaluate spiculation in LIDC-IDRI, proposed in [15], is the degree to which the nodule exhibits spicules, spike-like structures, along its border - a spiculated margin.

As in lobulation, the ratings shown in Table 4.1 or in Table 3.2, are categorized inversely. Has been reported that there are 399 known cases in the LIDC-IDRI dataset for which a subset of 100 may have been annotated using that inconsistent rating system for spiculation and lobulation [29]. The feature ratings in LIDC-IDRI evaluations are from 1 to 5, where, correcting the proposed

definitions in [15], class 1 is none spiculation and 5 spiculation marked. So far, only the classes extremes have been defined. Examples of class 1 could be seen in Figure 4.9a, and of class 5 in Figure 4.9b.

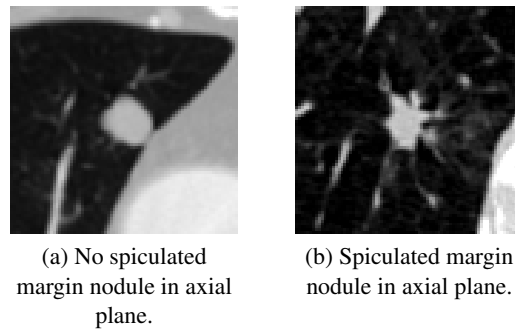


Figure 4.9: No spiculated and spiculated nodules examples from LIDC-IDRI dataset [4].

As lobulation, there is not a definition in the literature for what is each degree of spiculation and how to define it. For that reason, rating a nodule in LIDC-IDRI in the range from 2 to 4 (including) is very subjective. That results in a high disparity between the doctors' ratings regarding the same nodule. To reach to the different classes definitions it will be indispensable the use of the dataset review tool.

Malignancy relation In Hancock *et al.* work [29] was proved that spiculation is the most relevant feature for lung nodule malignancy prediction, from the eight evaluated characteristics in LIDC-IDRI (excluding malignancy). A nodule with a spiculated margin is a highly predictive of malignancy with a positive predictive value up to 90% [69], much more likely to be malignant than one with smooth edges, round, or polygonal shape [60, 65, 67, 68, 71].

4.1.2.7 Subtlety

Definitions Subtlety is related to the difficulty of nodule detection - refers to the contrast between the lung and its surroundings [15]. Subtlety is a subjective visibility rating of a lesion in a medical image, ranging from not visible to obvious. In LIDC-IDRI this feature has a division in 5 classes, defined in [15], see Table 4.1. Class 1 is defined as extremely subtle, 2 as moderately subtle, 3 as fairly subtle, 4 as moderately obvious, and 5 as obvious.

Malignancy relation From the eight characteristics evaluated in LIDC-IDRI (excluding malignancy), this characteristic is the third with most relevance for lung nodule malignancy classification [29]. Was not found information in the literature on how this feature is relevant for malignancy, being most relevant for segmentation. However, given the importance that it has concerning the malignancy prediction, this feature was chosen to be used to the dataset reassessment.

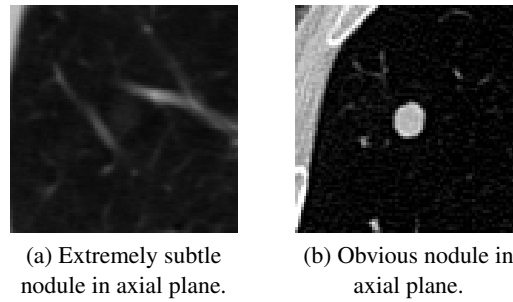


Figure 4.10: Extremely subtle and obvious nodules examples from LIDC-IDRI dataset [4].

4.1.2.8 Texture

Definitions In LIDC-IDRI dataset the defined ratings to evaluate this characteristic were from 1 to 5 with clear definitions proposed in [15], in class 1 as non-solid, class 3 as part-solid/mixed, and class 5 as solid. As described previously in section 2.2, solid nodules have uniform intensity distribution, part-solids are ground-glass nodules with density components, and non-solids are pure ground-glass nodules. Images of these three types could be seen in Figure 2.5.

Malignancy relation From the eight characteristics evaluated in LIDC-IDRI (excluding malignancy), this characteristic is the third with less relevance for lung nodule malignancy prediction [29]. In a study by Henschke *et al.* [59], was found that part-solid nodules have a malignancy rate of 63%, pure ground glass nodules of 18% and solid nodules only 7%. With this, it is concluded that nodules that contain a ground-glass component are more likely to be malignant than solid nodules [72, 73].

After this research, was decided to only create evaluation norms for six characteristics (calcification, lobulation, margin, spiculation, subtlety, texture), excluding this way the internal structure and sphericity.

4.2 Dataset Reassessment

After presenting the first step of methodology and its two methods, used to create the evaluation norms, the second step is presented. The tool created to the LIDC-IDRI analysis, had also an important role in this step. The dataset reassessment was based on the evaluations norms, 2.5.D .png nodule images, and the evaluation given by the doctors and its average. The axial view was favored, in cases of uncertainty in the classification, to removing the doubts generated by the visualization of the 2.5D images. The process used for the data reassessment, was made as could be seen in Figure 4.11. In this method, $m = [0, 6]$, where m is the number of characteristic being reassessed; $l = [0, N]$, where l is the number of the class being analysed, and N is total number of classes for the characteristic m ; and $i = [0, J]$, where i is the abnormality in question and J is the total number of abnormalities in LIDC-IDRI for l class.

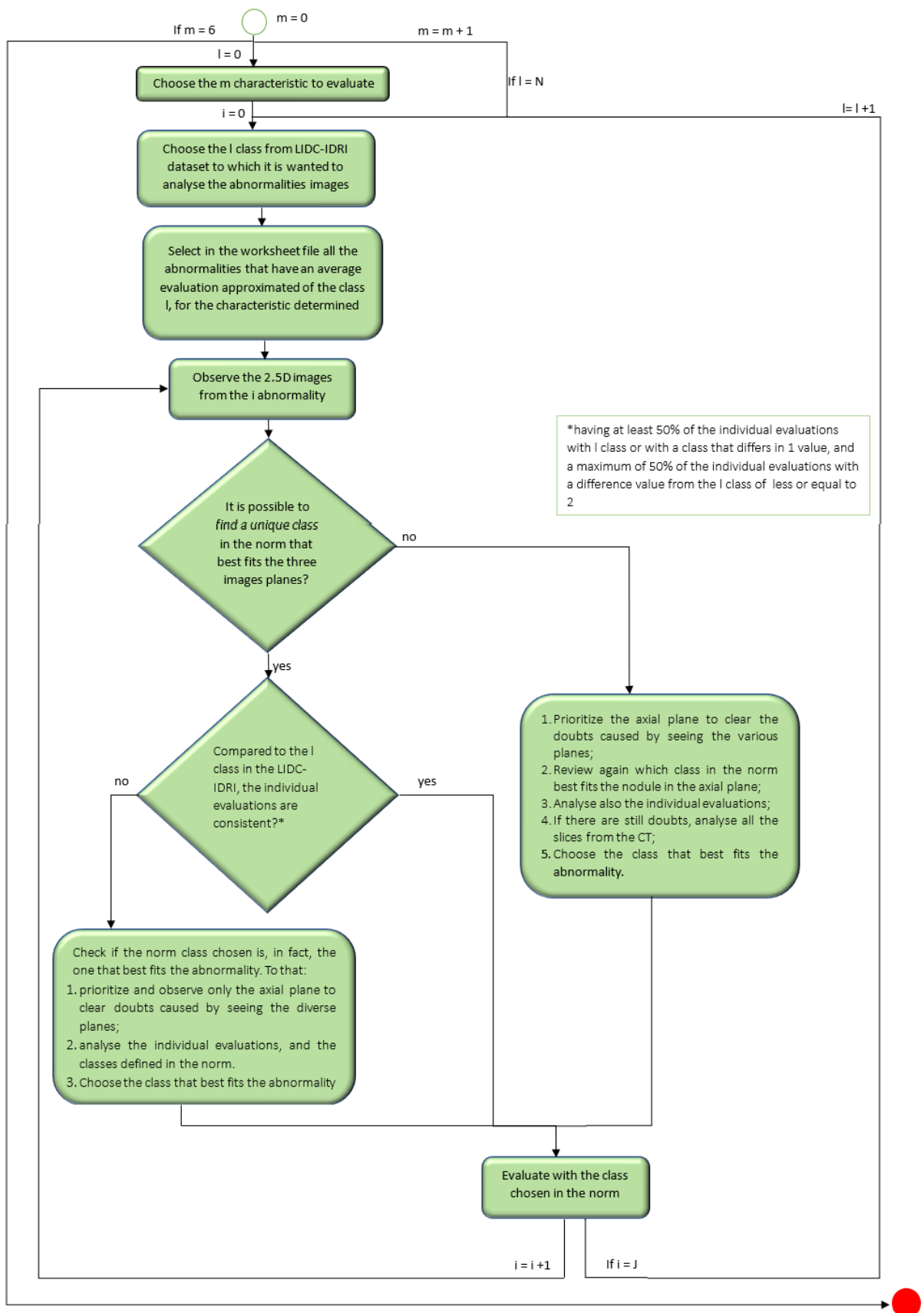


Figure 4.11: Method used to do the dataset reassessment.

Employing this method was possible to reassess the full dataset in a more consistent form.

4.3 Impact Evaluation

After having with the second method a more coherently reassessed dataset it was needed to see the impact that it had in the malignancy prediction. To evaluate this impact, the resultant dataset was used as input in HSCNN, an interpretable hierarchical semantic convolutional neural network. This way, comparing the results obtained with those of Shen *et al.* [3] it will be possible to answer to the question done in chapter 1. All the information about the architecture was obtained from [3].

This network architecture utilizes as input a 3D cube, capturing the lung nodule, and has two outputs, the prediction of lung nodule semantic diagnostic features and malignancy. To do that, this architecture is composed by 3 modules: feature learning, low-level task, and high-level task. To a better understanding, see Figure 3.13. Each component has a fundamental role in the two outputs predictions. The feature learning module learns the image features, in an adaptive form, which are generalizable over different tasks. The second module, low-level task, predicts the semantic features. In this dissertation, this module will predict six semantic features, which are texture, margin, subtlety, spiculation, lobulation, and calcification. The high-level task module concatenates information from the feature learning module and all the low-level tasks to predict the lung nodule malignancy.

Feature learning The feature learning component is represented in the Figure 3.13. The first processing unit is composed of two blocks, having each two 3D convolutional layer, with a kernel size of 3x3x3, each one followed by a batch normalization. At the end of each block, there is a 3D max-pooling layer. This approach is performed on the input feature maps along all the dimensions of the input patch, to produce an output defined as in Equation 4.1:

$$f^j = \sum_i c^j * f^i + b^j \quad (4.1)$$

where f^j and f^i are the j^{th} output feature map i^{th} input feature map, respectively. c^j is the j^{th} convolution kernel and represents the 3D convolutional operation between the convolution kernel and input feature map. b^j is the j^{th} bias corresponding to the j^{th} convolution kernel. Batch normalizations [46] were applied to the output feature maps, after each 3D convolutional layer. This way, the training process was accelerated, and the features maps normalized, by reducing the internal covariate shift. See more about batch normalization in section 3.1.3. The activation function chosen to take the output from batch normalization, was ReLu [42], see more in section 3.1.3. The Equation of this nonlinear function could be seen in 3.7.

In the first block, for both 3D convolution layers, 16 features maps were used. For the second, again for both 3D convolutional layers, were used 32. To control overfitting, and reduce the number of parameters, a 3D max-pooling layer was applied at the end of each block. By employing

it, the spatial size of the feature maps was progressively reduced. The 3D max-pooling layer is defined as:

$$\begin{aligned} \hat{f}_{x,y,z}^i = \max\{f_{x',y',z'}^i; & x' \in [x \cdot s_x, x \cdot s_x + d_x - 1], \\ & y' \in [y \cdot s_y, y \cdot s_y + d_y - 1], \\ & z' \in [z \cdot s_z, z \cdot s_z + d_z - 1]\} \end{aligned} \quad (4.2)$$

where x , the row index, y , the column index, and z , the depth index, start from zero. Here, s is the stride size (downscale factor) and d is the size of the max pooling window. The pooling window size used was the same used by Shen *et al.*, $d = (2,2,2)$. The stride used was $s = (1,1,1)$ down sampling the input features maps by a factor of 1, across all the image cube dimensions. This stride dimensions were different from the used by Shen *et al.*, that were $(2,2,2)$, since it has shown better results. The pooling layer has no learnable parameters.

In the end, the resulted output features were fed into the low and high-level task, at the same time.

Low-level task module The low-level task module predicts six semantic features and provides information from it to the high-level task module. This low-level module is composed by six equal branches, each one addressed to a different semantic feature. The semantic features predicted in this module are texture, margin, subtlety, spiculation, lobulation, and calcification. These were the characteristics chosen to reassess, in chapter 4.

Analyzing the output predictions of this module, will be seen if a more coherent evaluated LIDC-IDRI dataset, will improve the semantic features predictions results, compared to those of Shen *et al.*

In each of the six branches, there are two equal blocks, each one initialized by a fully-connected layer, see more about what it in section 3.1.2. In the first block, the fully-connected layer was employed with 256 neurons, and in the second with 64. Each one was followed by a batch normalization. ReLu was the activation function chosen to take the output from batch normalization. To control overfitting, along with the batch normalization, was also used the dropout technique between the two blocks. For more information about dropout, see section 3.1.3.1.

High-level task module The function of the high-level task module is to predict the lung nodule malignancy. With this output prediction it will be possible to know if, a more consistent LIDC-IDRI dataset, will improve the malignancy prediction in a nodule.

This task receives as input the concatenation of two outputs from the modules before: from the feature learning component and the low-level task first block. To a better understanding, see Figure 3.13.

The first block from the low-level task was chosen instead of the second because comparing the two, the first block output has less specific, but salient information, generalizable across tasks, for lung nodule malignancy prediction. The second block, on the contrary, is trained to extract more specific representations for each sub task.

This way is taken advantage of the features learned from the feature learning module, and of the general representations learned in each sub task.

These outputs concatenated were used as input for the high-level task module. This component starts with a fully-connected layer with 256 neurons, followed by a batch normalization. The activation function used was the soft max. Allied to the batch normalization, a dropout was also applied, to reduce overfitting. This resulted in a lung nodule malignancy prediction.

Loss functions The global loss function used to optimize the HSCNN was the one proposed by Shen *et al.*. This proposed function maximize the probability of predicting the correct label for each task. The equation that defines this loss function could be seen as follows:

$$L_{global} = \frac{1}{N} * \sum_{i=1}^N \left(\sum_{j=1}^6 \lambda_j \cdot L_{j,i} + \lambda_M \cdot L_{M,i} \right) \quad (4.3)$$

where N is the total number of training samples and i indicates the i_{th} training sample. j is the j_{th} sub task and $j \in [1, 6]$. λ_j is the weighting hyper parameter for the j_{th} sub task. $L_{j,i}$ represents the loss for sample i and task j . $L_{M,i}$ is the loss for the malignancy prediction task for the i_{th} sample. λ_M represents the weighting hyper parameter for the malignancy task.

Each loss component, $L_{j,i}$ and $L_{M,i}$, is defined as weighted cross entropy loss, as could be seen in the Equation 4.4.

$$L_{j,i} = -\log\left(\frac{e^{f_{y_i,j}}}{\sum_n e^{f_{y_n,j}}}\right) \cdot \omega_{y_i,j} \quad (4.4)$$

where y_i is true label for the i_{th} sample (x_i, y_i) . Here, y_i equals 0, 1. $f_{y_i,j}$ is the prediction score of the true class y_i for task j and $f_{y_n,j}$ represents a prediction score for class y_n . They use $\omega_{y_i,j}$ to represent the weight of class y_i for task j . The new dataset, as the original LIDC-IDRI, has a lot of label imbalance in all the tasks. To reduce the training bias, introduced by the imbalance, is fundamental the use of $\omega_{y_i,j}$. The Equation that defines the weight for class 0, in each task j could be seen in Equation 4.5 and the Equation that defines it for class 1, in 4.6.

$$\omega_{y_i=0,j} = \frac{N_{y_i=1,j}}{N_{y_i=0,j} + N_{y_i=1,j}} \quad (4.5)$$

$$\omega_{y_i=1,j} = \frac{N_{y_i=0,j}}{N_{y_i=0,j} + N_{y_i=1,j}} \quad (4.6)$$

With $l \in [0, 1]$, $N_{y_i=l,j}$ represents the total count of samples in the training data for task j , where the true class label equals l .

To minimize the L_{global} , the gradient of it is computed iteratively during the training process, over the learnable parameters, and updates the parameters through back-propagation. For the weights initialization was used in the low and high level task the Xavier uniform initializer [74]. This algorithm, also know as glort uniform initializer, draws samples from a uniform distribution within an interval $[-limit, limit]$, where limit is $\sqrt{6/(fan_{in} + fan_{out})}$, where fan_{in} is the

number of input units in the weight tensor, and fan_{out} is the number of output units in the weight tensor. The weights are updated using the Adam stochastic optimization algorithm [43], an extension to stochastic gradient descent. To estimate the model performance, the statistic model 4-fold cross validation was used, as done in [3].

4.4 Summary

The methodology employed in this dissertation is presented, composed of three modules: LIDC-IDRI analysis, dataset reassessment, and impact evaluation. For the first methodology module, were present the two methods used to do the analysis. They were a dataset tool review and a bibliographic research. The first was employed to have a better understanding of how the physicians evaluated the dataset semantic features. The second was used to gather specific information about each LIDC-IDRI semantic feature and its classes. From this method was concluded that sphericity and internal structure in LIDC-IDRI have no impact on the malignancy prediction, thus not being used for the dataset reassessment. Was also seen that in the literature was not found any objective definition in how to evaluate the different ratings of lobulation and spiculation, and so it will be indispensable to use the dataset review tool to reach to those definitions.

In the second module was presented the method used to reassess the LIDC-IDRI dataset using the resultant evaluated norms from the first step. In the last module, was explained the HSCNN architecture used to measure the impact of the reassessed dataset in the malignancy prediction, and the loss function used.

Chapter 5

Results and Discussion

This chapter presents conclusions from the research made in chapter 4, and the resultant evaluation norms, see section 5.1. In section 5.2 are presented the baselines for the reclassification, and held a discussion about the differences found between the LIDC-IDRI and the reassessed dataset, see section 5.2. In section 5.3 is measured the impact of using the reassessed dataset as the input of HSCNN. The dataset preparation and how the training was performed are explained. Results are shown and discussed.

5.1 Evaluation Norms

This section presents the results from the LIDC-IDRI analysis, with the tool and the bibliographic review. After the bibliographic research, were decided to review only the semantic features calcification, lobulation, margin, spiculation, subtlety, and texture. To a better dataset reassessment, the evaluations norms had to be defined in a consistent form. To that end, the way of classifying some characteristics had to change, according to the research made in section 4.1.2. This way was possible to more consistently reclassify the LIDC-IDRI dataset, after getting validation of the evaluation norms.

After a meeting with Hospital Sao João doctors and some doubts about the definitions were withdrawn, it was concluded that instead of 5 classes, as in LIDC-IDRI dataset, doctors tend to evaluate the characteristics or in 2 ("unmarked" or "marked"), or in 3 ("unmarked", "intermediate" or "marked") classes. One of the conclusions obtained in the reference work, HSCNN [3], was that the use of only two classes for the characteristics could have excluded information considered relevant for the malignancy results. Trying to not exclude much information in the coherence analysis, the evaluations were made with 3 classes, for each characteristic. The classes used were 1, 3, and 5, trying to represent this way the scale used in LIDC-IDRI dataset. It was attempted to order the classes, in all the characteristics, according to the relevance that each had for the malignancy. In this ordination, class 1 has the lowest correlation with malignancy, and class 5 has the most. Class 3 usually has a mean correlation between the two extremes.

Based on the knowledge acquired in the LIDC-IDRI analysis, it was possible to represent each characteristic in three different classes, obtaining definitions for each characteristic and each type of it. The conclusions obtained for each semantic feature, after the LIDC-IDRI analysis, were as follows:

Calcification's final definition was characterized as the calcium deposited in a nodule. In LIDC-IDRI dataset, calcification is divided in 6 classes, by calcification types.

After a review in LIDC-IDRI dataset, with the tool review, it was observed that this semantic feature classes were conceived differently from the others - distributed by each type of calcification, and not by a scale that was in agreement with the amount of characteristic's presence. With this lack of association, this class was prone to errors, often having doctors characterizing the nodules by the amount of calcification present rather than by type. An example of that was the nodule present in Figure 5.1, that has an average evaluation by doctors as central and following the definitions, is not central. The LIDC-IDRI classes distribution for this characteristic is also without an order of malignancy associated with the classes order.

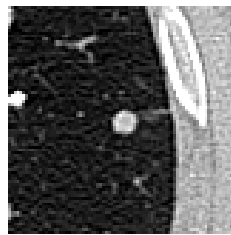


Figure 5.1: Nodule with a non-central calcification that has an LIDC-IDRI's doctors averages evaluation of central calcification from [4].

After research, see section 4.1.2, it was found that the classes 1, 2, 3, and 5 from LIDC-IDRI were correlated with benignity and class 4 correlated with malignancy. In addition to the non-central type being correlated with malignancy, in the literature was found another type related to it, known as "stippled". While non-central and stippled calcification correlates with malignancy, absent calcification is dependent on other characteristics to obtain an assertion about its correlation with malignancy. According to this, absent was defined as an intermediate class, class 3. Ordinating these classes with its malignancy correlation were obtained three classes as follows.

1. Popcorn, laminated, solid, central;
3. Absent;
5. Non-central, stippled.

As this characteristic presents visually substantial differences for each class, the best way to describe it would be to present image examples for each of the classes. The classes definitions could be seen in Figure B.1.

Lobulation's final definition was determined as the presence of lobules that come from the nodule, excrescences of the tumor as its leading edge. In LIDC-IDRI this characteristic is divided in 5 classes. After reviewing the worksheet file and done the research present in section 4.1.2, it

was concluded that class 1 is a nodule with "no lobulation", and 5 a nodule "lobulation marked". The ratings 2, 3, and 4 are intermediate values between these extremes, which have their definition subject to the opinion of each radiologist.

There is a uncertainty about how much lobulated margins relates with malignancy, see section 4.1.2.3. However, a lobulated margin has more malignancy suspicion than a smooth, well-marginated margin [65]. For this reason, the order of LIDC-IDRI classes was kept, being class 1 a no lobulated margin, and class 5 a lobulated margin.

After reviewing the worksheet file, it was noted that there are cases that are not explicitly lobulated or without lobulation, and that therefore it is not fair to classify them as an extreme. Thereby, it was decided to define a intermediate class - class 3 as a moderately lobulated margin.

Analysing the worksheet file, it was also notable that the main difference between evaluations were the quantity present of lobules in the nodules margin. Analysing all the nodules, and trying to divide them in 3 different cases, it was possible to reach to three definitions.

The following definitions were the basis for the reclassification of LIDC-IDRI nodules:

1. No lobulated - 0% of nodule's perimeter has lobules;
3. Moderately lobulated - between 25% to 50% of nodule's perimeter has lobules; number of grooves present in the nodule less or equal to 2;
5. Lobulated - more than 50% of nodule's perimeter with lobules; number of grooves present in the nodule greater than 2.

These classes definitions could be seen in Figure B.4.

Margin, was chosen to be described in the norm as the boundary line or the area immediately inside the boundary. In the LIDC-IDRI dataset it is classified from 1 to 5, with only the extremes defined by Opulencia *et al.* [15]: class 1 as "poorly defined", and 5 as "sharp". After reviewing this characteristic in section 4.1.2.4, it was concluded that the class that is more correlated with benignity is 5 (sharp, with well-defined margins), being the class more correlated with malignancy the 1 (poorly defined).

However, it was necessary to associate this characteristic with the malignancy scale, to synchronize it with the others. Therefore, to reorganize it, it was decided that class 1 would be defined as "sharp" and class 5 by "poorly defined", inverting the scale present in the LIDC-IDRI.

Analysing the worksheet file, it was verified that there were nodules that neither were well-defined nor poorly-defined. For them, an intermediate class was created, "defined". Thus, were established 3 classes defined as follows.

1. Sharp - margins well-defined in the nodule's entirety;
3. Defined - it is possible to see only some defined margins in the nodule - margins are no defined in the nodule's entirety;
5. Poorly defined - it is not possible to see definitions on nodule margins: it is possible at most to identify a nodule (e.g. a mass/substance having a higher density than the surrounding background).

These classes definitions could be seen in Figure B.7.

Spiculation's norm definition was characterized as a nodule that is covered with or having small needle-like structures (spicules) or fine fleshy points. In LIDC-IDRI the evaluation of this characteristic is divided into 5 categories, from 1 to 5. In the right order, based on Opulencia *et al.* [15] definitions, the extreme classes in LIDC-IDRI dataset, are 1 as "none" spiculation present, and 5 as "marked" spiculation present. Being the class 5, the class with more correlation with malignancy, and class 1 with less correlation, the order of the ratings remained the same, since it is synchronized with the malignancy scale. Thereby, in this redefinition, the class 1 was defined as a nodule with no spiculated margin and 5 as a nodule with a spiculated margin.

Visualizing the worksheet file, it was seen that there are lot of intermediate cases of spiculation, being neither very spiculated nor without any spiculation. To take advantage of the information that those nodules type could provide, was created an intermediate class, "moderately spiculated". Thus, three classes were defined. Another notable analysis in the file, was that the main difference between evaluations were the number of spicules and its radius. Examining the different nodules, and trying to combine these two variables with the 3 different classes, it was possible to supply a definition for each class. They were characterized as follows.

1. No spiculated - There is no spiculation present in the nodule margins;
3. Moderately spiculated - Low spiculation present in the margins - only 1 or 2 spicules with at least one spicule radius greater or equal to the nodule's radius, or more than 2 subtle spicules all with a smaller radius than the nodule's radius;
5. Spiculated - Extremely spiculated margins - with various (more than 2) spicules with at least one spicule radius larger or equal to the nodule's radius.

These definitions could also be seen in Figure B.10.

Subtlety, was described in the norm as the difficulty of detection of the nodule relatively to the surrounding. In LIDC-IDRI was divided in 5 classes, see section 4.1.2. After reviewing the worksheet file was seen that a division in 3 classes was achievable. Thus, maintaining the classes 1, 3, and 5, as the same defined by Opulencia *et al.* [15] the division was made as follows:

1. Extremely subtle - Little visible difference between structures and surroundings of the lung and nodule;
3. Fairly subtle - Visible difference between structures and surroundings of the lung and nodule. Even if the nodule is subtle, it is visible;
5. Obvious - Obvious difference between structures and surroundings of the lung and nodule. Obvious and easy visibility of the nodule.

These norm definitions could be seen in the Figure B.13.

Lastly, **texture's** definition in the norm was the internal density of the nodule in terms of the ratio of solid present. In LIDC-IDRI the ratings to this characteristic were from 1 to 5, with a clear definition by Opulencia *et al.* [15] in the classes 1, 3, and 5, see Table 4.1.

As observed in section 4.1.2, the distribution of the texture's LIDC-IDRI ratings is completely reversed, according to the correlation that it has with malignancy. In the texture's case, the part-solid nodules are the ones that have most correlate with the malignancy, being after the pure ground-glass nodules and finally the solid nodules. It will become a less intuitive classification because it goes into a disagreement with the way the characteristic develops. However, this way the relationship with malignancy is more coherent and facilitates the use of this characteristic in the HSCNN network for the malignancy classification.

Thus, synchronizing this feature with the malignancy scale, the classes definitions were:

1. Solid - Nodule composed with a greater proportion of solid than non-solid components;
3. Non-solid - Low visible nodule with low attenuation, with a greater proportion of non-solid than solid components;
5. Part-solid - Nodule with a higher percentage density located centrally or peripherally, with an equal proportion of solid and non-solid components.

These definitions could also be seen in Figure B.16.

The classes division obtained could be seen in Table 5.1.

Table 5.1: Ratings division obtained from the analysis.

Characteristic	Classes	Designations
Calcification	1	Popcorn, laminated, solid, central
	3	Absent
	5	Non-central, stippled
Lobulation	1	No lobulated
	3	Moderately lobulated
	5	Lobulated
Margin	1	Sharp
	3	Defined
	5	Poorly defined
Spiculation	1	No spiculated
	3	Moderately spiculated
	5	Spiculated
Subtlety	1	Subtle
	3	Fairly subtle
	5	Obvious
Texture	1	Solid
	3	Non-solid
	5	Part-solid

After reach to more standardized definitions for each semantic feature and each of this classes, the evaluation norms were created. Along with the definitions, there were provide some image examples for each class, to help to relate real nodules with the definitions given. The image examples provided were from a very small nodule and two different bigger nodules (with the difference in size - one small and another larger), with the images in the three different anatomical

planes. Along with these images were provide also with the very small nodule images with the average binary mask from doctors to help to see, in doubt, where the abnormality is. These images were obtained in the same way as explained in section 4.1.1. The evaluation norms could be seen in Appendix B.

Validation Created these norms, was held a meeting with Hospital São João doctors to validate them. They validated the norms as they were, but made a suggestion for a better future application of these for lobulation and spiculation characteristics. They considered that its classes definitions, as they were, could, in the future, restrict doctors' assessment too much, and that a more ambiguous description would be better. For this reason, they suggested the characteristics definitions as follows. For lobulation:

1. No lobulated - 0% to of lobulation present;
3. Moderately lobulated - 25 to 50% of lobulation present;
5. Lobulated - more than 50% of lobulation present.

For spiculation the definitions recommended were as follows.

1. No spiculated - 0% to of spiculation present;
3. Moderately spiculated - 25 to 50% of spiculation present;
5. Spiculated - more than 50% of spiculation present.

However, these definitions were too ambiguous to implement an objective reclassification. The only reason provided by doctors to suggest norms different from the initially conceived, was that these were too objective. Since the purpose of the created norms is to make a classification more objective, the definitions presented before were maintained for the dataset reassessment.

Having the evaluation norms validated by physicians, a dataset reassessment using them was possible.

5.2 Dataset Reassessment

Based on the procedure explained in section 4.2, the dataset was reclassified with the evaluation norms in a more standardized form. The reassessment was done on the same worksheet file of the tool review, made in section 4.1.1, analysing 2.5D images since the 3D images analysis is prone to fatigue, and these give a good perspective of the anatomical planes.

All the abnormalities were analysed and used for reclassification, even if in the evaluations some have been identified by some physicians as non-nodules. In LIDC-IDRI dataset evaluations, if a radiologist identifies a non-nodule, he does not evaluate its semantic features. However, sometimes non-nodules could have noticeable features that could help to explain to the network what each characteristic is. To try to collect all the possible information, the entire dataset was reclassified.

For the reclassification 2,660 of the 2,668 nodules of the dataset were used. The remaining 8 nodules were not used for reclassification because they were not visible in the 2.5D images

acquired for the worksheet file. The reclassification was made of characteristic in characteristic, for each one of the 2,660 nodules, being made a total of 15,960 evaluations. The evaluations were after converted to a .csv file, for a future ease use as input in the network, for the impact evaluation.

Results and Discussion After the reassessment, was made a comparison between the total number of nodules in each class, in each characteristics, of the original and the reclassified datasets.

For the original dataset, to count the nodules present in each class, it was taken into account the average of the doctors' evaluations. To make a comparison among those values was done a conversion between the original LIDC-IDRI classes and the evaluation norms classes. This translation is not completely linear, employed only to have a perception of the differences between the two datasets. However, to do this comparison had to be made some assumptions.

For lobulation and spiculation, when a nodule was evaluated in the LIDC-IDRI dataset with class 2, thus having some sign of spiculation or lobulation in its margin, it was considered to be counted as an intermediate lobulation/spiculation, class 3 in the evaluation norms. When a nodule in the LIDC-IDRI dataset was classified with 4, as it has a strong sign of spiculation/lobulation in the nodule margins, it was counted as class 5 in the evaluation norm. For subtlety, as the difference between classes 1 and 2 and between 3 and 4 in the original dataset is not as relevant to malignancy, comparing with those of lobulation/spiculation, were converted respectively, into classes 1 and 3 of the evaluation norm. To margin, the equivalent happened, converting the ratings regarding the evaluation norm. Calcification not needed assumptions, since the translation was linear, according to the types of calcification. In texture, the major part of the nodules evaluated with 2, 3, and 4 in LIDC-IDRI dataset were part-solids thus being converted to the class correspondent in the evaluation norm, class 5. The conversion and numbers of nodules for each class, for each characteristic could be seen in Table 5.2.

With this information, it was possible to note the differences between the original and reassessed dataset. In calcification, the main difference seen between the original and the reclassified dataset, was the number of nodules with non-central calcification (class 4 in the LIDC-IDRI, and class 5 in the reassessed). In the original dataset, most nodules with an average of evaluations of 4 (non-central class) were incorrectly classified. These nodules had a very high disparity in physicians' ratings, caused by a lack of class definitions or by error, when physicians incorrectly classified according to the amount of calcium presence, instead of calcium type. It often resulted in average evaluations of 4, even if it was not the correct assessment. With the reassessment, it was possible to evaluate these nodules correctly which resulted in more nodules classified in the original ratings 1,2,3,5, and 6. Although stippled was used to define the class 5 in calcification norm, was not found any nodule in LIDC-IDRI dataset with this type of calcification. This way, all the nodules evaluated with class 5 have non-central calcification.

In lobulation, was observed a high disparity between physician evaluations in the original dataset. Moreover, it was visually confirmed, observing LIDC-IDRI dataset, that many nodules evaluated with class 1 (non-lobulated margins) had instead small grooves in the margins. It was concluded that when physicians assessed these type of nodules, they choose to evaluate the nodule

Table 5.2: Comparison between the nodules number in the original and reclassified datasets

Characteristic	LIDC-IDRI Ratings	Norm Ratings	Nodules in LIDC-IDRI	Nodules in Norm
Calcification	1,2,3,5	1	296	331
	6	3	2282	2311
	4	5	82	18
Lobulation	1	1	1452	853
	2,3	3	1082	1603
	4,5	5	126	204
Margin	5	1	1057	1315
	3,4	3	1271	1123
	1,2	5	332	222
Spiculation	1	1	1665	1656
	2,3	3	848	852
	4,5	5	147	152
Subtlety	1,2	1	374	310
	3,4	3	1506	1104
	5	5	780	1246
Texture	5	1	1886	1452
	1	3	175	214
	2,3,4	5	599	994

as having no lobules, instead of evaluating with class 2 or 3. That could result in a loss of valuable information. By reassessing the dataset, it was decided to evaluate those nodules' types with the intermediate class, class 3. So, comparing the reassessed to the original dataset, the number of those classified as 3 increased, and the number of those evaluated as 1 decreased.

In margin was not observed much disparity between the physicians' evaluations. It was noticed a not very significant increase in the number of nodules, in the "sharp" class (class 5 in the original and class 1 in the reassessed dataset). Analysing the worksheet file, it was seen that some nodules that had an average rating of 3 and 4 in the original dataset, had, however, sharp margins, fitting more into the class sharp. The equivalent happened to some nodules that had an average classification of 1 and 2 in the original dataset, but had defined margins, fitting more into the class 3, and 4.

For spiculation, was observed the same phenomenon that occurred in lobulation. When seeing a minimal sign of spicules in the nodule margins, the physicians tended to evaluate it with 1, without spicules, instead of evaluating with 2 or 3. Since it can result in a loss of valuable information when the reassessment was made these cases were assessed with the intermediate class of the evaluation norm, class 3. Although this was done, little difference was noticed between the numbers of the original and reclassified dataset. Analysing the evaluations given to spiculation in LIDC-IDRI was concluded that in some abnormalities, despite the individual evaluations were not accurate, its average was. An example of it would be, a moderately spiculated nodule with an average classification of 3 (so, correct), but when analysing the individual evaluations, they were 5, 1, 1, and 5. Even though it had not been rated with 3 in the individual evaluations, it achieved

the right classification in the average evaluations.

In subtlety, the relevant differences between the two datasets were seen in classes 3, 4, and 5, from the original dataset. There was seen a disparity, even if small, within the physicians' evaluations that caused many nodules that were obvious to visualize, class 5, to had an average evaluation of 3, or 4. In the reassessment, it was possible to solve that, being all the nodules that were absolutely obvious to visualize, evaluated with 5. That increased the nodules' number of this class.

For texture the main differences were between the classes solid and part-solid. In the reclassification, there was an increase in the number of nodules with a texture part-solid. That happened because, when reclassifying, were taken into account for the class solid, only nodules that had a texture completely solid. Thus, some nodules originally classified as 5 were assigned to classes 2, 3, and 4 of the original dataset.

Concluding, most of the errors present in LIDC-IDRI original classification, happened because of the disparity in the radiologists' nodules evaluations, that caused many times a wrong average evaluation result, excluding in spiculation. Most of the differences between the nodule numbers in the two datasets are due to a readjustment between consecutive classes. In LIDC-IDRI this feature is evaluated differently from the others, by type and not by the amount of calcification. With it, many physicians have misclassified it, annotating it according to quantity. With that, some nodules had an average of 4, not-central calcification, when they were not.

Having a more standardized classification in the LIDC-IDRI dataset, regarding all the relevant characteristics, it was then possible to move to the next phase, the implementation of HSCNN having this new dataset as input.

5.3 Impact Evaluation

In this section are presented the results from HSCNN prediction with the reassessed dataset. In section 5.3.1 is described the data processment done to use the reassessed dataset in HSCNN. In section 5.3.3 are shown the results, and in section 5.3.4 these are discussed.

5.3.1 Dataset Preparation

For the data preparation first was made a ternary to binary conversion. Second, the division for 4-fold cross validation was accomplished. Finally, a data preprocessment was made.

Binary Conversion The data used as input for the HSCNN in [15] was binary. LIDC-IDRI ratings are between 1 and 5, or between 1 and 6 in calcification. To convert these evaluations in binary, they used a conversion method that could be seen in Table 3.3. In this translation, all the nodule semantic features that were classified with 0 were correlated with benignity, as the ones classified with 1 were correlated with malignancy. As Shen *et al.* used the HSCNN with binary

labels, to make a comparison with their results, to know if the new dataset has an impact on the malignancy prediction, it was necessary to convert the ternary reassessed dataset to binary.

To convert the classes from ternary to binary in the new dataset, was done a conversion based on [3, 75]. The main difference was the binary conversion done to the intermediate class. While in the conversion of Shen *et al.* all the nodules classified with 3 were converted to the binary class 0, in this work they were converted to the binary class 1. Following the logic used in the Shen *et al.* conversion, this way the intermediate class was considered malignant, in the binary classification. This was considered preventive, avoiding this way false negatives diagnosis. All the nodules were converted to the respective class in binary, following the Table 5.3. The nodules' total number for each class in each characteristic could be seen in Table 5.5.

Table 5.3: Characteristics conversion from ternary to binary classification.

	Malignancy	
	0	1
Calcification	1	3,5
Lobulation	1	3,5
Margin	1	3,5
Sphericity	1	3,5
Spiculation	1	3,5
Subtlety	1	3,5
Texture	1	3,5

The designations for each semantic feature binary rating could be seen in Table 5.4.

Table 5.4: Ratings division obtained from the binarisation.

Characteristic	Classes	Designations
Calcification	0	Popcorn, laminated, solid, central
	1	Absent, Non-central, stippled
Lobulation	0	No lobulated
	1	Moderately lobulated, Lobulated
Margin	0	Sharp
	1	Defined, Poorly defined
Spiculation	0	No spiculated
	1	Moderately spiculated, Spiculated
Subtlety	0	Subtle
	1	Fairly subtle, Obvious
Texture	0	Solid
	1	Non-solid, Part-solid

Malignancy in LIDC-IDRI is evaluated between ratings 1 and 5. Opulencia *et al.* proposed terms for each rating, where class 1 is highly unlikely, 2 is moderately unlikely, 3 is indeterminate, 4 moderately suspicious, and 5 highly suspicious malignancy. These evaluations are only suspicion levels of physicians, like the semantic features evaluations. The actual diagnostic data in

Table 5.5: Total number of nodules for each characteristic

	Classes	
	0	1
Calcification	331	2329
Lobulation	853	1807
Margin	1315	1345
Spiculation	1656	1004
Subtlety	310	2350
Texture	1452	1208

LIDC-IDRI only exists for 96 patients. As so, the evaluations used were the likelihood of malignancy. To use malignancy as one of the HSCNN inputs, it was needed to make a conversion from the malignancy LIDC-IDRI assessments into binary. For this were used the malignancy LIDC-IDRI average evaluations given by physicians. To transform the malignancy ratings into binary, Shen *et al.* converted the ratings 1,2,3 to class 0, and 4, 5 to class 1, see Table 3.3. Class 0 is related to a benign nodule, and class 1 to a malign. In this work, the transformation was made by converting 1, and 2 to class 0, and 3,4, and 5 to class 1. The difference between the Shen *et al.* conversion and this work, was the relevance given to the class 3 for the malignancy. This was made to avoid false-negative cases in malignancy diagnosis. This transformation to binary gave a total of 899 benign nodules (class 0), and 1,761 malign nodules (class 1). These ratings were all converted in the worksheet file, being converted after to a .csv file.

Dataset division for 4-fold cross validation To obtain the final assessment of the model performance was used a 4-fold cross validation. In each fold, there were two subsets with the train set, one subset with the validation set, to tune the hyperparameters, and another for holdout testing, that reported the final model performance. In each fold each subset used for training, validation and test, changed, to make sure the test set was independent from model training and parameter optimizations. Another advantage of changing the test set in each fold was that, unlike other studies that not holdout the test set [10, 75], there was no information leakage and in such way there was not a model performance over-estimation.

Initially, to implement 4 fold cross-validation on the model, it was necessary to split the input dataset. The input dataset utilized in the architecture consisted of LIDC-IDRI 3D images (CTs) with dimension of 80x80x80 voxels, and binary labels for each image. All the image information and respective labels were load. The data were shuffled randomly, with a specific seed. The images were split into four groups, as happened with the labels. Initially, the folds were created with the information of the 2,660 abnormalities. After some training, and validation it was concluded that as Shen *et al.* [3] made, the HSCNN achieve better results when picking only the abnormalities that had in LIDC-IDRI original dataset, more or equal to 3 evaluations. This is because when there is no evaluation, the doctor do not consider the abnormality as a nodule, being sometimes so little,

that possibly the network does not learn anything from it. As so, the process was made for a total of 1,386 nodules, representing the first created fold 346, the second 347, the third 346, and the fourth 347 nodules.

The total nodules number in each class, for each semantic feature and malignancy, could be seen in Table 5.6.

Table 5.6: Total number of nodules for each characteristic and malignancy, after selecting only the abnormalities that had 3 or more evaluations in LIDC-IDRI

	Classes	
	0	1
Calcification	190	1196
Lobulation	308	1078
Margin	776	610
Spiculation	751	635
Subtlety	58	1328
Texture	960	426
Malignancy	335	1051

Data Preprocessing Before using the split data as input of the HSCNN, it was necessary to preprocess it. For each of the four image group, the pixel values were normalized, converting them from (-1000, 500) HU, Housfield scale, to a range from (0, 1). After, each one of the 3D images present in each image group, were cut in the borders, reducing each image to a dimension of 52x52x52 voxels. This size was the chosen because, in this shape, even the largest nodules are entirely contained in the cube, thus not losing relevant information [3]. It was also the dimension used in the work of Shen *et al.*.

5.3.2 Training

After this data preprocessing, the 4-fold cross validation began. A procedure was made to vary in each fold the four subsets. See the disposition in Figure 5.2. In this way, the subsets assigned for training, validation, and testing were not repeated at any fold. In each fold, when knowing which subsets were designated for training, the images and labels from the two chosen training subsets were combined, creating this way one unique subset with the train images, and train labels. For the chosen subsets for validation and test, the images and respective labels were also put together.

Models were trained for 300 epochs during each fold of cross-validation. Was used a batch size of 32. The learning rate that gave the best model results was 0.00005. Were used the regularization methods dropout with a value of 0.8, and L2 with a value of 0.08.

The weights for each task, present in the Equation 4.3, were chosen first by the relevance that each feature had to malignancy prediction. After, by trial end error it was seen that it achieved better results with the weights for texture, spiculation, subtlety, lobulation, margin, calcification and malignancy as 0.1, 0.2, 0.2, 0.2, 0.1, 0.2, 0.6, respectively. This weights were applied to

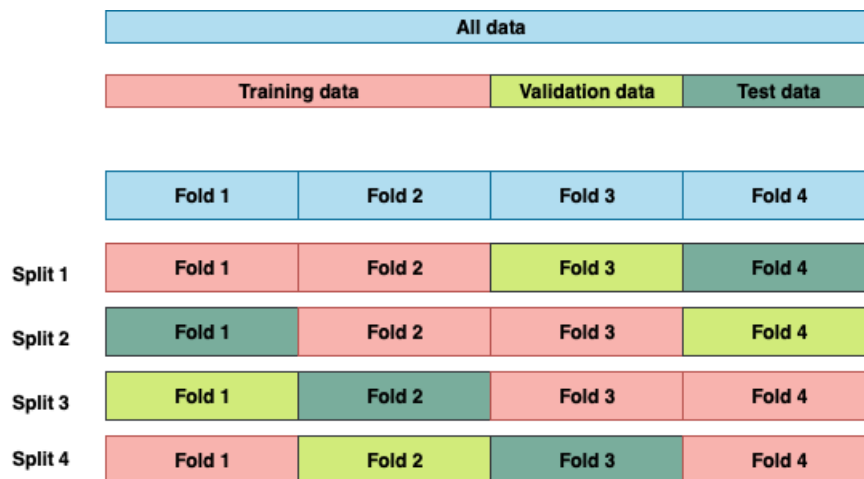


Figure 5.2: Implemented 4 fold cross-validation.

the loss function. To choose the best model for each fold, it was used the checkpoint function, being the best model the one that achieved the lowest validation malignancy loss. This tuning to the hyperparameters was performed with a fit generator function, where data augmentation was executed on the training and validation dataset. The augment consisted of artificially add samples, generated by randomly picking one or multiple operations such as rotations of 45 degrees, width, or/and height shifts of 1.82mm (0.05% each of the total volume), or random vertical, or/and horizontal flips. The dataset augmentation was used as a form of regularization, enabling the mode to generalize better, see section 3.1.3.1.

Only the independent test dataset was used to calculate end model performance. This network was implemented in Python 2.7, with Tensorflow and the Keras toolkit. All experiments were performed using NVIDIA TESLA P100, with a GPU of 16GB, and a 32GB RAM.

5.3.3 Results

To measure the results, the metrics used were AUC, specificity, sensitivity and accuracy. These were used to make a comparison with the work of Shen *et al.*, to see if with a more coherently assessed dataset were obtained better results in the HSCNN predictions. Although accuracy does not measure well the performance of the network, since this is an imbalanced dataset, it was used, only with the purpose to compare its result with of the Shen *et al.*. After implementing the new dataset as input in the HSCNN, were obtained the metrics.

The results for the semantic features used in the HSCNN could be seen in Table 5.7. These results were obtained by obtaining the average of all 4 fold cross-validation results.

Table 5.7: Average test results for each semantic feature.

Characteristic	Accuracy	AUC	Sensitivity	Specificity
Calcification	0.71	0.64	0.75	0.50
Lobulation	0.54	0.71	0.45	0.84
Margin	0.58	0.67	0.72	0.46
Spiculation	0.69	0.76	0.59	0.77
Subtlety	0.58	0.85	0.57	0.79
Texture	0.58	0.72	0.75	0.50

The best results obtained in the test set for the semantic features, could be seen in Table 5.8. These results were obtained using fold 2 as test set.

Table 5.8: Best test results for each semantic feature, using fold 2.

Characteristic	Accuracy	AUC	Sensitivity	Specificity
Calcification	0.73	0.66	0.84	0.54
Lobulation	0.50	0.69	0.38	0.89
Margin	0.60	0.62	0.63	0.57
Spiculation	0.69	0.74	0.75	0.63
Subtlety	0.76	0.80	0.76	0.45
Texture	0.62	0.69	0.74	0.57

The results for malignancy prediction for each fold used as test set, could be seen in Table 5.9.

Table 5.9: Test results for each fold, for lung nodule malignancy prediction.

Fold	Accuracy	AUC	Sensitivity	Specificity
Fold 1	0.68	0.67	0.74	0.46
Fold 2	0.78	0.74	0.83	0.54
Fold 3	0.70	0.66	0.80	0.43
Fold 4	0.65	0.66	0.71	0.48

The confusion matrix and ROC curve of each semantic feature and malignancy were also analysed. For this analysis and further discussion, were chosen the ROC curves and confusion matrices of the test results using fold 2. It was chosen this since was the fold that achieved overall best results.

The confusion matrices values are disposed in the same form as Table 3.1. They were made for a total of 347 different images as it is the total number of data in fold 2. To see the low-level tasks confusion matrices see Figure 5.3. For the high-level task confusion matrix see Figure 5.10.

Calcification		Prevision		Lobulation		Prevision		Margin		Prevision	
		0	1			0	1			0	1
Disease	0	22	19	Disease	0	74	9	Disease	0	74	9
	1	74	232		1	163	101		1	163	101

(a) Calcification

Spiculation		Prevision		Subtlety		Prevision		Texture		Prevision	
		0	1			0	1			0	1
Disease	0	111	65	Disease	0	5	6	Disease	0	142	108
	1	43	128		1	79	257		1	25	72

(d) Spiculation

(e) Subtlety

(f) Texture

Figure 5.3: Confusion matrices for each semantic feature, using fold 2 as test set.

Table 5.10: Confusion matrix for malignancy, using fold 2 as test set

Malignancy		Prevision	
		0	1
Disease	0	37	32
	1	46	232

To analyse the resulting ROC curves were used the results of the same fold, fold 2, as test set. The ROC curves of low-level tasks are present in Figure 5.4, and of high-level task, see Figure 5.5.

Examining the ROC curves and confusions matrices using fold 2 as test set, the relevant results for each semantic feature and malignancy were collected.

Analysing the calcification ROC curve see 5.4a, and knowing its AUC is 0.66, was concluded that this low-level task predicted 66% correctly the test. From its confusion matrix, see Figure 5.3a, was observed that it obtained a sensitivity of 0.76, a specificity of 0.54, and an accuracy of 0.73.

Lobulation had an AUC of ROC curve of 0.69, meaning that the task predicted correctly 69% of the test set, see Figure 5.4b. Analysing its confusion matrix, see Figure 5.3b, this low-level task achieved a sensitivity of 0.38, a specificity of 0.89, and an accuracy of 0.5.

For margin, the AUC of its ROC curve, see Figure 5.4c, was of 0.62, meaning that the task predicted 62% correctly the test set. Observing its confusion matrix, see Figure 5.3c, was seen that achieved a sensitivity of 0.63, and a specificity of 0.57. The accuracy obtained was of 0.6.

For spiculation, analysing the ROC curve and its AUC, see Figure 5.4d, was concluded that this classifier predicted 74% of the test set correctly. From the confusion matrix, see Figure 5.3d, was observed that obtained a sensitivity of 0.75, a specificity of 0.63, and an accuracy of 0.69.

Subtlety presented a ROC curve with an AUC of 0.80 see Figure 5.4e, predicting 80% of the test set correctly. Analysing its confusion matrix, see Figure 5.3e, was concluded that this low-level task obtained a sensitivity of 0.76, a specificity of 0.45, and an accuracy of 0.76. The texture had an AUC value for the ROC curve of 0.69, presented in Figure 5.4f, which means that

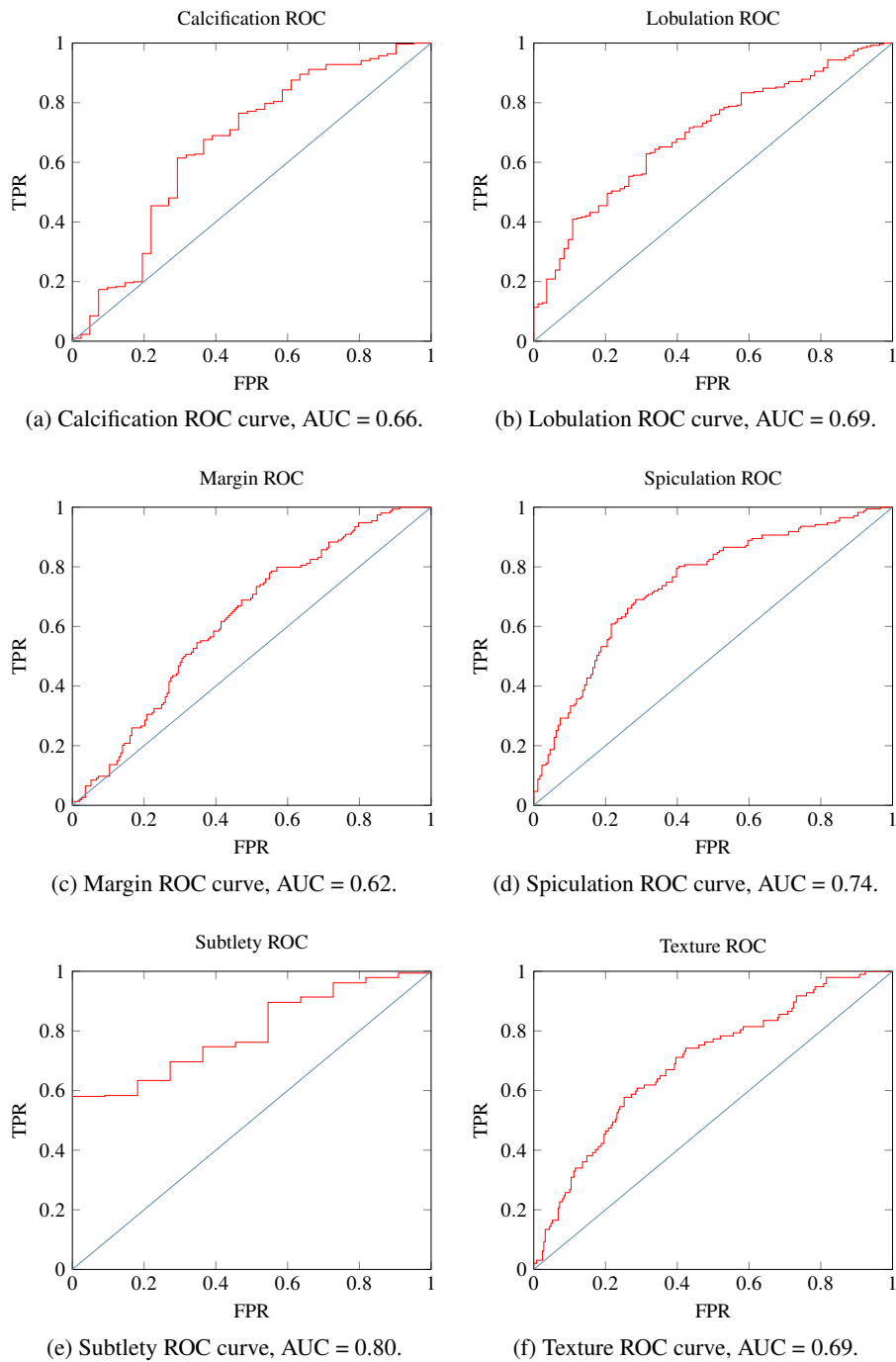


Figure 5.4: ROC curve for each semantic feature, using fold 2 as test set.

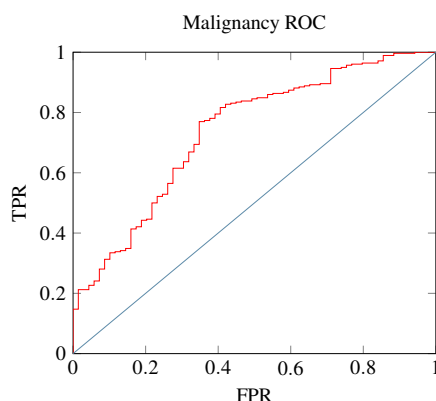


Figure 5.5: Malignancy ROC curve for fold 2, with an AUC = 0.74.

the texture low-level task predicted 69% of the test set correctly. Seeing its confusion matrix, see Table 5.3f, was concluded that its sensitivity was 0.74, and its specificity 0.57.

5.3.4 Discussion

After relating the results in section 5.3.3, in this section these were analysed and discussed.

Observing the confusion matrix for calcification, see Figure 5.3a, was concluded that this low-level task was better to predict when the calcification in a nodule was absent or non-central, predicting correctly 76% of those, than when the calcification was benign, predicting correctly 54% of it. Analysing the confusion matrix of lobulation, see Figure 5.3b, was concluded that, when were presented nodules with no lobulation to the task, it predicted correctly in 89% of those, predicting incorrectly 11%. When lobulated nodules were presented to the network, it predicted correctly 38% of these. By observing the confusion matrix of margin, see Figure 5.3c, 57% of the sharp nodules were correctly identified, being 43% incorrectly predicted as poorly defined. It was also concluded that 63% of the nodules that were ill-defined were correctly predicted. For spiculation, the confusion matrix results, see Figure 5.3d, shown that when presented nodules with spiculated margins the network predicted 75% of them correctly. When presented nodules with no spiculation in the margins to the network, it predicted well 63% of them correctly. In the subtlety confusion matrix, see Figure 5.3e, is shown that when presented obvious nodules to the task, it predicted 76% of them correctly. However, when presented nodules that were subtle, it predicted 45% of them correctly, 5 in a total of 11 subtle nodules from this test set. In texture, by observing its confusion matrix, see Figure 5.3f, it was concluded that when nodules non-solids/part-solids were presented to the network, were predicted as such in 74% of it, and 26% wrongly, as solids. From the results it was also observed that the network predicted correctly the nodules as solid in 57% of the solid nodules present in the test set, being 43% predicted incorrectly.

Malignancy obtained an AUC of ROC curve of 74%, predicting correctly 74% of the test set, see Figure 5.5. Had a sensitivity of 84% being that, the percentage of malignant nodules that were correctly predicted as such. However, when predicting correctly benignity this classifier had worst performance, predicting correctly 54% of the benign nodules.

An analysis to all the nodules from this test set was made, to understand which could be the causes of the wrong predictions. From this analysis was concluded that all the tasks had difficulty to predict correctly nodules that were small, and had a lot of matter around. All the tasks, excepting margin had trouble in predicting juxta-pleural nodules. However, margin had trouble with irregular margins, along with lobulation and spiculation tasks. Texture, spiculation and lobulation had difficulty to predict correctly nodules with solid calcification. Another conclusion obtained from observing the wrongly predicted nodules from the tasks lobulation, spiculation, margin, and subtlety was that the way as the dataset binarisation was performed, may have misled the tasks. That happened because, most of the false-negative cases resultant of those tasks, were, before the binarisation, from the intermediate class in each respective norm. When applied the binarisation, these and the nodules from one of the extreme classes were joined, in one class. Judging by the results, this merge contributed to a rather heterogeneous class, which could have mistaken the network. The nodules types that were more times correctly predicted were solitary or with pleura-tail.

Examples of nodules and their truth and predicted outputs from the fold 2, could be seen as follows, in Figure 5.6. These nodules examples were chosen as of some with the best and worst trade off between prediction and truth labels.

Figure 5.6: Nodules images with their predictions and truth labels.

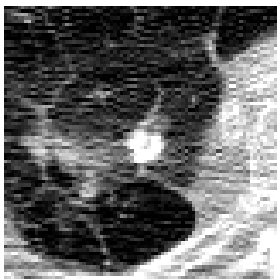
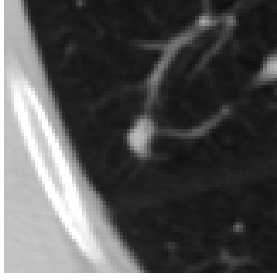
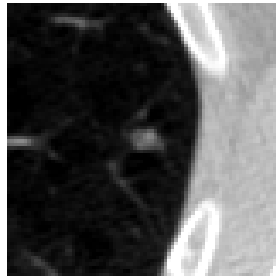
(a) Nodule in axial plane.	(b) Results of true and predicted outputs.																								
	<table border="1"> <thead> <tr> <th>Semantic feature</th> <th>True</th> <th>Predicted</th> </tr> </thead> <tbody> <tr> <td>Calcification</td> <td>0</td> <td>1</td> </tr> <tr> <td>Lobulation</td> <td>1</td> <td>1</td> </tr> <tr> <td>Margin</td> <td>0</td> <td>1</td> </tr> <tr> <td>Spiculation</td> <td>1</td> <td>1</td> </tr> <tr> <td>Subtlety</td> <td>1</td> <td>0</td> </tr> <tr> <td>Texture</td> <td>1</td> <td>1</td> </tr> <tr> <td>Malignancy</td> <td>1</td> <td>1</td> </tr> </tbody> </table>	Semantic feature	True	Predicted	Calcification	0	1	Lobulation	1	1	Margin	0	1	Spiculation	1	1	Subtlety	1	0	Texture	1	1	Malignancy	1	1
Semantic feature	True	Predicted																							
Calcification	0	1																							
Lobulation	1	1																							
Margin	0	1																							
Spiculation	1	1																							
Subtlety	1	0																							
Texture	1	1																							
Malignancy	1	1																							
(c) Nodule in axial plane.	(d) Results of true and predicted outputs.																								
	<table border="1"> <thead> <tr> <th>Semantic feature</th> <th>True</th> <th>Predicted</th> </tr> </thead> <tbody> <tr> <td>Calcification</td> <td>1</td> <td>0</td> </tr> <tr> <td>Lobulation</td> <td>1</td> <td>0</td> </tr> <tr> <td>Margin</td> <td>0</td> <td>0</td> </tr> <tr> <td>Spiculation</td> <td>1</td> <td>0</td> </tr> <tr> <td>Subtlety</td> <td>1</td> <td>1</td> </tr> <tr> <td>Texture</td> <td>0</td> <td>1</td> </tr> <tr> <td>Malignancy</td> <td>1</td> <td>1</td> </tr> </tbody> </table>	Semantic feature	True	Predicted	Calcification	1	0	Lobulation	1	0	Margin	0	0	Spiculation	1	0	Subtlety	1	1	Texture	0	1	Malignancy	1	1
Semantic feature	True	Predicted																							
Calcification	1	0																							
Lobulation	1	0																							
Margin	0	0																							
Spiculation	1	0																							
Subtlety	1	1																							
Texture	0	1																							
Malignancy	1	1																							

Figure 5.6: Nodules images with their predictions and truth labels.

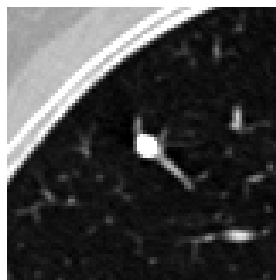
(e) Nodule in axial plane.



(f) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	0	0
Margin	1	0
Spiculation	0	0
Subtlety	1	1
Texture	1	0
Malignancy	1	0

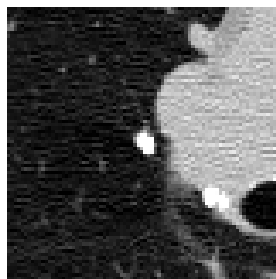
(g) Nodule in axial plane.



(h) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	0	1
Lobulation	0	1
Margin	0	0
Spiculation	1	1
Subtlety	1	1
Texture	0	0
Malignancy	0	0

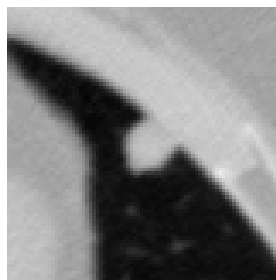
(i) Nodule in axial plane.



(j) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	0	0
Lobulation	1	0
Margin	0	1
Spiculation	0	1
Subtlety	1	1
Texture	0	1
Malignancy	0	0

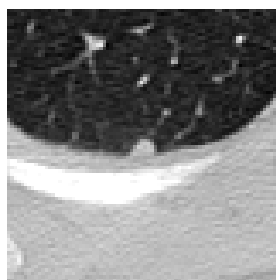
(k) Nodule in axial plane.



(l) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	0
Lobulation	1	0
Margin	0	1
Spiculation	0	1
Subtlety	1	0
Texture	0	1
Malignancy	1	1

(m) Nodule in axial plane.

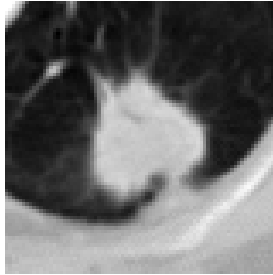


(n) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	0	0
Margin	0	0
Spiculation	0	1
Subtlety	1	1
Texture	0	0
Malignancy	0	1

Figure 5.6: Nodules images with their predictions and truth labels.

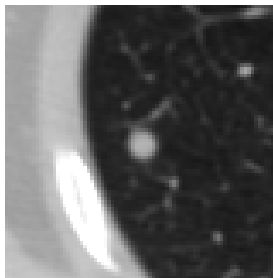
(o) Nodule in axial plane.



(p) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	1	1
Margin	0	0
Spiculation	1	1
Subtlety	1	1
Texture	0	0
Malignancy	1	1

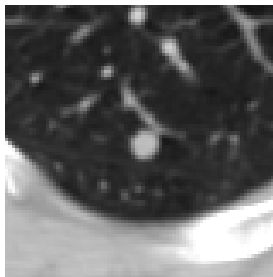
(q) Nodule in axial plane.



(r) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	0	0
Margin	0	0
Spiculation	0	0
Subtlety	1	1
Texture	0	0
Malignancy	1	1

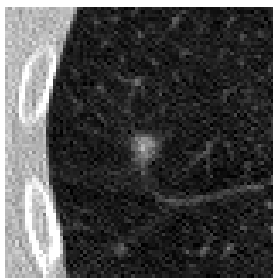
(s) Nodule in axial plane.



(t) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	0	0
Margin	0	1
Spiculation	0	1
Subtlety	1	0
Texture	0	1
Malignancy	1	1

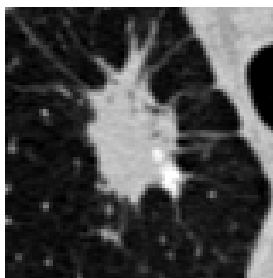
(u) Nodule in axial plane.



(v) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	0	0
Margin	1	1
Spiculation	0	0
Subtlety	1	0
Texture	1	1
Malignancy	1	1

(w) Nodule in axial plane.



(x) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	1	1
Margin	0	1
Spiculation	1	1
Subtlety	1	1
Texture	0	0
Malignancy	1	1

Figure 5.6: Nodules images with their predictions and truth labels.

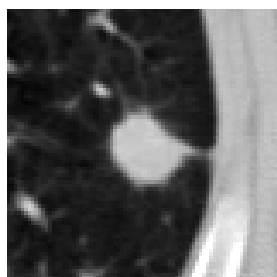
(y) Nodule in axial plane.



(z) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	1	1
Margin	0	0
Spiculation	1	1
Subtlety	1	1
Texture	0	0
Malignancy	1	1

(o) Nodule in axial plane.



(o) Results of true and predicted outputs.

Semantic feature	True	Predicted
Calcification	1	1
Lobulation	1	1
Margin	0	0
Spiculation	1	1
Subtlety	1	1
Texture	0	0
Malignancy	1	1

In the work of Shen *et al.* were obtained as best results for malignancy an accuracy of 0.845, an AUC of ROC of 0.856, a sensitivity of 0.705 and a specificity of 0.889. Comparing the best results of this dissertation with those, our dataset only improved the sensitivity, obtaining 0.83, meaning that, that improved the prediction of positives in malignancy. Semantic features result were not improved, also. Comparing with the works reported in chapter 3, this work only presented better results than MC-CNN [10] in the sensitivity. The results for subtlety and margin compared to MC-CNN results were also not improved. Although, in their work they evaluated the results using only training and validation subsets, not having an independent holdout test set which could overestimate the model performance.

Some hypotheses are given to explain the differences between the results of Shen *et al.* work. First, in this work were only used 1,386 images, against the 4,252 used in the Shen *et al.* work. Although the reassessed dataset had more consistent labels, reducing the dataset for only the abnormalities evaluated by 3 or more doctors, made it smaller, that could possibly limited the HSCNN learning. The second hypothesis is the high imbalance present in this dataset classes. Although in the loss equation weights are implemented to balance this problem (see Equations 4.4, 4.5, 4.6), the worst task results are a reflection of when the classes imbalance is higher. A third hypothesis is that converting the intermediate class to the binary class 1 could have mistaken the network, as it created a heterogeneous class. A plan to solve it is to transform ternary to binary in a different way, converting the classes 0, and 1 to the binary class 0 and the class 2 to the binary class 1. After observing the results, it is believed that this last conversion introduces less heterogeneity to the classes, thus generating better results. A fourth hypothesis is not doing the binarisation, using

the ternary original classification. Binarising the dataset caused the loss of valuable information. The reassessment of LIDC-IDRI was made in three classes because it was considered that those were important to have a more consistent dataset, since the classes can vary a lot within each other. Taking advantage of the ternary classification will probably improve the results.

An improvement to these results, also suggested by Shen *et al.* in [3], would be to use the exact malignancy diagnosis of the lung nodules present in LIDC-IDRI. In LIDC-IDRI the malignancy labels used were only suspicion levels of the physicians, which also have lot of divergence in the ratings, regarding the same nodule. LIDC-IDRI dataset has the diagnostic data available for 96 patients. As almost of the nodules incorrectly predicted were juxta-pleural another suggestion is to employ a juxta-pleural nodules segmentation.

5.4 Summary

Conclusions about the first method presented in chapter 4 were given for each relevant semantic feature, creating thus the evaluation norms. Each evaluation norm created was represented by three classes. Using those, a dataset reassessment was done, following the procedure explained in chapter 4. This dataset reassessment was done to the 2,660 nodules evaluated in LIDC-IDRI dataset, for the six relevant semantic features. After the reassessment, a discussion about the differences found between the new and LIDC-IDRI datasets was made. There was found a lot of disparity in the LIDC-IDRI physicians individual evaluations, except in margin. The errors found in calcification assessments were corrected. The implementation of HSCNN with the reassessed dataset was shown, explaining the data preprocessing, demonstrating the results, and discussing it. Comparing the results to malignancy prediction between the folds, the best obtained were from fold 2 with an AUC of 0.74, accuracy of 0.78, sensitivity of 0.83 and an specificity of 0.54. The overall best results to the semantic features were also obtained using fold 2 as test, reported in Table 5.8. The results of this work only surpass the sensitivity results of Shen *et al.* work, concluding with this that a more consistent classification only improved the prediction of positives in the malignancy prediction. Shen *et al.* work obtained as accuracy 0.84, AUC of 0.86, sensitivity of 0.71 and a specificity of 0.89. Comparing the two works, the binarisation in the present was done in a different manner, giving more consideration to malignancy, converting the intermediate classes to the binary class 1, related to malignancy. Having concluded that HSCNN performed better with abnormalities that were evaluated by 3 or more doctors, the dataset was modified, using 1,386 nodules from the 2,660. In Shen *et al.* work were used 4,252, see section 3.3. Despite having more consistent labels, the dataset was smaller, which could also limited the network learning. After seeing the results and observing in which nodules the network had predicted incorrectly was concluded that were in most part nodules those were small, had a lot of matter around, with solid calcification, juxta-pleural, or had ill-defined/irregular margins. It predicted correctly in solitary or obvious nodules, or with pleural-tail. Hypothesis to future improve the network results were given.

Chapter 6

Conclusions and Future Work

The diagnosis of lung cancer is often made by reading CTs. To reach to a malignancy diagnosis, physicians' often analyse semantic features that the abnormality present in the CT has. However, what each semantic feature is and how it could behave is not accurately defined in the literature. The semantic features analysis is very subjective, which generates diverse opinions regarding the diagnosis of the same nodule. This is one of the reasons for an initial diagnosis incorrect. Thereat, CADx aim to help the physicians, giving a second opinion regarding the nodules' malignancy. However, even that its efficiency has already been proved, doctors end up not using it because they can not understand the "how and why" of the system diagnose results. To solve that, Shen *et al.* proposed HSCNN, a network that besides the nodule malignancy diagnose, outputs predictions for the semantic features present in the nodule. In the work proposed by Shen *et al.* the LIDC-IDRI dataset was used. This dataset presents in some semantic features high variability in physicians assessments, especially in lobulation and spiculation, the two features more correlated to malignancy prediction. With that motivation, Shen et al. raised the question: will the decrease of variability in LIDC-IDRI semantic features evaluations improve the HSCNN malignancy prediction? To answer to that question was proposed a methodology composed of three steps that had as results the creation of evaluation criteria for each relevant semantic feature, a more consistent reassessed dataset based on those criteria, and an analyse of the impact in malignancy prediction by using that new dataset in HSCNN. The overall results of HSCNN trained with new dataset only shown improvements compared with Shen *et al.* publication in sensitivity, with a better prediction of when the nodules are truly malignant. The malignancy prediction of this work had an accuracy of 0.78, an AUC of 0.74, a sensitivity of 0.83, and a specificity of 0.54. In the Shen *et al.* work were obtained as accuracy 0.84, for AUC 0.86, for sensitivity 0.71 and for specificity 0.89. However, this comparison is not linear for two main reasons. First, while in this dissertation were used 1,386 abnormalities, in Shen *et al.* work were used 4,252. Despite these new dataset assessments presenting more consistency, the dataset is smaller, which can limit the network learning. Also, the binary conversion was done in a different form in the two works, which may have influenced the final results. Despite these differences, it is concluded that the approach used to implement the new dataset in HSCNN, do not improve the lung nodule malignancy prediction, comparing to

Shen *et al.* results.

Discussion on the results declared that these could have been influenced by the use of binary classes, that cause the loss of relevant information for semantic features and malignancy prediction; by the conversion made from ternary to binary classification; by the small amount of data present; or by the classes imbalance in each task of the dataset. Use the real malignancy diagnosis of the nodules could also have future impact in this work. LIDC-IDRI only presents the suspicion of malignancy by the physicians in each nodule, that is also subject to each one subjectivity. Making segmentation of juxta-pleural nodules could also be a future work implementation.

Summarizing, for future work the following improvements are suggested:

- Change how the classes binarisation was performed;
- Use as input of HSCNN the original reassessed dataset, evaluated with three classes for each semantic feature;
- Exploit more approaches to data augmentation;
- Use the real malignancy diagnosis of LIDC-IDRI dataset lung nodules;
- Employ more approaches to reduce the dataset imbalance;
- Segmentation of juxta-pleura nodules.

Despite the detected problems, all the objectives were overall completed, and the implementation of the previously referred suggestions and consequent performance improvement will further increase the importance of the integration of the proposed method on a CADx system.

Appendix A

Worksheet File

id	patient	no. of evaluation	Evaluation's average							
			texture	calcification	margin	sphericity	lobulation	subtle	spiculation	internalstructure
1	1	4	4.75	6	3.25	3.75	3	5	4.25	1
2	2	2	1.5	6	1.5	4	1	1.5	1	1
3	3	4	4	6	3.25	4	2	5	3	1
4	3	4	4.75	6	4	3.5	1.75	3.5	1.5	1
5	3	4	5	6	5	4.5	1	4	1	1

Figure A.1: Nodule information, and evaluations' average given by the doctors, in the worksheet file.

Evaluation 1								
internalstructure ▾	texture ▾	calcification ▾	margin ▾	sphericity ▾	lobulation ▾	subtlety ▾	spiculation ▾	internalstructure ▾
1	5	6	3	3	3	5	4	1
1	1	6	1	5	1	2	1	1
1	3	6	3	3	2	5	2	1
1	5	6	4	3	1	4	1	1
1	5	6	5	4	1	4	1	1

Figure A.2: Classifications given by a doctor in evaluation 1, in the worksheet file.

image_name	axial	image_name	coronal	image_name	sagittal	image_name	axial	image_name	coronal	image_name	sagittal
nod1_ax.png		nod1_cor.png		nod1_sag.png		nod1_marked_ax.png		nod1_marked_cor.png		nod1_marked_sag.png	
nod2_ax.png		nod2_cor.png		nod2_sag.png		nod2_marked_ax.png		nod2_marked_cor.png		nod2_marked_sag.png	
nod3_ax.png		nod3_cor.png		nod3_sag.png		nod3_marked_ax.png		nod3_marked_cor.png		nod3_marked_sag.png	
nod4_ax.png		nod4_cor.png		nod4_sag.png		nod4_marked_ax.png		nod4_marked_cor.png		nod4_marked_sag.png	
nod5_ax.png		nod5_cor.png		nod5_sag.png		nod5_marked_ax.png		nod5_marked_cor.png		nod5_marked_sag.png	

Figure A.3: Nodule's images without and with the average binary mask of the doctors, in the worksheet file.

Appendix B

Norms




Type	Description	Subtypes	Description
Calcification	Calcium deposited in a nodule	Popcorn, laminated, solid, central 1	Calcification present in the nodule like the examples in the figures below 
		Absent 3	Nodule does not present any type of calcification 
		Non-central, stippled 5	Calcification represented by numerous small dots or calcification with a regular shape in the non-central zone of the nodule 

Figure B.1: Calcification and its classes definition.

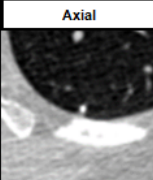
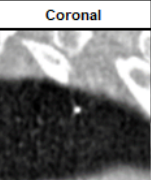

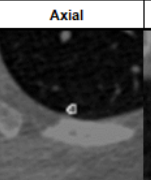
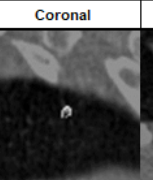
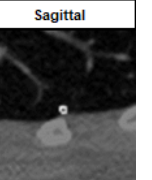
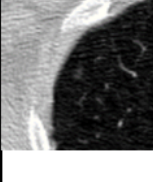
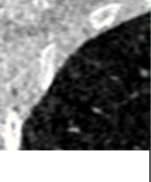
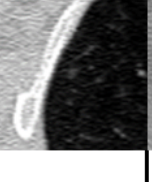
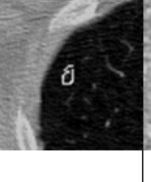
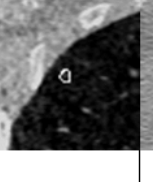
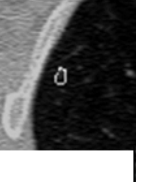
Subtypes	Axial	Coronal	Sagittal	Axial	Coronal	Sagittal
Popcorn, laminated, solid, central 1						
Absent 3						
Non-central, stippled 5	x	x	x	x	x	x

Figure B.2: Very small nodules for each class in calcification.

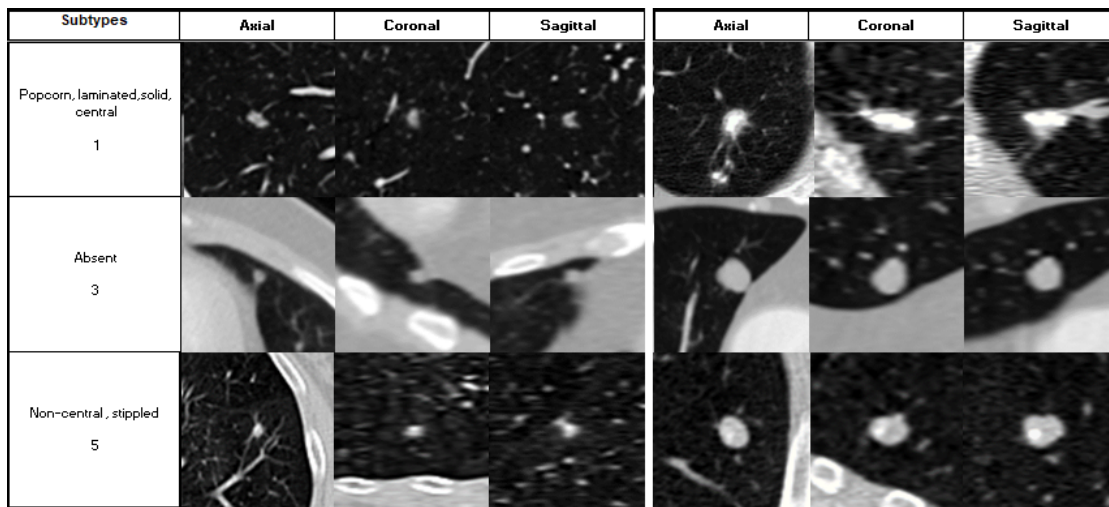


Figure B.3: Small and big nodules for each class in calcification.

Type	Description	Subtypes	Description
Lobulation	Presence of lobules that come from the nodule, excrescences of the tumor at its leading edge	No Lobulated 1	0% of nodule's perimeter has lobules
		Moderately Lobulated 3	Between 25% to 50% of nodule's perimeter has lobules; Number of grooves present in the nodule less or equal to 2
		Lobulated 5	More than 50% of nodule's perimeter with lobules; Number of grooves present in the nodule greater than 2

Figure B.4: Lobulation and its classes definition.

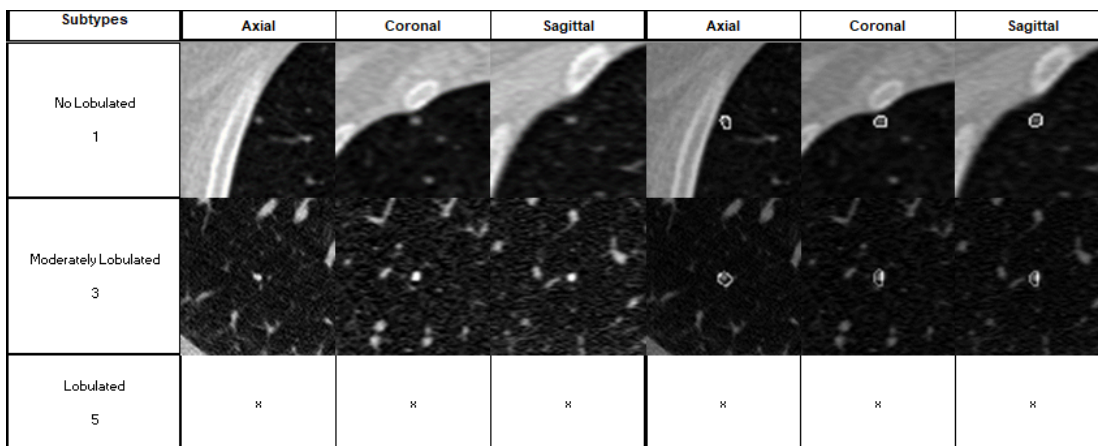


Figure B.5: Very small for each class in lobulation.

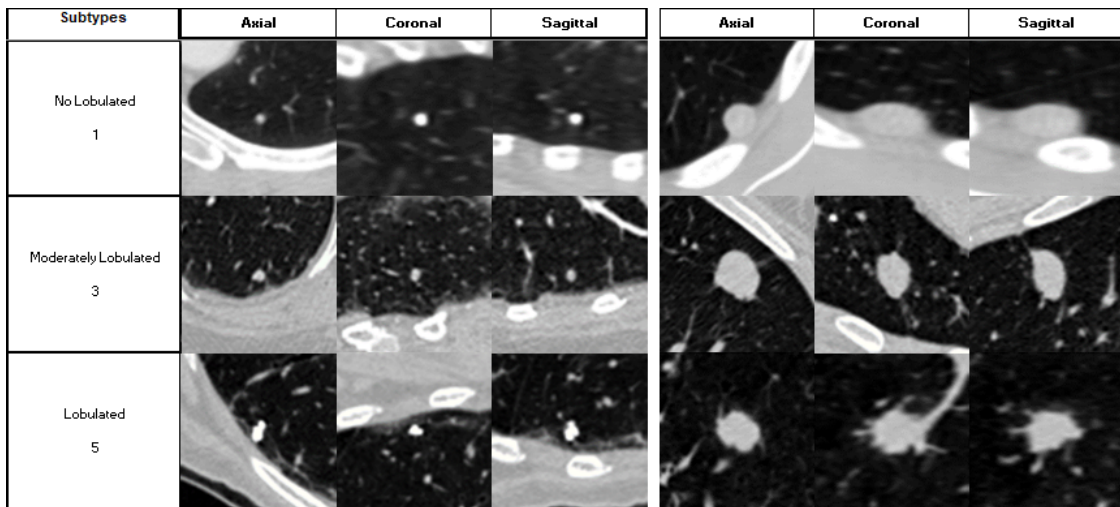


Figure B.6: Small and big nodules for each class in lobulation.

Type	Description	Subtypes	Description
Margin	A boundary line or the area immediately inside the boundary	Sharp 1	Margins well-defined, in the nodule's entirety
		Defined 3	It is possible to see only some defined margins in the nodule - margins are not defined in the nodule's entirety
		Poorly defined 5	It is not possible to see definition on nodule margins; it is possible at most to identify a nodule (e.g. a mass/substance having a higher density than the surrounding background)

Figure B.7: Margin and its classes definition.

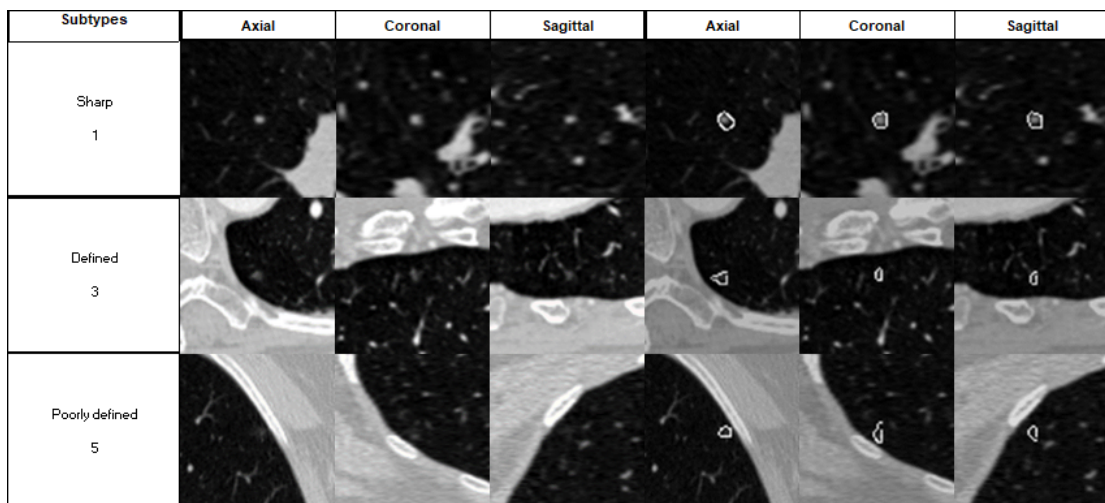


Figure B.8: Very small for each class in margin.

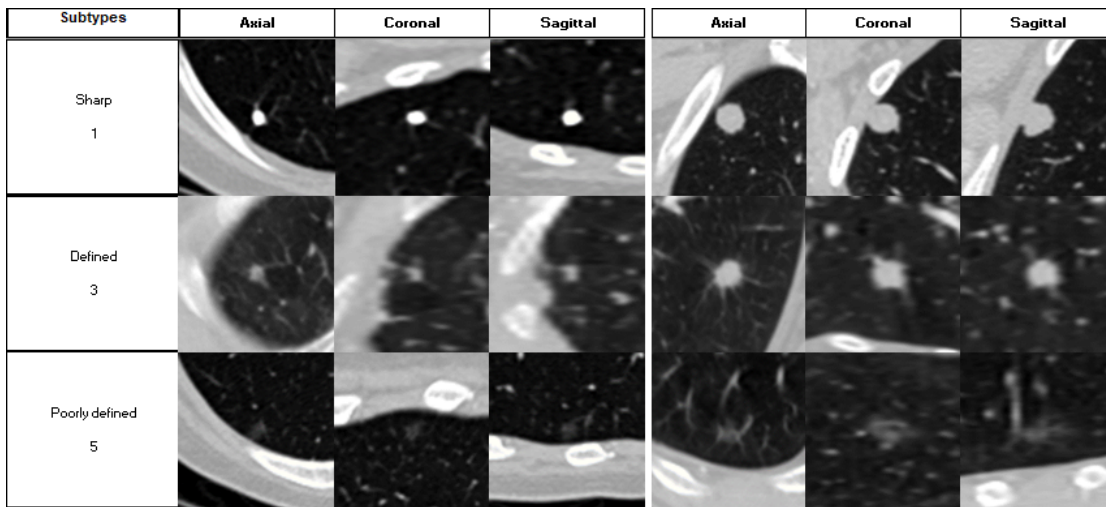


Figure B.9: Small and big nodules for each class in margin.

Type	Description	Subtypes	Description
Spiculation	Covered with or having small needlelike structures (spicules) or fine fleshy points	No Spiculated 1	There is no spiculation present in the nodule margins
		Moderately Spiculated 3	Low spiculation present in the margins - only 1 or 2 spicules with at least one spicule radius greater or equal to the nodule's radius, or more than 2 subtle spicules all with a smaller radius than the nodule's radius
		Spiculated 5	Extremely spiculated margins - with various (more than 2) spicules with at least one spicule radius larger or equal to the nodule's radius.

Figure B.10: Spiculation and its classes definition.

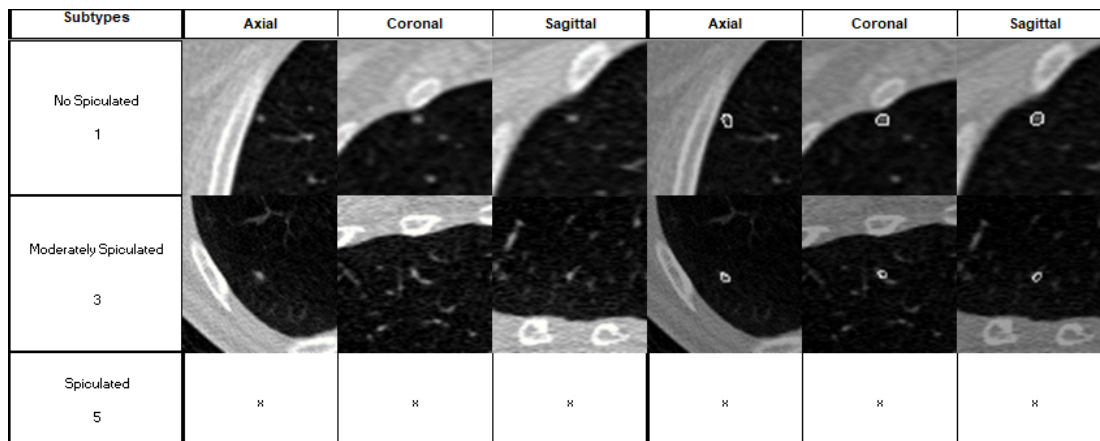


Figure B.11: Very small for each class in spiculation.

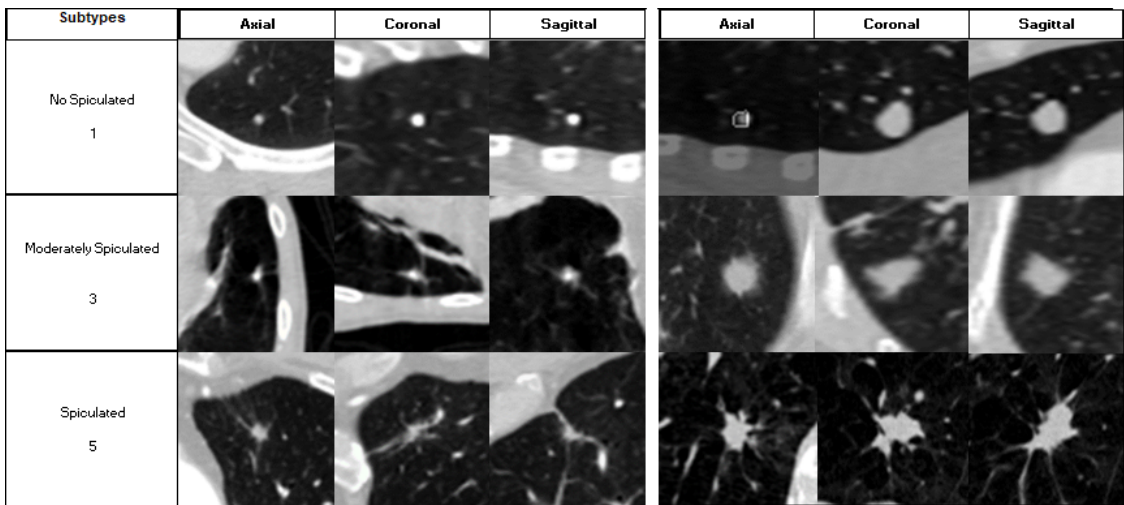


Figure B.12: Small and big nodules for each class in spiculation.

Type	Description	Subtypes	Description
Subtlety	Difficulty of detection relative to surround	Subtle 1	Little visible difference between structures and surroundings of the lung and nodule. Subtly visible nodule
		Fairly subtle 3	Visible difference between structures and surroundings of the lung and nodule. Even if the nodule is subtle, it is visible
		Obvious 5	Obvious difference between structures and surroundings of the lung and nodule. Obvious and easy visibility of the nodule

Figure B.13: Subtlety and its classes definition.

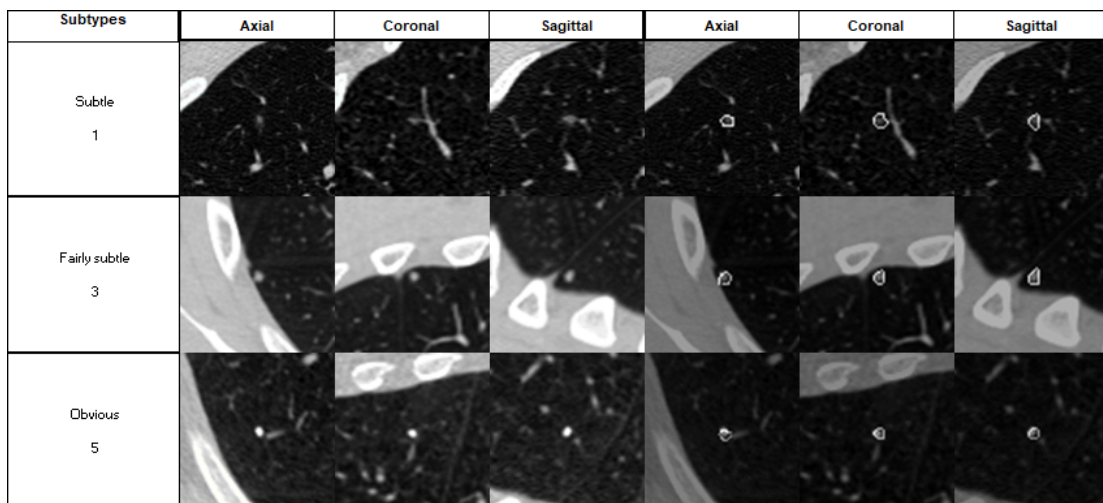


Figure B.14: Very small for each class in subtlety.

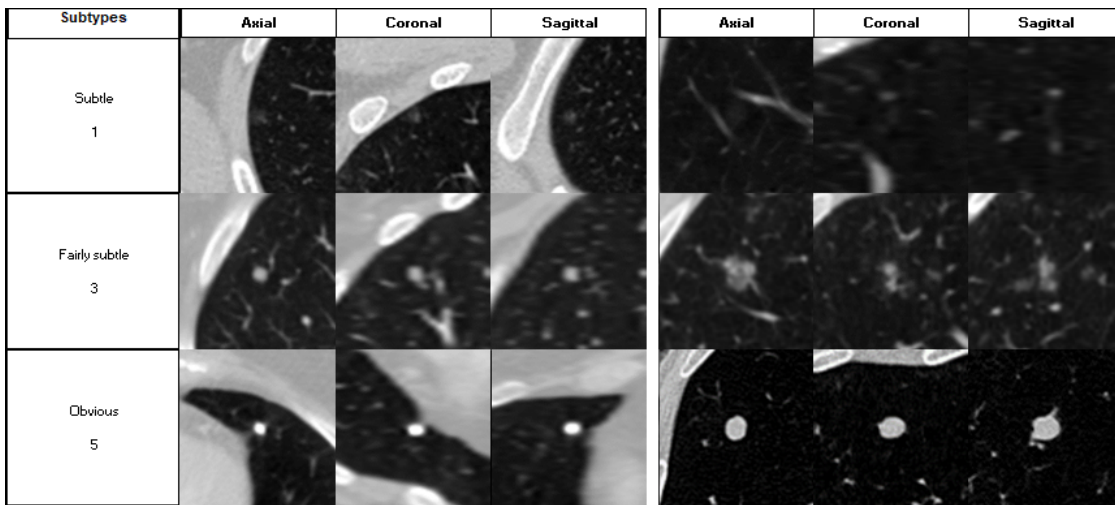


Figure B.15: Small and big nodules for each class in subtlety.

Type	Description	Subtypes	Description
Texture	Internal density of the nodule in terms of the ratio of solid present	Solid 1	Nodule composed with a greater proportion of solid than non-solid components
		Non-solid 3	Low visible nodule with low attenuation, with a greater proportion of non-solid than solid components
		Part-solid 5	Nodules with a higher percentage density located centrally or peripherally, with an equal proportion of solid and non-solid components

Figure B.16: Texture and its classes definition.

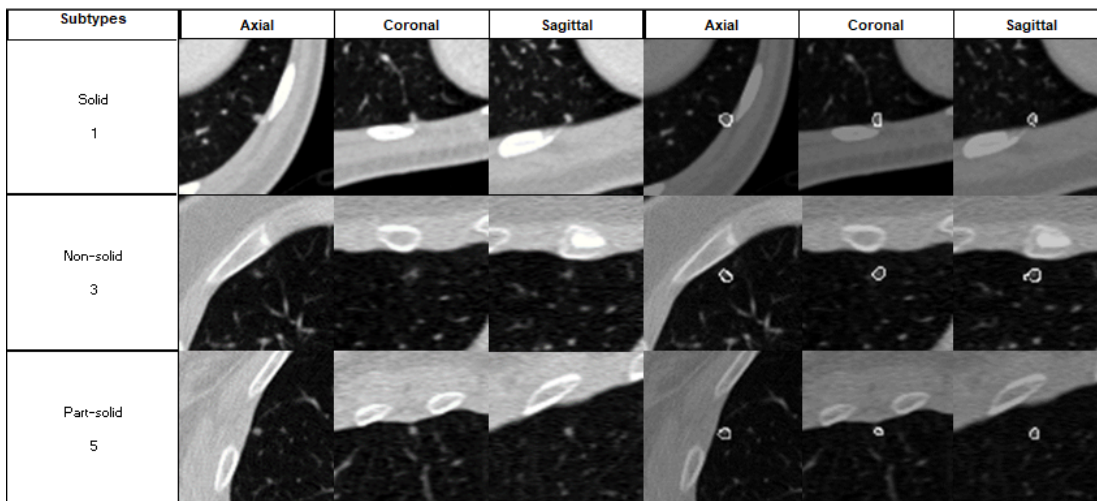


Figure B.17: Very small for each class in texture.

Subtypes	Axial	Coronal	Sagittal	Axial	Coronal	Sagittal
Solid 1						
Non-solid 3						
Part-solid 5						

Figure B.18: Small and big nodules for each class in texture.

References

- [1] S. Sone M. Hasegawa and S. Takashima. Growth rate of small lung cancels detected on mass ct screening. *Tire British Journal of Radiology*, pages 1252–1259, December 2000.
- [2] W Jorritsma, Fokie Cnossen, and Peter Van Ooijen. Improving the radiologist-cad interaction: Designing for appropriate trust. *Clinical Radiology*, 70, October 2014.
- [3] Shiwen Shen, Simon X. Han, Deni Aberle, Alex A. T. Bui, and Willliam Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. June 2018.
- [4] 3rd Armato, Samuel G, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, Edwin J R Van Beeke, David Yankelevitz, Alberto M Biancardi, Peyton H Bland, Matthew S Brown, Roger M Engelmann, Gary E Laderach, Daniel Max, Richard C Pais, David P Y Qing, Rachael Y Roberts, Amanda R Smith, Adam Starkey, Poonam Batrah, Philip Caligiuri, Ali Farooqi, Gregory W Gladish, C Matilda Jude, Reginald F Munden, Iva Petkowska, Leslie E Quint, Lawrence H Schwartz, Baskaran Sundaram, Lori E Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Castele, Sangeeta Gupte, Maha Sallamm, Michael D Heath, Michael H Kuhn, Ekta Dharaiya, Richard Burns, David S Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, and Barbara Y Croft. Data from lidc-idri. the cancer imaging archive. 2015. Available in <http://doi.org/10.7937/K9/TCIA.2015.L09QL9SX>.
- [5] International Agency for Research on Cancer. Lung-fact-sheet, March 2019. Global Cancer Observatory: Available in <http://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf>.
- [6] Edgar R., Mazor Y., Rinon A., Blumenthal J., Golan Y., Buzhor E., Livnat I., Ben-Ari S., Lieder I., Shitrit A., Gilboa Y., Ben-Yehudah A., Edri O., Shraga N., Bogoch Y., Leshansky L., Aharoni S., West MD., Warshawsky D., and Shtrichman R. Lifemap discoveryTM: the embryonic development, stem cells, and regenerative medicine research portal. *PLoS One* 17;8(7), July 2013. Available in <https://discovery.lifemapsc.com/library/images/lung-anatomy-and-structure>.
- [7] Martin Dolejší and Jan Kybic. Automatic two-step detection of pulmonary nodules. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, 6514, March 2007.
- [8] Ashis Kumar Dhara, Sudipta Mukhopadhyay, and Niranjana Khandelwal. Computer aided detection and analysis of pulmonary nodule from ct images: A survey. *IETE Technical Review*, November 2011.

- [9] CS231N. Convolutional neural networks for visual recognition, 2014. Available in <http://cs231n.github.io/convolutional-networks/>.
- [10] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663 – 673, 2017.
- [11] Kui Liu and Guixia Kang. Multiview convolutional neural networks for lung nodule classification. *International Journal of Imaging Systems and Technology*, 27(1):12–22, March 2017.
- [12] Kang G., Liu K., Hou B., and Zhang N. 3d multi-view convolutional neural networks for lung nodule classification. *PLoS ONE 12(11): e0188290*, November 2017.
- [13] George E Kikano, Andre Fabien, and Robert Schilz. Evaluation of the solitary pulmonary nodule. *American family physician*, 92:1084–1091, December 2015.
- [14] Marcelo Sánchez, Mariana Benegas, and Ivan Vollmer. Management of incidental lung nodules <8 mm in diameter. *Journal of thoracic disease*, 10(Suppl 22):S2611–S2627, August 2018.
- [15] Pia Opulencia, David Channin, Daniela Raicu, and Jacob Furst. Mapping lidc, radlex, and lung nodule image features. *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, 24:256–70, March 2010.
- [16] Luís Gonçalves, Jorge Novo, António Cunha, and Aurélio Campilho. Evaluation of the degree of malignancy of lung nodules in computed tomography images. March 2017.
- [17] International Agency for Research on Cancer. Lung-fact-sheet, 2018. Global Cancer Observatory: Available in <http://gco.iarc.fr/today/online-analysis-map>.
- [18] Stojan Trajanovski, Dimitrios Mavroeidis, Christine Leon Swisher, Binyam Gebrekidan Gebre, Bas Veeling, Rafael Wiemker, Tobias Klinder, Amir Tahmasebi, Shawn M. Regis, Christoph Wald, Brady J. McKee, Heber MacMahon, and Homer Pien. Towards radiologist-level cancer risk assessment in CT lung screening using deep learning. *CoRR*, abs/1804.01901, 2018.
- [19] Stefan Diederich, Dag Wormanns, Michael Semik, Michael Thomas, Horst Lenzen, Nikolaus Roos, and Walter Heindel. Screening for early lung cancer with low-dose spiral ct: Prevalence in 817 asymptomatic smokers. *Radiology*, 222(3):773–781, 2002. PMID: 11867800.
- [20] CI. Henschke. Early lung cancer action project: overall design and findings from baseline screening. *Cancer*, vol. 89, pages 2474–2482, 2000.
- [21] Screaton N.J Breads Moore, C.J. Classification, staging and prognosis of lung cancer. *European Journal of Radiology*, 45, pages 8–17, 2003.
- [22] D. J. Bell S. E. Marley P. Guo H. Mann M. L. Scott L. H. Schwartz D. C. Ghiorghiu B. Zhao, Y. Tan. Exploring intra-and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on ct scans reconstructed at different slice intervals. *European journal of radiology* 82, page 959–968, 2013.
- [23] H.T Winer-Muram. The solitary pulmonary nodule 1. *Radiology*, 239, pages 39–49, 2006.

- [24] D.P. Godoy, M.C. Naidich. Subsolid pulmonary nodules and the spectrum of peripheral adenocarcinomas of the lung: Recommended interim guidelines for assessment and management 1. *Radiology*, 253, pages 606–622, 2009.
- [25] R. Yan J. Lee L. C. Chu C. T. Lin A. Hussien J. Rathmell B. Thomas C. Chen et al. P. Huang, S. Park. Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study. *Radiology* 286, page 286–295, 2017.
- [26] Yann LeCun, Y Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, May 2015.
- [27] Tom Brosch, Youngjin Yoo, David Li, Anthony Traboulsee, and Roger Tam. Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning. volume 17, September 2014.
- [28] Aydın Kaya and Ahmet Burak Can. A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. *Journal of Biomedical Informatics*, 56:69 – 79, 2015.
- [29] Matthew C. Hancock and J Magnan. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: Probing the Lung Image Database Consortium dataset with two statistical learning methods. *Journal of Medical Imaging*, 3:44504, 2016.
- [30] Disabled World. Lung cancer: Symptoms causes treatment, March 2015. Available in <https://www.disabled-world.com/health/cancer/lung>.
- [31] Liga Portuguesa contra o Cancro. A melhor forma de prevenir o cancro do pulmão é deixar de fumar ou nunca começar, 2010. Available in <https://www.ligacontracancro.pt>.
- [32] American Lung Association Scientific. Lung cancer fact sheet, 2018. Available in <https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html>.
- [33] Team National Lung Screening Trial Research, D.R. Aberle, and A.M. Adams. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*, 365, January 2011.
- [34] Furqan Shaukat, Gulistan Raja, Ali Gooya, and Alejandro F. Frangi. Fully automatic detection of lung nodules in ct images using a hybrid feature set. *Medical physics*, 44 7:3615–3629, 2017.
- [35] Gilbert Ferretti, Laure Felix, Géraldine Serra-Tosio, Christian Brambilla, Denis Moro-Sibilot, Pierre-Yves Brichon, M Coulomb, and Sylvie Lantuejoul. [non-solid and part-solid pulmonary nodules on ct scanning]. *Revue des maladies respiratoires*, 24:1265–76, January 2008.
- [36] MD Peter B. O’donovan. The radiologic appearance of lung cancer. *Lung Cancer, Oncology Journal* 11, September 1997.
- [37] Macedo Firmino, Giovani Angelo, Higor Morais, Marcel R. Dantas, and Ricardo Valentim. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *BioMedical Engineering OnLine*, 15(1):2, January 2016.

- [38] Luís Gonçalves, Jorge Novo, and Aurélio Campilho. Feature definition, analysis and selection for lung nodule classification in chest computerized tomography images. April 2016.
- [39] Francesco Ciompi, Bartjan de Hoop, Sarah J. van Riel, Kaman Chung, Ernst Th. Scholten, Matthijs Oudkerk, Pim A. de Jong, Mathias Prokop, and Bram van Ginneken. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, 26(1):195 – 202, 2015.
- [40] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943.
- [41] Hassan Ramchoun, Mohammed Amine, Mohammed Amine Janati Idrissi, Youssef Ghanou, and Mohamed Ettaouil. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4:26–30, January 2016.
- [42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.
- [43] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.
- [44] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130, 1988.
- [45] Yann Lecun, Leon Bottou, Y Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, December 1998.
- [46] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.
- [47] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.
- [48] F. Mosteller and J. W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology*, Vol. 2. Addison-Wesley, 1968.
- [49] 3rd Armato, Samuel G, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, Edwin J R Van Beeke, David Yankelevitz, Alberto M Biancardi, Peyton H Bland, Matthew S Brown, Roger M Engelmann, Gary E Laderach, Daniel Max, Richard C Pais, David P Y Qing, Rachael Y Roberts, Amanda R Smith, Adam Starkey, Poonam Batrah, Philip Caligiuri, Ali Farooqi, Gregory W Gladish, C Matilda Jude, Reginald F Munden, Iva Petkovska, Leslie E Quint, Lawrence H Schwartz, Baskaran Sundaram, Lori E Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Castele, Sangeeta Gupte, Maha Sallamm, Michael D Heath, Michael H Kuhn, Ekta Dharaiya, Richard Burns, David S Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, and Barbara Y Croft. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, February 2011.

- [50] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. pages 318–362, 1986.
- [51] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5:4006 EP –, June 2014.
- [52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
- [53] Qi Dou, Hao Chen, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Multi-level contextual 3d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, PP:1–1, September 2016.
- [54] Annette McWilliams, Martin C Tammemagi, John R Mayo, Heidi Roberts, Geoffrey Liu, Kam Soghrati, Kazuhiro Yasufuku, Simon Martel, Francis Laberge, Michel Gingras, Sukhinder Atkar-Khattra, Christine D Berg, Ken Evans, Richard Finley, John Yee, John English, Paola Nasute, John Goffin, Serge Puksa, and Stephen Lam. Probability of cancer in pulmonary nodules detected on first screening ct. *The New England journal of medicine*, 369:910–9, September 2013.
- [55] Stephen Swensen, Marc Silverstein, Duane Ilstrup, Cathy Schleck, and Eric Edell. The probability of malignancy in solitary pulmonary nodules. application to small radiologically indeterminate nodules. *Archives of internal medicine*, 157:849–55, April 1997.
- [56] Omnia Elsayed, Khaled Mahar, M Kholief, and Hatem A. Khater. Automatic detection of the pulmonary nodules from CT images. pages 742–746, 2015.
- [57] Patricia Lorena Arancibia Hernández, Teresa Taub Estrada, Alejandra López Pizarro, María Lorena Díaz Cisternas, and Carla Sáez Tapia. Breast calcifications: Description and classification according to BI-RADS 5th edition. *Revista Chilena de Radiología*, 22(2):80–91, 2016.
- [58] Annemie Snoeckx, Pieter Reyntiens, Damien Desbuquoit, Maarten J Spinhoven, Paul E Van Schil, Jan P van Meerbeeck, and Paul M Parizel. Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology. *Insights into imaging*, 9(1):73–86, February 2018.
- [59] Robin Smithuis and Ottob van Delden. The radiology assistant, inc, May 2007. Available in <http://www.radiologyassistant.nl/en/p420ccf4d2c5c0/click-for-more-information.html>.
- [60] J W Gurney. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. *Radiology*, 186(2):405–413, February 1993.
- [61] M C Mahoney, R T Shipley, H L Corcoran, and B A Dickson. CT demonstration of calcification in carcinoma of the lung. *AJR. American journal of roentgenology*, 154(2):255–258, February 1990.

- [62] Dr Jayanth Keshavamurthy and A.Prof Frank Gaillard et al. Fat containing solitary pulmonary nodule. 2019. Available in <https://radiopaedia.org/articles/fat-containing-solitary-pulmonary-nodule?lang=us>.
- [63] S S Siegelman, N F Khouri, W W Jr Scott, F P Leo, U M Hamper, E K Fishman, and E A Zerhouni. Pulmonary hamartoma: CT findings. *Radiology*, 160(2):313–317, August 1986.
- [64] Agnieszka Choromańska and Katarzyna J Macura. Evaluation of solitary pulmonary nodule detected during computed tomography examination. *Polish journal of radiology*, 77(2):22–34, 2012.
- [65] Ko JP, Truong MT1, Sabloff BS. Multidetector CT of solitary pulmonary nodules. *Radiologic clinics of North America*, 48(1):141–155, January 2010.
- [66] J W Gurney. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. *Radiology*, 186(2):405–413, February 1993.
- [67] D R Baldwin, S W Duffy, N J Wald, R Page, D M Hansell, and J K Field. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. *Thorax*, 66(4):308–313, April 2011.
- [68] Daiwei Han, Marjolein A. Heuvelmans, and Matthijs Oudkerk. Volume versus diameter assessment of small pulmonary nodules in ct lung cancer screening. *Translational Lung Cancer Research*, 6(1), 2017.
- [69] Helen T Winer-Muram. The solitary pulmonary nodule. *Radiology*, 239(1):34–49, April 2006.
- [70] J J Erasmus, J E Connolly, H P McAdams, and V L Roggli. Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 20(1):43–58, 2000.
- [71] M D Seemann, A Staebler, T Beinert, H Dienemann, B Obst, M Matzko, C Pistitsch, and M F Reiser. Usefulness of morphological characteristics for the differentiation of benign from malignant solitary pulmonary lesions using HRCT. *European radiology*, 9(3):409–417, 1999.
- [72] Jing Yang, Hailin Wang, Chen Geng, Yakang Dai, and Jiansong Ji. Advances in intelligent diagnosis methods for pulmonary ground-glass opacity nodules. *Biomedical engineering online*, 17(1):20, 2018.
- [73] Tomoyuki Takeguchi Sumiaki Matsumoto Yoshiharu Ohno Kota Aoyagi Hitoshi Yamagata Atsushi Yaguchi, Tomoya Okazaki. Semi-automated segmentation of solid and ggo nodules in lung ct images using vessel-likelihood derived from local foreground structure. 9414, 2015.
- [74] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

- [75] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. In Sebastien Ourselin, Daniel C Alexander, Carl-Fredrik Westin, and M Jorge Cardoso, editors, *Information Processing in Medical Imaging*, pages 588–599, Cham, 2015. Springer International Publishing.